HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of Physics

have examined a dissertation entitled

Simulating and Imaging Supermassive Black Hole Accretion Flows

presented by Andrew Alan Chael

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature _____

Typed name: Professor Cora Dvorkin, Chair

Signature _____

Typed name: Professor Ramesh Narayan, Co-Chair

Signature _____

Typed name: Dr. Sheperd Doeleman

Signature _____

Typed name: Professor Peter L. Galison

Signature _____

Typed name: Dr. Michael D. Johnson

Date: April 30, 2019

Simulating and Imaging Supermassive Black Hole Accretion Flows

A dissertation presented
by
Andrew Alan Chael
to
The Department of Physics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Physics

Harvard University
Cambridge, Massachusetts
April 2019

Simulating and Imaging Supermassive Black Hole Accretion Flows

Abstract

Supermassive black holes exist in the centers of nearly all galaxies. They are most frequently surrounded by hot, thick, and low-radiative-efficiency accretion flows, including in the Galactic Center radio source Sagittarius A* (Sgr A*) and at the base of the relativistic jet in the giant galaxy M87. In this thesis, I study these objects in two ways: with numerical simulations and with image reconstruction of data from the Event Horizon Telescope (EHT), a mm-wavelength Very Long Baseline Interferometry (VLBI) array. In the first part, I simulate both Sgr A* and M87 using two-temperature, radiative, general relativistic magnetohydrodynamics (GRMHD). Including radiation and thermodynamics in GRMHD simulations is necessary to predict the electron temperatures and emission from these objects, as electrons and ions in hot flows are far from mutual equilibrium. I also develop a method for moving beyond thermal equilibrium in simulations by evolving a full population of nonthermal electrons. In the second part, I describe a framework for imaging VLBI data with regularized maximum likelihood methods, and I detail its implementation in the `eht-imaging` software library. This framework allows VLBI data to be imaged with no a priori calibration, using only robust closure quantities. Finally, I present images from the first full EHT campaign on M87 reconstructed using `eht-imaging` and other methods. I conclude by describing measurements of the black hole shadow and mass from these first images of a supermassive black hole.

Page intentionally left blank

Contents

Acknowledgments

This thesis, and my entire graduate school career, would not have been possible without the ceaseless support of a large and diverse community of people. More than any particular insight or work of my own, it was their advice, friendship, and perspective that brought me to this point. I hope you'll bear with a little sentimentality as I thank them in turn.

First of all, I would like to thank my three primary advisers – Ramesh Narayan, Michael D. Johnson, and Shep Doeleman. My unique path through graduate school in fusing research in both numerical simulations and VLBI imaging would have left me completely lost without their constant guidance. First, I'd like to thank Ramesh for his patience, wisdom, and overall example of how to engage with research with curiosity and flexibility. I have never met anyone better able to pose an incisive question that gets right at the heart of an issue; Ramesh's example in how to listen and think when reading papers, listening to talks, and writing code has transformed me into a much better researcher. Additionally, I would probably still be finishing my first paper right now if not for Ramesh's ability to get me moving with a well-placed deadline. Second, I thank Michael for his patience, his advice, his leadership, and his tireless work on my behalf and on behalf of the entire EHT project. I grew up in the EHT and in astronomy with Michael; watching his success over the last five years has always pushed me to work harder and live up to his example (and his ever-increasing expectations). Third, I wouldn't have had any of the opportunities I've enjoyed

over the last five years if Shep hadn't taken me on board the fledgling SAO EHT group in the summer of 2014. Shep's energy, enthusiasm, and near-inhuman drive in pushing the EHT project forward has been a consistent source of inspiration to me and the whole team at Harvard. I never would have made it to this point without the belief all three of these men put in me.

Thank you also to the two other members of my committee: Cora Dvorkin, for her generosity in staying on as an advisor even after I decided my path wasn't in cosmology, and Peter Galison, for providing his wisdom and perspective on science, writing, and collaboration over the last year. Thank you also to Olek Sądowski, who taught me how to run my first GRMHD simulation and trusted me with KORAL. To my fellow graduate students in Ramesh's group – Michael Rowan, Xinyi Guo, and Brandon Curd – thank you for setting such high standards of dedication and passion for your research, for keeping me company in the office (very) late into the night, and for putting up with all the bugs I've introduced into KORAL. In particular, the simulations in Chapters 2- 3 would not have been possible without Michael's collaboration and help in understanding the results of his reconnection simulations.

The gratitude I feel toward the whole EHT group at the CfA is impossible to express completely in words. Through long days and nights of work, trips around the world, discoveries, arguments, venting sessions, and many, many, many telecons, we accomplished something truly amazing and developed into a family. Most of all, I have to thank Katie Bouman, my partner in imaging crime for the last five years. Katie taught me more or less everything I know about image reconstruction, and she motivates me every day with her insight, persistence, inhuman drive, and amazing PowerPoint art. I also want to particularly thank Kazu Akiyama for making my imaging work better through his friendly competition, Lindy Blackburn for setting a high bar of rigor and quality for the whole group, Sara Issaoun for knowing basically everything about VLBI and for our shared taste in late night karaoke, Daniel Palumbo for his humor and encouragement through incredible work

on thankless tasks, Dom Pesce for his unflappable cheer and for providing us with a constant supply of Oreos, and Maciek Wielgus for making the group more entertaining with his uniquely wry perspective. My gratitude extends to everyone else in the Boston-area EHT group, past and present, especially: Mislav Baloković, Rodrigo Córdova, Joseph Farah, Vincent Fish, Kari Haworth, David James, Atish Kamble, Jim Moran, Rurik Primiani, Alex Raymond, Hotaka Shiokawa, Jason SooHoo, Laura Vertatschitsch, Jonathan Weintroub, and André Young. Furthermore, none of our group's progress would have been possible without the incredible support of the staff at the CfA and BHI; I'd particularly like to thank Barbara Elfman and Nina Zonnevylle for their help and patience with me whenever I messed up on reimbursement forms. Working with a group this talented has been an education and an adventure, and I can't wait for what comes next.

Outside of the local group, I'd particularly like to thank Jason Dexter for his help developing and using his `grtrans` code and for sharing many useful insights, John Wardle for sharing his wisdom on VLBI polarimetry and polarimetric imaging, Lorenzo Sironi for sharing his expertise on plasma microphysics and patiently correcting my misunderstandings, Chi-kwan Chan for his kindness and his example in how to code elegantly and efficiently, and Sasha Tchekhovskoy for his candid advice on my postdoc applications and for his overall mentorship. In addition, many other members of the global EHT collaboration educated me on topics from GRMHD theory to VLBI data calibration, and I'd like to thank them all – particularly David Ball, Geoffrey Bower, Avery Broderick, Ilje Cho, Jordy Davelaar, Roger Deane, Heino Falcke, Charles Gammie, José Gómez, Mareki Honma, Shoko Koyama, Michael Janßen, Svetlana Jorstad, Jae-Young Kim, Laurent Loinard, Sera Markoff, Dan Marrone, Alan Marscher, Lia Medeiros, Monika Mościbrodzka, Gisela Ortiz-León, Feryal Özel, Oliver Porth, Dimitrios Psaltis, Venkatessh Ramakrishnan, Freek Roelofs, Ben Ryan, Fumie Tazaki, and Paul Tiede – for their help, conversation, and guidance. I'm excited to keep making discoveries with all of you.

I also extend my gratitude to all of my friends in Cambridge, particularly my intrepid board game group – James Mitchell, Ron and Lan Alexander, and Ceci Mancuso – for their company and skill in saving the world many times over. I also thank James and Maryrose Barrios, my original G1 cat crew, for all the midnight CVS runs, dinner parties, and tea tastings that launched this crazy graduate school experience. My thanks as well to all the other physics friends who were there for me when I needed it, particularly Ellen Klein and Joe Olson; my thanks as well to Keith Mason for his friendship and for sharing many new musicals with me. Finally, I would not have survived graduate school without the friendship, gossip, and snacks provided by the real deans of the Harvard physics graduate program: Lisa Cacciabaudo and the legendary Carol Davis.

My time at Harvard was indelibly shaped by my four years as a resident tutor in Dunster House. I'd like to thank the whole house staff, particularly Sean Kelly, Cheryl Chen, Michael Uy, Roger Porter, Diana Hovsepian, and Rachel Barbarisi. They welcomed me into the house, allowed me to live in rent-free comfort for four years, and helped build a warm and stimulating community I was always excited to come home to. I also want to acknowledge my debt to both Ann Porter and John Pomeroy, two Dunster fixtures who passed away in my time in the House; thank you for anchoring this community, and for your conversation and trust. Furthermore, my time in Dunster was shaped by all of the amazing undergraduates I was privileged to briefly mentor. I have been consistently inspired by their drive, their curiosity, and their kindness over the past four years. Finally, I thank all of my fellow resident tutors for all of their help, support, and advice. I particularly thank my fellow 2015 tutors: Jordan and Katie Anderson for all their help with fellowship advising, Gregory Davis for his kindness, forthright humor, and many unexpected anecdotes, and Jennifer Hsaio for her all-around virtuosity and for always being there to lend a listening ear over a brain break snack.

I would be remiss to not acknowledge that my love for science and my whole outlook on education was shaped by Carleton College and everyone I met there. I thank the entire Carleton physics and

astronomy department for laying the foundation for everything I learned in graduate school. In particular, I am forever grateful to my undergraduate research advisor, Joel Weisberg, for opening so many opportunities for me and for his strong example in both how to conduct research and live ethically in the world. As always, thank you to my Carleton friends, who made college everything it should have been and who continue to support me and brighten my life: Tom Callister, for keeping me sharp; Dan Ackerman, for his sardonic wit and earnest soul; Amelia Schlossberg, for her emotional support and moral example; Rachel Porcher, for her brilliance and exactly-my-wavelength humor; Mallika Jayaraman, for her tireless work in support of what's right; Steve Moran, for his heart; Fadi Hakim, for his insight; and Charlotte Pfeifer, for being there to talk about anything and everything and for understanding me completely.

I love my home state of New Mexico, and I express my deepest thanks to everyone there that touched my life and set me on this path, especially my teachers and friends. Thank you especially to Kristin Cordwell, for being my friend since high school and for always being there with a welcoming floor whenever I decided to come see a show in New York, and to Katie (KT) Morgans, for anchoring me in home through her friendship over more than a decade.

Finally, thank you to the four most important people in my life, who support me and ground me in and through everything. To my partner Jason Cohn: you are brilliant, funny, and the person I most want to talk to every day. Through all the trials of this process and the difficulties of living apart, you gave me strength, support, and the occasional necessary rude awakening. As Shevek and Takver also discovered, my love for you is felt most in the unique "exhilaration in finding the bond is stronger, after all, than all that tries the bond." To my brother Nathan: I've been so proud watching you become the empathetic and brilliant person you are. Thank you for your curiosity, your care for others, and for your pride in me. Finally, to my parents: thank you for giving me the foundations of who I am. Ten years ago I had absolutely no intention of following you into science

or into a physics PhD program. I don't know if you always knew I'd come around eventually, but now I realize how much of me is rooted in your example, and how I am better for it. Thank you for always being there to listen and advise me, for your support in choices you probably didn't understand at the time, and most of all for steeping me in science, books, music, cooking, and love. This thesis – and these past twenty-eight years – are due to you.

To all those named here, and to everyone else who has offered me their inspiration, kindness, company, and laughter throughout my life, thank you.

* * * *

Craig Walker and Jae-Young Kim provided the VLBA and GMVA images of M87 in Chapter 3. Hotaka Shiokawa, Avery Broderick, Roman Gold, Monica Mościbrodzka, and C.K. Chan provided the GRMHD and semi-analytic model images used in Chapter 5, and Svetlana Jorstad and Alan Marscher provided the 43 GHz VLBA image of 3C 273. The ALMA data set used in Chapter 5 is ADS/JAO.ALMA#2011.0.00015.SV, and Craig Walker provided the 7 mm VLBA M87 data used in the same chapter. The 2017 EHT data and images in Chapters 6– 7 are from The Event Horizon Telescope Collaboration et al. (2019c) and The Event Horizon Telescope Collaboration et al. (2019d). Paul Tiede and Avery Broderick provided the simulated data sets used for the mass calibration in Figure 7.9.

I was supported at various times in this work by NSF grants AST-1440254 and AST-1312651. The simulations in Chapter 2– 4 were done on the TACC Stampede/Stampede2 computers using computational support from NSF via XSEDE resources (grant TG-AST080026N). This work was conducted at the Black Hole Initiative at Harvard University, which is supported by a grant from the John Templeton Foundation. It never would have been finished without the support of the BHI's reliable coffee machine.

To my parents.

Page intentionally left blank

"Jamais je ne restais plus d'une journée à Tipasa. Il vient toujours un moment où l'on a trop vu un paysage, de même qu'il faut longtemps avant qu'on l'ait assez vu. Les montagnes, le ciel, la mer sont comme des visages dont on découvre l'aridité ou la splendeur, à force de regarder au lieu de voir. Mais tout visage, pour être éloquent, doit subir un certain renouvellement. Et l'on se plaint d'être trop rapidement lassé quand il faudrait admirer que le monde nous paraisse nouveau pour avoir été seulement oublié."

–Albert Camus, *Noces à Tipasa*

"She wept in pain, because she was free. What she had begun to learn was the weight of liberty. Freedom is a heavy load, a great and strange burden for the spirit to undertake. It is not easy. It is not a gift given, but a choice made, and the choice may be a hard one. The road goes upward toward the light; but the laden traveller may never reach the end of it."

–Ursula K. Le Guin, *The Tombs of Atuan*

"Black holes collect problems faster than they collect matter."

–Carl Sagan, *Contact*

Page intentionally left blank

Introduction

Supermassive Black Holes: Light and Shadow

In my first year of graduate school, I read Ursula K. Le Guin's *The Left Hand of Darkness* for the first time. Arguably her most consequential work, the 1969 novel draws its title from a poem shared between two characters near death on the icy tundra of a planet called Winter:

> Light is the left hand of darkness
> and darkness the right hand of light.

This ancient idea, the dualism of darkness and light, life and death, the clear and the obscure, fascinated Le Guin throughout her life and work. Le Guin's voice became my constant companion as I read my way through her novels and short stories amidst the ups and downs, stress and exhilaration, darkness and light of six long years of graduate school. Through it all, I have increasingly drawn a connection between this idea and the subjects of my research. After all, no object has much more to do with the interplay of light and darkness than a black hole.

0.1 Black holes

In the midst of the darkness of World War I, Schwarzschild (1916) derived the first (non-trivial) exact solution of Einstein's field equations of general relativity (GR: Einstein, 1916). Schwarzschild's

solution describes the spacetime around a spherically symmetric point mass M. The line element of a particle traversing this spacetime in spherical polar coordinates (t, r, θ, ϕ) is

$$ds^2 = -\left(1 - \frac{2\,GM}{c^2 r}\right) c^2 dt^2 + \left(1 - \frac{2\,GM}{c^2 r}\right)^{-1} dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta\; d\phi^2, \tag{0.1}$$

where G is Newton's constant and c is the speed of light. The Schwarzchild metric has a characteristic length-scale, or gravitational radius, $r_{\rm g} = GM/c^2$.

Equation 0.1 diverges at two radii: at $r = 0$ and at the Schwarzschild radius $r = r_{\rm S} = 2r_{\rm g}$. While the singularity at the origin is a true point of infinite curvature, in 1933 Lemaître showed that the apparent singularity at $r = r_{\rm S}$ is a coordinate artifact. An observer traversing this radius would experience nothing abnormal. Yet to an observer at infinity, an object falling toward $r_{\rm S}$ would appear to take an infinite time, and the signals they emit would infinitely redshift to perfect darkness. Finkelstein (1958) called this surface at $r_{\rm S}$ an event horizon, a "perfect unidirectional membrane," or a one-way causal boundary from which even light cannot escape. Later in the century, more and more solutions to the GR field equations with similar event horizons began to emerge – most notably the Kerr (1963) solution describing a point mass M with nonzero angular momentum J. Now, modern relativists take this dark, one-way surface as the defining feature of all black holes.

Einstein himself thought that true black holes could never form in the universe, and that the Schwarzchild metric would only be a valid solution outside the surface of objects with finite radius $r > r_{\rm S}$. However, the black hole solutions were eventually seen to be almost inevitable. Oppenheimer & Volkoff (1939) predicted that neutron stars above a certain mass should inevitably collapse into black holes. Later, Penrose (1965) and Hawking & Penrose (1970) showed that generic initial conditions in GR produce singularities, seemingly always shadowed by a black hole event

horizon. But while the classical picture of a black hole may be as an object defined by a dark event horizon,[1] the physical reality of these objects was first realized in astronomical sources that emit tremendous amounts of light.

Schmidt (1963) identified the optical counterpart to the radio quasar 3C 273, and showed that its emission originated at a cosmological redshift $z = 0.158$. At this cosmological distance, its bolometric luminosity was computed to be $\sim 10^{45}$ erg s^{-1}, orders of magnitude higher than the total stellar luminosity of a galaxy like the Milky Way. The extreme luminosities of quasars and other Active Galactic Nuclei (AGN) are generated by liberating a large fraction of the gravitational potential energy of material falling onto a central compact object with a mass in the range $10^6\ M_\odot \lesssim M \lesssim 10^{10}\ M_\odot$: a supermassive black hole (SMBH: Salpeter, 1964; Lynden-Bell, 1969; Rees, 1984). Supermassive black holes are now thought to exist at the center of nearly every galaxy, and to play a key role in the evolution of galaxies themselves (King, 2003; Kormendy & Ho, 2013; King & Pounds, 2015).

Evidently, black holes are common in the universe. Aside from the SMBH that live at the hearts of most galaxies, observations of bright X-ray emission from binaries like Cygnus X-1 (Webster & Murdin, 1972) suggest that each galaxy may host tens to hundreds of millions of stellar mass black holes with masses $M \lesssim 100\,M_\odot$ (Remillard & McClintock, 2006). Until 2015, all black holes were identified by the electromagnetic radiation from their accreting material, but the LIGO Scientific Collaboration, et al. (2016) detection of gravitational waves from two merging stellar mass black holes has opened up a new window on these objects, using "dark" ripples in spacetime itself. The latest results from LIGO (LIGO Scientific Collaboration, et al., 2018) indicate that stellar mass black holes merge frequently in the universe, with an event rate $\sim$100,000 yr^{-1}.

[1]When quantum effects are added to GR, even this dark surface starts to emit a very faint light (Hawking, 1974).

0.2 Sgr A*, low-luminosity AGN, and hot accretion flows

Closer to home than the distant quasars or merging stellar-mass black holes observed by LIGO, the Galactic Center radio source Sagittarius A* (Sgr A*: Balick & Brown, 1974) is now known to host a relatively small SMBH of mass $M = (4.1 \pm 0.03) \times 10^6\, M_\odot$ (GRAVITY Collaboration et al., 2018a). In contrast to the bright quasars, Sgr A* has a low luminosity $\sim 10^{35}$ erg s^{-1}, or $\sim 100\, L_\odot$. As a result, Sgr A* is classified as an (extremely) Low-Luminosity AGN (LLAGN). Most SMBH in galaxies throughout the local universe are in a similar low-luminosity state (Greene & Ho, 2007; Ho, 2008). While wafer-thin accretion disks power quasars and bright AGN (Shakura & Sunyaev, 1973; Novikov & Thorne, 1973), LLAGN are fed by Advection-Dominated Accretion Flows (ADAFs: Ichimaru 1977; Rees et al. 1982; Narayan & Yi 1994, 1995a,b; Abramowicz et al. 1995; Blandford & Begelman 1999). In these systems (also called Radiatively Inefficient Accretion Flows, or RIAFs), most of the gravitational potential energy liberated by the infall of gas is lost by advection across the black hole event horizon instead of being released in light, leaving the accretion flow relatively dark.

Accretion systems are described primarily by their accretion rate $\dot{M}$. Accounting for the large range in the masses of the central objects, $\dot{M}$ is usually expressed in dimensionless form as a fraction of the Eddington accretion rate $\dot{M}_{\rm Edd}$. An object accreting at the Eddington rate, with infalling matter emitting with an efficiency η (usually $\eta \approx 0.1$), will shine at the Eddington luminosity where the radiation pressure on ions is in equilibrium with gravity:

$$L_{\rm Edd} = \frac{4\pi\, GMc\, m_{\rm p}}{\sigma_{\rm T}} = 1.3 \times 10^{47} \left(\frac{M}{10^9\, M_\odot}\right) \,{\rm erg\,s^{-1}}, \tag{0.2}$$

$$\dot{M}_{\rm Edd} = \frac{L_{\rm Edd}}{\eta c^2} = \frac{2.2}{\eta} \left(\frac{M}{10^9\, M_\odot}\right) M_\odot\,{\rm yr^{-1}}, \tag{0.3}$$

where $m_{\rm p}$ is the proton mass and $\sigma_{\rm T}$ is the Thomson scattering cross-section. Thin disks in AGN usually have an accretion rate only a little below Eddington, $10^{-3}\dot{M}_{\rm Edd} \lesssim \dot{M} \lesssim \dot{M}_{\rm Edd}$, while ADAFs have $\dot{M} \lesssim 10^{-3}\dot{M}_{\rm Edd}$. Notably, Sgr A* has a measured accretion rate of $\dot{M} \sim 10^{-7}\,\dot{M}_{\rm Edd}$ (Agol, 2000; Bower et al., 2003; Marrone et al., 2007).

The three key properties of low-accretion rate ADAF models are that they are (1) hot, (2) diffuse and geometrically thick, and (3) optically thin (see Yuan & Narayan 2014 for a review). In ADAF models, the ion temperature of infalling plasma at radius r is only slightly less than the virial temperature:

$$T_{\rm i} \sim \frac{m_{\rm p}\, c^2}{3k_{\rm B}} \frac{r}{r_{\rm g}} \sim 10^{12} \frac{r}{r_{\rm g}}\,{\rm K}, \tag{0.4}$$

where $k_{\rm B}$ is Boltzmann's constant. As a result of these high temperatures, pressure support makes the gas in the disk "puff up" to the point where the height-to-radius ratio $h/r \sim 0.5$. Consequently, the plasma density becomes quite low ($n \lesssim 10^6$ cm^{-3} in Sgr A*; Genzel et al. 2010). As a result of these low densities, the optical depth $\tau \ll 1$, so radiation never equilibrates into a blackbody spectrum as it does locally in thin disks. Instead, the spectra from ADAFs are dominated by individual radiative processes, most notably synchrotron emission. The magnetic fields that source the observed synchrotron radiation are also critical in enabling accretion in the first place. Ordinary molecular viscosity in these diffuse, nearly-collisionless systems is too weak to transport angular momentum out of the accreting plasma. Instead, turbulence generated by the magnetorotational instability (MRI: Balbus & Hawley, 1998; de Villiers et al., 2003; Narayan et al., 2012), is thought

to act as an effective viscosity and enable accretion.

Semi-analytic ADAF models emitting via thermal (or nonthermal) synchrotron radiation can describe the bulk of the dim emission from Sgr A* in the 'submm bump' around 10^{12} Hz (e.g., Narayan et al., 1995, 1998; Mahadevan, 1998; Yuan et al., 2002, 2003). To make detailed comparisons with Sgr A*'s spectrum and rapid variability requires simulating the Sgr A*accretion flow with General Relativistic Magnetohydrodynamic (GRMHD) simulations (e.g., Mościbrodzka et al., 2009; Dexter et al., 2010; Shcherbakov et al., 2012; Mościbrodzka et al., 2014; Chan et al., 2015a; Ressler et al., 2017; Chael et al., 2018a). These simulations are complicated by the fact that the low densities and high temperatures in ADAFs prevent ions and electrons from coming into thermal equilibrium with each other. Chapters 1 and 2 address this problem, and describe radiative GRMHD simulations of Sgr A* that self-consistently simulate electron and ion populations that are not in mutual equilibrium.

When even the electron-electron collision time-scale becomes long in hot flows (Mahadevan & Quataert, 1997), nonthermal electrons accelerated by plasma processes like shocks (e.g., Guo et al., 2014) and magnetic reconnection (e.g., Sironi & Spitkovsky, 2011, 2014) can persist in the accretion flow. These rapidly injected nonthermal electrons may generate the observed near-infrared (NIR; Genzel et al., 2003), and X-ray flares (Neilsen et al., 2013; Zhang et al., 2017) seen from Sgr A*. The first spatially resolved observations of NIR flares by the GRAVITY interferometer revealed circular motion (GRAVITY Collaboration et al., 2018b), indicating that the compact flares may originate in "hot spots" of plasma near the black hole (Broderick & Loeb, 2006). The first steps toward simulating these nonthermal particles self-consistently is the focus of Chapter 4.

0.3 M87, extragalactic jets, and magnetically arrested disks

Only a few years after Schwarzschild, Curtis (1918) detected a linear feature emanating from the giant elliptical galaxy Messier 87 (M87). Only much later was this observation appreciated as the first detection of an astrophysical "jet" (Baade & Minkowski, 1954). The M87 jet extends outside its host galaxy, up to 65 kpc, and it is pointed nearly directly at the Earth (the inclination angle is $\approx$17°; Mertens et al. 2016).

Synchrotron emission from extragalactic jets can dominate LLAGN spectra (Blandford & Königl, 1979; Falcke & Biermann, 1995). Very-long-baseline interferometry (VLBI) observations at radio and millimeter wavelengths (e.g., Palmer et al., 1967; Reid et al., 1982; Junor et al., 1999; Kovalev et al., 2007; Ly et al., 2007; Asada & Nakamura, 2012; Hada et al., 2016; Walker et al., 2018; Kim et al., 2018) show that the M87 jet remains collimated on sub-parsec scales into the heart of the galaxy, where it terminates in a bright radio core.

At frequencies higher than ~86 GHz, Hada et al. (2011) showed that the radio core in M87 is coincident with a supermassive black hole. The SMBH mass has been measured to be $6.2 \times 10^9 M_\odot$ from stellar dynamics in the surrounding galaxy nucleus (Gebhardt et al. 2011; assuming a distance of $D = 16.9$ Mpc; Mei et al. 2007), but the mass measured from gas dynamics is a factor of two smaller (Walsh et al., 2013). Like Sgr A* the M87 accretion flow is hot and dim. While Sgr A* apparently lacks jets in radio images (unless they are launched directly toward Earth; Issaoun et al. 2019), jets are launched from many AGN (see e.g., Bridle & Perley, 1984; Blandford et al., 2019, for reviews).

The jet in M87 carries a large kinetic energy in the range $P \sim 10^{42} - 10^{45}$ erg s^{-1} (Reynolds et al., 1996; Owen et al., 2000; Stawarz et al., 2006; de Gasperin et al., 2012). This power may be extracted from the black hole spin itself by magnetic fields threading the event horizon (Blandford

& Znajek, 1977; Tchekhovskoy et al., 2011). In thin disk models, the field lines extract rotational energy from the black hole at a rate proportional to the square of both the dimensionless black hole spin $a = Jc/GM^2$ and the magnetic flux on the event horizon $\Phi_{\rm BH}$. This Blandford-Znajek mechanism produces a jet power (Tchekhovskoy et al., 2010)

$$P_{\rm BZ} \approx 2.8\, a^2 \left(\frac{\Phi_{\rm BH}}{50\sqrt{\dot{M}c}\, r_{\rm g}}\right)^2 \dot{M}c^2$$
$$\approx 3.6 \times 10^{42}\, a^2 \left(\frac{\Phi_{\rm BH}}{50\sqrt{\dot{M}c}\, r_{\rm g}}\right)^2 \left(\frac{\dot{M}}{10^{-6}\dot{M}_{\rm Edd}}\right) \left(\frac{M}{10^9\, M_\odot}\right) {\rm erg\,s^{-1}}, \qquad (0.5)$$

where $\Phi_{\rm BH}/\sqrt{\dot{M}c}\, r_{\rm g}$ is the dimensionless magnetic flux on the horizon.

Simulations of thick ADAFs that launch relativistic jets show that they obey the same scaling relation with spin and flux as Equation 0.5, but with a smaller prefactor (e.g., Sądowski et al., 2013b; The Event Horizon Telescope Collaboration et al., 2019e). Strong jets like that from M87 are most easily launched by Magnetically Arrested Disks (MADs: Bisnovatyi-Kogan & Ruzmaikin, 1976; Narayan et al., 2003; Igumenshchev et al., 2003). MADs represent the opposite limit to the weak magnetic flux mode of black hole accretion (Standard and Normal Evolution, or SANE; Narayan et al. 2012). In MADs, coherent magnetic flux builds up on the black hole, saturating at a dimensionless flux on the horizon $\Phi_{\rm BH}/\sqrt{\dot{M}c}\, r_{\rm g} \approx 50$. This buildup of flux limits accretion via magnetic pressure. In GRMHD simulations (e.g., Igumenshchev et al., 2003; Tchekhovskoy et al., 2011; McKinney et al., 2012; Sądowski et al., 2013b), MADs launch jets powered by the black hole spin with wide opening angles and large jet powers. The large jet power and wide opening angle observed in the M87 jet ($\sim$55° at 43 GHz ; Walker et al., 2018) suggest that M87 may have a magnetically arrested disk at its core. Chapter 3 explores MAD simulations of the M87 disk and jet in radiative GRMHD simulations with self-consistent two-temperature evolution.

0.4 The black hole shadow and the Event Horizon Telescope

Until this year (2019), all black holes were unresolved. To test models of the accretion flow or jet-launching region, one had to rely primarily on the integrated emission from the accretion flow. Resolving an image of the near-horizon region within a few $r_{\rm g}$ of the event horizon was impossible, due to the small angular scales involved. Prior to 2019, the Event Horizon Telescope (EHT), a global millimeter VLBI array (The Event Horizon Telescope Collaboration et al. 2019b, hereafter Paper II), constrained the compact structure in the cores of Sgr A* (Doeleman et al., 2008) and M87 (Doeleman et al., 2012; Akiyama et al., 2015) to the scale of a few Schwarzschild radii, but these early observations could not produce an image of this horizon-scale structure.

The geodesics of the Schwarzschild metric (Equation 0.1) indicate that photons travel on circular orbits close to the black hole at $r_{\rm p} = 3r_{\rm g}$. The time (t) and azimuthal (ϕ) symmetries of the Schwarzschild metric imply that the geodesics of a particle with four-velocity u^μ has two conserved quantities: $E = -u_t$, the energy at infinity, and $L = u_\phi$, the angular momentum at infinity (Carroll, 2004). Thus, working in units where $c = 1$, a photon on a circular orbit has a covariant four-velocity $u_\mu = (-E, 0, 0, L)_\mu$.

Photon geodesics are null: $u^\mu u_\mu = 0$. Expanding the null condition at the equatorial circular photon orbit ($\theta = \pi/2$, $r = 3r_{\rm g}$) gives:

$$\begin{aligned} g^{tt}\,(u_t)^2 + g^{\phi\phi}\,(u_\phi)^2 &= 0, \\ -\left(1 - 2\,\frac{r_{\rm g}}{r}\right)^{-1} E^2 + \frac{1}{r^2 \sin^2\theta} L^2 &= 0, \\ L^2 &= 27\, r_{\rm g}^2 E^2. \end{aligned} \tag{0.6}$$

Thus, at the last photon orbit, photons have a specific angular momentum $\ell = L/E = \sqrt{27}\, r_{\rm g}$. For

a particle traveling at c at a large distance D from the black hole, the specific angular momentum ℓ corresponds to the impact parameter b. Thus, in the Schwarzschild metric, photons with an impact parameter $b_c = \sqrt{27}\, r_g$ are captured on the unstable circular photon orbit. Photons with an impact parameter less than b_c eventually plunge into the black hole (Hilbert, 1917). If the black hole were illuminated from behind from the perspective of an observer at a large distance D, the black hole would therefore have a silhouette with an apparent angular diameter θ_{BH}:

$$\theta_{BH} = \frac{2\sqrt{27}\, r_g}{D} = 37.6 \left(\frac{M}{6.2 \times 10^9 M_\odot} \right) \left(\frac{D}{16.9\,\mathrm{Mpc}} \right)^{-1} \mu\mathrm{as}, \tag{0.7}$$

where the values in Equation 0.7 are scaled relative to the values for the black hole in M87.[2]

Outside θ_{BH}, light reaches the observer. Inside is darkness; a "shadow" cast by the black hole on the surrounding emission. In the Kerr (1963) metric, which describes black holes with angular momentum, b_c changes with a ray's orientation relative to the angular momentum vector, so the black hole no longer appears circular (Bardeen et al., 1972). However, this change is small, $\lesssim$4% over the full range of black hole spin and viewing inclination (Chandrasekhar, 1983; Takahashi, 2004).

The shadow should be visible for astrophysical black holes surrounded by optically thin emission from an accretion disk or jet (Luminet, 1979; Falcke et al., 2000). In 2017, the EHT observed M87 at 230 GHz with a full array of eight telescopes at six distinct geographic sites, with a nominal resolution of $\approx$25 μas (The Event Horizon Telescope Collaboration et al. 2019c, hereafter Paper III). These observations were synthesized into an image (The Event Horizon Telescope Collaboration et al. 2019d, hereafter Paper IV), using cutting-edge image reconstruction methods described in Chapters 5–7, to produce the first image of the black hole shadow in M87. The emission in these first

[2] A microarcescond, 1 μas $= 2.8 \times 10^{-10}$ degrees, is approximately the angular size of a grain of rice on the surface of the moon as viewed from Earth.

EHT images of M87 is consistent with the physical predictions from a wide range of simulations of the synchrotron emission from an accretion flow and jet a few Schwarzchild radii away from the black hole event horizon (The Event Horizon Telescope Collaboration et al. 2019e, hereafter Paper V), including the simulations presented in Chapter 3 of this work. The measured shadow diameter of $\sim$40 μas was used to determine the black hole mass as $(6.5 \pm 0.7) \times 10^9\, M_\odot$ (The Event Horizon Telescope Collaboration et al. 2019f, hereafter Paper VI), consistent with the Gebhardt et al. (2011) results from stellar dynamics. Figure 0.1 presents an image of the supermassive black hole in M87 from EHT observations reconstructed with the `eht-imaging` library (Chapter 6), as well as an image from a radiative, two-temperature GRMHD simulation (Chapter 4), reconstructed with the same technique. The EHT 2017 image is dominated by a ring of emission at the radius of the lensed photon orbit of the supermassive black hole in M87 (The Event Horizon Telescope Collaboration et al., 2019a).

The EHT also observed Sgr A* extensively during 2017. Sgr A* should have an even larger shadow (with a mean diameter of 48 μas depending on the inclination and spin), but producing an image of Sgr A*'s shadow from EHT data is complicated by its rapid variability (e.g., Yusef-Zadeh et al., 2009) and interstellar scattering along the line of sight (e.g., Johnson et al., 2018). Imaging of Sgr A* using 2017 EHT data is currently ongoing.

0.5 Simulating, imaging, outlook

The realization of the decades-long drive to produce an image of the black hole in M87 would not have been possible without a confluence of experts from many fields: in the advanced engineering and hardware development that made millimeter VLBI possible at all 8 EHT telescopes, in the analysis and imaging of difficult millimeter VLBI data, and in the theoretical understanding of

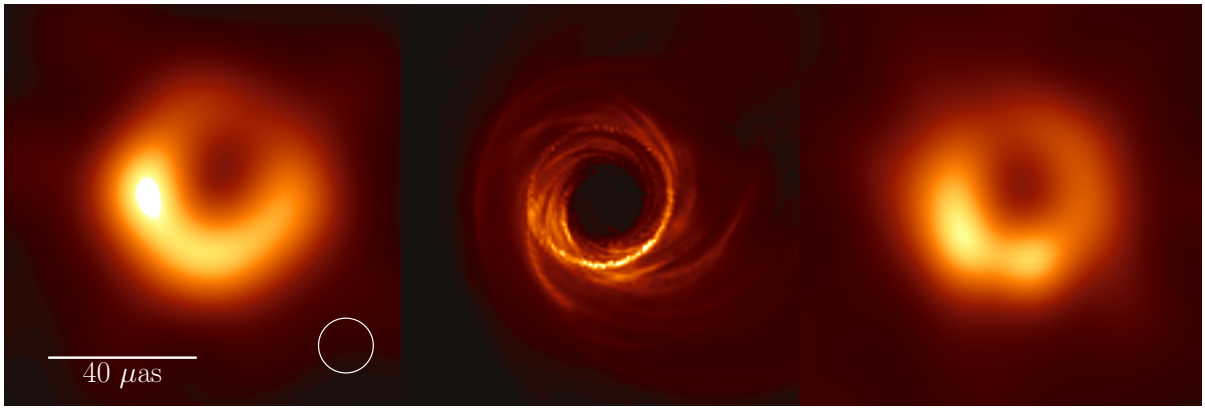


Figure 0.1: (Left) the black hole shadow of M87 as resolved by the EHT on April 6, 2017. This image was reconstructed with the `eht-imaging` library (Chapter 6) and has been convolved with a 15 μas FWHM Gaussian kernel (Paper IV). (Center) a simulated image of the M87 black hole from a radiative, two-temperature GRMHD simulation (Chapter 3). (Right) a reconstruction of simulated data from the model image on EHT 2017 baselines, using the same fiducial `eht-imaging` script used on the real data, and convolved with the same 15 μas kernel.

supermassive black hole accretion flows and jets. In my time with the global EHT collaboration, I have established a unique position linking the last two of these domains. These two threads of research – performing accretion simulations to investigate the role of particle heating and acceleration in the plasmas around M87 and Sgr A* and designing new algorithms for reconstructing images from EHT data – have never seemed less than essentially intertwined in my mind. Interpreting EHT images in the context of LLAGN accretion models and simulations requires understanding the processes that contribute to the emission that we see; understanding how different models and theories can be tested with EHT data requires understanding how images are formed from these challenging data and what features in these images are robust and which are more suspect.

This dissertation presents the most significant projects in these two strands of my research over the past five years. Part I focuses on simulations of the accretion flows and jets around Sgr A* and M87. In Chapter 1, I summarize the method used in the code `KORAL` to self-consistently solve for the temperature of the emitting electrons in a simulation given a prescription for the underlying plasma microphysics. Chapter 2 applies this method to four two-temperature simulations of Sgr A* using two different mechanisms for electron heating and compares the resulting images and spectra

with VLBI observations and early EHT data. Chapter 3 repeats this exercise for the first MAD simulations applied to M87. It finds (before any EHT 2017 data are considered) that MAD models are natural fits to explain the large jet power and wide jet opening angle observed in VLBI images of M87 at frequencies $\leq$86 GHz. In Chapter 4, I detail the algorithm for evolving nonthermal electron populations I developed in `KORAL`, and discuss initial tests of this method. This method will soon enable self-consistent simulations of the nonthermal particle injection and evolution that may source flares and rapid variability in Sgr A*.

Part II focuses on the imaging techniques and software I developed and that were critical in producing the first image of M87 with the EHT. In Chapter 5, I discuss the principles of interferometry and imaging with Regularized Maximum Likelihood (RML) techniques. Chapter 6 presents the structure of the `eht-imaging` software library I developed to implement these new imaging methods. Finally, in Chapter 7 I present the results of my work contributing to the global team effort to produce the first image of a black hole (The Event Horizon Telescope Collaboration et al., 2019d).

The synthesis between advanced EHT imaging and accretion flow simulations has a bright future in the emerging field of horizon-scale astrophysics. In the coming years, techniques and methods developed in this thesis will be used as part of the community-wide effort to link EHT observations with information from across the electromagnetic spectrum. These new data, images, and simulations will help us explore and better understand the dynamics and energetics of material around supermassive black holes, the conditions which set how these black holes launch jets and outflows to galactic scales, and what new avenues will become possible, using images of the black hole shadow, to test the nature of spacetime close to the black hole's event horizon.

Page intentionally left blank

Part I

Simulations

Page intentionally left blank

Text in this chapter was previously published in *MNRAS* 470 (2017), 2, pp 2367-2386 (A. Chael, R. Narayan, and A. Sądowski), *MNRAS* 478 (2018), 4, pp 5209-5229 (A. Chael, M. Rowan, R. Narayan, M. Johnson, and L. Sironi), and in *MNRAS* 486 (2019), 2, pp 2873-2895 (A. Chael, R. Narayan, and M. Johnson).

1

Black hole accretion simulations with radiation and thermodynamics

General relativistic magnetohydrodynamic (GRMHD) simulations evolve an ionized, magnetized plasma in a fixed background metric $g_{\mu\nu}$. ADAF models for LLAGN accretion flows are optically thin and geometrically thick, making them particularly attractive targets for these grid-based fluid simulations. Multiple GRMHD codes have been developed to evolve the plasma flow around black holes in the Kerr metric (e.g., Komissarov, 1999; de Villiers et al., 2003; Gammie et al., 2003; Tchekhovskoy et al., 2010; Sądowski et al., 2013a, 2014; McKinney et al., 2014; Ryan et al., 2015; White et al., 2016; Chandra et al., 2017). GRMHD codes have been successfully used to demonstrate that viscosity in these collisionless systems can arise from turbulence generated by the magnetorotational instability (MRI: Balbus & Hawley, 1998; de Villiers et al., 2003; Narayan et al., 2012) and to investigate the launching of jets powered by the black hole spin (e.g., McKinney, 2006; Komissarov et al., 2007; McKinney & Blandford, 2009; Tchekhovskoy et al., 2011). They have also been used to simulate emission and compare theory to observations in low accretion rate systems like Sgr A* (e.g., Mościbrodzka et al., 2009; Dexter et al., 2010; Mościbrodzka & Falcke, 2013; Mościbrodzka et al., 2014; Chan et al., 2015a,b) and M87 (e.g., Dexter et al., 2012; Mościbrodzka

et al., 2016a, 2017).

The GRMHD equations on their own include no feedback from radiation on the dynamics or energetics of the plasma flow. Since ADAF flows with accretion rates $\dot{M} \lesssim 10^{-6}\dot{M}_{\rm Edd}$ are optically thin, radiative feedback is likely dynamically unimportant and pure GRMHD codes are a good choice for studying the dynamics of these systems. At intermediate accretion rates $10^{-6}\dot{M}_{\rm Edd} \lesssim \dot{M} \lesssim 10^{-3}\dot{M}_{\rm Edd}$, radiative cooling affects the gas temperature, and at high accretion rates $\dot{M} \gtrsim 10^{-3}\dot{M}_{\rm Edd}$, efficient cooling collapses the disk into an optically thick and geometrically thin state described by the Shakura & Sunyaev (1973) solution.[1]

Without modifications to account for radiative cooling, GRMHD simulations are ill-suited to explore disks in any state but the most optically thin ADAFs. While some studies have explored higher accretion rate regimes by adding artificial cooling functions to the gas (e.g., Shafee et al., 2008; Penna et al., 2010; Noble et al., 2011), "radiation GRMHD" or "GRRMHD" codes (e.g., Farris et al., 2008; Roedig et al., 2012; Sądowski et al., 2013a, 2014; Ryan et al., 2015) allow for a more self-consistent treatment of the physical interactions between gas and radiation at all accretion rates, and they show particular promise for studying disks that transition between different accretion rates and corresponding spectral states (e.g., in X-ray binaries Esin et al., 1997). Of particular note for this work is the fact that while Sgr A* has an accretion rate $\lesssim 10^{-7}\dot{M}_{\rm Edd}$ (Marrone et al., 2007), M87 accretes more efficiently and radiative feedback may begin to be important in setting the gas temperature (Ryan et al., 2018).

While the radiation field may be *dynamically* unimportant in low accretion rate ADAFs, radiative cooling is nevertheless important in determining the observable properties of these objects. In these hot, diffuse flows, Coulomb coupling between electrons and ions is inefficient (Mahadevan, 1998).

[1]At the highest Eddington or super-Eddington accretion rates, photons become trapped in an extremely optically thick disk. In these "slim-disk" systems (Abramowicz et al., 1988), most of the photons are advected into the black hole, making these systems again radiatively inefficient, although radiation is dynamically important.

Electrons and ions will have different temperatures, with the ratio set by the balance between the viscous heating rates of the two species, the rate of energy transfer from ions to electrons by Coulomb coupling, and the rate of radiative cooling of the electrons. Thus, to obtain spectra and images from GRMHD simulations for comparison with data, it is usually necessary to impose ad hoc assumptions about the electron energy distributions in post-processing. Often, the electron-to-gas temperature ratio is set at a constant value throughout the simulation. (e.g., Mościbrodzka et al., 2009; Dexter et al., 2010), or the simulation is manually divided into jet and disk regions with different temperatures (e.g., Mościbrodzka & Falcke, 2013; Mościbrodzka et al., 2014; Chan et al., 2015a,b). In this approach, the electron temperature is adjusted after the fact to find the best fit to the measured spectrum, with no input from the physics of the underlying electron-ion thermodynamics.

Another approach to simulating these sources is to *self-consistently* evolve populations of ions and electrons, each with its own thermodynamics and interactions with each other and the radiation field. This approach has been pursued with several different codes (Ressler et al., 2015; Sądowski et al., 2017; Ryan et al., 2017), all of which evolve a thermal electron population alongside the other GRMHD or GRRMHD fluid variables. In the GRRMHD code `KORAL` in particular, both ion and electron populations are evolved simultaneously along with the total gas and radiation in a GRRMHD simulation; the particles gain and lose energy from adiabatic compression/expansion, Coulomb coupling, and radiative cooling. In this way, the electron temperature is obtained 'on the fly' during the simulation and is not modeled during post-processing. `KORAL` and other two-temperature GRMHD methods have been used successfully to produce spectra and images from several recent simulations of Sgr A* (Ressler et al., 2017; Chael et al., 2018a) and of M87 (Ryan et al., 2018; Chael et al., 2019b).

While the physics of electron cooling is well understood and relatively straightforward to incor-

porate in these simulations, the full range of physical processes that govern dissipation and particle heating at the smallest scales in hot accretion flows is still unconstrained. The dissipation occurs on scales much smaller than the finest resolved by the grid. At current resolutions, while the total amount of viscous dissipation in a cell can be estimated numerically, the *fraction* of the dissipated energy that goes into the electrons or the ions must be determined by a sub-grid prescription. Most works in this field (Ressler et al., 2015; Sądowski et al., 2017; Ressler et al., 2017; Ryan et al., 2017) have used a single heating prescription based on the Landau-damping of weakly collisional MHD turbulence (Howes, 2010). The simulations of Sgr A* and M87 in the following chapters (2–3) investigate the importance of this choice by comparing results with a new prescription derived motivated by the assumption that magnetic reconnection is the process that truncates the turbulent cascade in these plasmas (Rowan et al., 2017).

This chapter reviews the full method for two-temperature particle evolution implemented in KORAL (Sądowski et al., 2013a, 2014, 2017). Except for some small modifications, this method is the same as that originally presented in Sądowski et al. (2017). Chapters 2 and 3 apply this method to simulations of Sgr A* and M87, and Chapter 4 extends it to evolve nonthermal, high-energy electrons in parallel with the thermal species. The GRMHD equations without radiation are introduced in Section 1.1, and radiative feedback with M1 closure is discussed in Section 1.2. Section 1.3 presents KORAL's method for evolving thermal electron and ion populations and handling their interactions. Finally, Section 1.4 introduces the plasma physics underlying the two prescriptions for electron heating (Landau damping and magnetic reconnection) that are investigated in the following chapters.

1.1 GRMHD equations

In an ionized plasma (assumed for the reminder of this work to be pure Hydrogen), charge neutrality demands that electrons and ions have the same number density n and four velocity u^μ everywhere. However, without efficient processes to bring them into equilibrium, the species can in principle have distinct local thermal energy densities $u_\mathrm{e} \neq u_\mathrm{i}$ and temperatures $T_\mathrm{e} \neq T_\mathrm{i}$. Standard single-fluid GRMHD simulations ignore the distinctions between the individual species, and treat electrons and ions together as a single fluid. This fluid is characterized by a mass density dominated by the ions, $\rho = m_\mathrm{p} n$, and a total internal energy density $u = u_\mathrm{e} + u_\mathrm{i}$.

The total pressure $p = p_\mathrm{e} + p_\mathrm{i}$ is related to the energy density u by the ideal gas law with an effective adiabatic index Γ_gas:

$$p = (\Gamma_\mathrm{gas} - 1)u. \tag{1.1}$$

In single-fluid GRMHD simulations, the adiabatic index Γ_gas is typically fixed at the non-relativistic, monatomic value $\Gamma_\mathrm{gas} = 5/3$. In a hot accretion flow, temperatures frequently reach values $> 10^{10}$ K in the inner regions (Yuan & Narayan, 2014). At these temperatures, electrons become relativistic, decreasing their adiabatic index from 5/3 towards 4/3. Thus, the effective gas adiabatic index Γ_gas takes on values in the range $4/3 \leq \Gamma_\mathrm{gas} \leq 5/3$ depending on the local temperatures and energy densities of the two component species. `KORAL` treats this variable adiabatic index self-consistently (Section 1.3), but this effect can also be approximated in other codes by fixing an intermediate value everywhere (e.g., $\Gamma_\mathrm{gas} = 13/9$; Ryan et al., 2018).

Under the ideal MHD assumption of infinite conductivity, the dual Faraday tensor $F^{*\mu\nu}$ can be completely described by a magnetic field four-vector b^μ:

$$F^{*\mu\nu} = b^\mu u^\nu - b^\nu u^\mu. \tag{1.2}$$

KORAL uses Heaviside-Lorentz units, so that the rest-frame magnetic field strength in Gauss is $|B| = \sqrt{4\pi b^\mu b_\mu}$.[2]

The MHD stress-energy tensor $T^\mu{}_\nu$ consists of contributions from the fluid variables (ρ, u, p, u^μ) as well as the magnetic field four-vector b^μ (Gammie et al., 2003):

$$T^\mu{}_\nu = \left(\rho + u + p + b^2\right) u^\mu u_\nu + \left(p + \frac{1}{2}b^2\right) \delta^\mu{}_\nu - b^\mu b_\nu. \tag{1.3}$$

In standard GRMHD simulations, the fluid is evolved using the conservation of the matter current ρu^μ and the stress-energy $T^\mu{}_\nu$, combined with the ideal MHD induction equation for b^μ (Gammie et al., 2003):

$$(\rho u^\mu)_{;\mu} = 0, \tag{1.4}$$

$$T^\mu{}_{\nu;\mu} = 0, \tag{1.5}$$

$$F^{*\mu}{}_{\nu;\mu} = 0. \tag{1.6}$$

GRMHD codes like KORAL typically use a second-order Runge-Kutta scheme to advance the fluid quantities and magnetic field at each time step. The advection of quantities across cells is handled explicitly by reconstructing Lax-Friedrichs fluxes at the cell walls using the Piecewise Parabolic Method (or a first-order flux-limited method). Geometrical terms (i.e., the covariant derivative terms involving Christoffel symbols in Equations 1.4–1.6) are added as source terms at cell centers. The full explicit advective algorithm used in KORAL is described in Sądowski et al. (2013a, 2014).

[2]This chapter and the KORAL code use units where $c = 1$.

1.2 Radiation GRMHD

KORAL incorporates radiation feedback by treating the frequency-integrated radiation field $R^{\mu}{}_{\nu}$ as a massless perfect fluid evolved in parallel to $T^{\mu}{}_{\nu}$. Under the M1 closure scheme (Levermore, 1984; Sądowski et al., 2013a, 2014; McKinney et al., 2014), the bolometric radiation field is fixed by its rest frame energy density $\bar{E}$ and a timelike four velocity $u^{\mu}_{\rm R} \neq u^{\mu}$ of the frame where the radiation is isotropic:

$$R^{\mu}{}_{\nu} = \frac{4}{3}\bar{E}u^{\mu}_{\rm R}u_{R\,\nu} + \frac{1}{3}\bar{E}\delta^{\mu}{}_{\nu}. \tag{1.7}$$

In addition to the radiation energy density and frame velocity, KORAL also tracks the rest frame photon number density $\bar{n}_{\rm R}$, which encodes information about the mean photon frequency and the radiation temperature $T_{\rm R}$. Under the assumption that the radiation spectrum is a grey body (Sądowski & Narayan, 2015), the radiation temperature is

$$T_{\rm R} = \frac{\hat{E}}{2.7012\,k_B\hat{n}_{\rm R}}, \tag{1.8}$$

where $\hat{E}$ and $\hat{n}_{\rm R}$ are the radiation energy density and photon number transformed to the fluid frame. Throughout this work, quantities in the radiation rest frame are denoted with bars, and quantities in the fluid frame are denoted with hats.[3]

The set of general relativistic, radiative, magnetohydrodynamic (GRRMHD) equations for evolving the total fluid, the magnetic field, the frequency-integrated radiation field, and the photon

[3]In contrast to this frequency-integrated approach, Ryan et al. (2015, 2017, 2018) use a Monte-Carlo approach which represents the radiation field with many individual particle "superphotons" with different frequencies that are emitted and absorbed in between the fluid evolution timesteps.

number are then (Sądowski et al., 2014; Sądowski & Narayan, 2015)

$$(\rho u^{\mu})_{;\mu} = 0, \tag{1.9}$$

$$T^{\mu}{}_{\nu;\mu} = G_{\nu}, \tag{1.10}$$

$$R^{\mu}{}_{\nu;\mu} = -G_{\nu}, \tag{1.11}$$

$$F^{*\mu}{}_{\nu;\mu} = 0, \tag{1.12}$$

$$(\bar{n}_{\rm R} u^{\mu}_{R})_{;\mu} = \dot{\bar{n}}_{\rm R}. \tag{1.13}$$

G_{ν} is the the four-force density that couples the radiation and gas. In the fluid rest frame:

$$\hat{G}^{\,0} = \rho\,(\kappa_{\rm P,a}\hat{E} - 4\pi\kappa_{\rm P,e}\hat{B}) + \hat{G}^{0}_{\rm IC}, \tag{1.14}$$

$$\hat{G}^{\,i} = (\rho\kappa_{\rm R} + \rho\kappa_{\rm es})\hat{F}^{i}. \tag{1.15}$$

In these equations, the κ factors are the frequency-averaged grey opacities for the thermal radiative processes, $\hat{G}^{0}_{\rm IC}$ is the fluid-frame thermal energy loss from inverse Compton scattering (see Sądowski & Narayan 2015 for the full expression), and $\hat{B} = \sigma_{\rm SB}T_{\rm e}^{4}/\pi$ is the electron blackbody radiance ($\sigma_{\rm SB}$ is the Stefan-Boltzmann constant). $\hat{F}^{i}$ is the fluid-frame radiation momentum flux, computed from the fluid-frame radiation tensor; $\hat{F}^{i} = \hat{R}^{\,0\,i}$.

Following Mihalas & Mihalas (1984), KORAL uses the Planck-averaged mean opacities $\kappa_{\rm P,e}$ and $\kappa_{\rm P,a}$ weighted for emission and absorption in $\hat{G}^{0}$ (the energy equation 1.14), and it uses the Rosseland-averaged mean opacity $\kappa_{\rm R}$ in $\hat{G}^{i}$ (the momentum equation 1.15). The full expressions for these opacities as a function of number density and temperature for both synchrotron and free-free emission are given in Sądowski et al. (2017).

The electron scattering opacity is $\kappa_{\rm es}$; it includes a Klein-Nishina factor that lowers the scattering

cross section at high photon energies (Buchler & Yueh, 1976):

$$\rho\kappa_{\rm es} = n\sigma_{\rm T}\left[1 + \frac{T_{\rm R}}{4.5\times10^8\,{\rm K}}\right]^{0.86}\ {\rm cm}^{-1}. \tag{1.16}$$

The frame-invariant photon number source term in Equation 1.13 is

$$\hat{\dot{n}}_{\rm R} = \hat{\dot{n}}_{\rm R,\,syn} + \hat{\dot{n}}_{\rm R,\,brem} - \rho\kappa_{n,{\rm a}}\hat{n}_{\rm R}. \tag{1.17}$$

The first term in Equation 1.17 is from synchrotron emission. Electrons at all energies produce the same number of synchrotron photons (Rybicki & Lightman, 1979):

$$\hat{\dot{n}}_{\rm syn} = n_{\rm e}\left[1.46\times10^5\left(\frac{B}{1\,{\rm G}}\right)\right]. \tag{1.18}$$

The second term is the production of photons from bremsstrahlung emission, and the last term is the photon loss rate from absorption by the thermal electrons, which is expressed in terms of a number absorption opacity $\kappa_{n,{\rm a}}$ (see Sądowski & Narayan 2015; Sądowski et al. 2017 for the full expressions).

KORAL solves Equations 1.9– 1.13 using a split explicit-implicit scheme. The advection of quantities across cells is first handled explicitly. The source terms in the evolution equations which represent the coupling between matter and radiation are then applied at each cell center (Sądowski et al., 2013a, 2014, 2017) using a Newton-Raphson solver to solve the implicit coupling equations for the evolved quantities.

1.3 Electron-ion thermodynamics

In moving beyond single-fluid GRMHD to evolve electrons and ions independently, both species are assumed to share the same fluid velocity u^μ and the same number density by charge neutrality: $n_\mathrm{e} = n_\mathrm{i} = \rho/m_\mathrm{p}$. The entropy per particle of each, s_e and s_i, are the fundamental variables, and the temperatures T_e, T_i and energy densities u_e, u_i are then functions of the species entropy and number density.

A given particle species s (electrons or ions) with a given temperature T_s and number density n_s has a pressure and energy density related by the ideal gas law:

$$p_s = n_s k_\mathrm{B} T_s, \tag{1.19}$$

$$u_s = \frac{p_s}{\Gamma_s(\Theta_s) - 1}. \tag{1.20}$$

If the particles are in a in a relativistic, nondegenerate thermal Maxwell-Jüttner distribution, the adiabatic index is a function of the dimensionless temperature $\Theta_s = k_\mathrm{B} T_s / m_s c^2$ (Chandrasekhar, 1939):

$$\Gamma_s(\Theta_s) - 1 = \Theta_s \left(\frac{3K_3\left(1/\Theta_s\right) + K_1\left(1/\Theta_s\right)}{4K_2\left(1/\Theta_s\right)} - 1 \right)^{-1}, \tag{1.21}$$

where $K_m(x)$ is the modified Bessel function of order m. The classical entropy per particle s_s is then

$$s_s/k_\mathrm{B} = \frac{K_1\left(1/\Theta_s\right)}{\Theta_s K_2\left(1/\Theta_s\right)} + \ln\left[\frac{\Theta_s K_2\left(1/\Theta_s\right)}{n_s}\right] + C, \tag{1.22}$$

where C is an integration constant.

Because the exact expressions in Equations 1.21 - 1.22 involve Bessel functions and are difficult

to invert, KORAL uses approximate forms. This self-consistent approximation is based on a fitting function to the specific heat at constant volume, which can be integrated to find expressions for the internal energy density and entropy per particle (see Appendix A of Sądowski et al. 2017). The approximate equation for the internal energy density used in KORAL is:

$$u_s(\Theta_s) = n_s m_s c^2 \frac{\Theta_s}{\Gamma_s(\Theta_s) - 1} \approx n_s m_s c^2 \Theta_s \left(3 - \frac{3}{5\Theta_s} \ln\left[1 + \frac{5\Theta_s}{2}\right]\right), \tag{1.23}$$

The entropy per particle under this approximation is then

$$s_s/k_{\rm B} \approx \ln\left[\frac{\Theta_s^{3/2}(\Theta_s + 2/5)^{3/2}}{n_s}\right] + C, \tag{1.24}$$

and the arbitrary integration constant can be set to $C = 0$. Equation 1.24 is easy to invert to obtain the species dimensionless temperature Θ_s from the number density and entropy:

$$\Theta_s \approx \frac{1}{5}\left(\sqrt{1 + 25\left[n_s \exp\frac{s_s}{k_{\rm B}}\right]^{2/3}} - 1\right). \tag{1.25}$$

Unfortunately, Equation 1.23 cannot be analytically inverted to solve for Θ_s from u_s. Since this inversion is only needed a few times per timestep, KORAL uses a Newton-Raphson solver to invert Equation 1.23 when necessary. This approach differs from the original treatment in Sądowski et al. (2017) which used a simpler but inconsistent fitting function for $u_s(\Theta_s)$.

Because the separate species pressures and energy densities must add to the total gas pressure and energy density ($p_{\rm i} + p_{\rm e} = p$, $u_{\rm i} + u_{\rm e} = u$), the gas temperature and adiabatic index are

$$T_{\rm gas} = \frac{1}{2}(T_{\rm i} + T_{\rm e}) \tag{1.26}$$

$$\Gamma_{\rm gas} - 1 = \frac{(\Gamma_{\rm i} - 1)(\Gamma_{\rm e} - 1)(T_{\rm i}/T_{\rm e} + 1)}{(T_{\rm i}/T_{\rm e})(\Gamma_{\rm e} - 1) + (\Gamma_{\rm i} - 1)}. \tag{1.27}$$

The species entropies are evolved in the simulation using the first law of thermodynamics in covariant form:[4]

$$T_{\rm e}\left(n s_{\rm e} u^{\mu}\right)_{;\mu} = \delta_{\rm e} q^{\rm v} + q^{\rm C} - \hat{G}^0, \tag{1.28}$$

$$T_{\rm i}\left(n s_{\rm i} u^{\mu}\right)_{;\mu} = (1-\delta_{\rm e}) q^{\rm v} - q^{\rm C}, \tag{1.29}$$

where $q^{\rm v}$ is the dissipative heating rate, $\delta_{\rm e}$ is the fraction of the dissipative heating that goes into electrons, $q^{\rm C}$ is the energy exchange rate from ions to electrons due to Coulomb coupling (Stepney & Guilbert, 1983), and $\hat{G}^0$ is the radiative cooling rate (Equation 1.14).

In the absence of Coulomb coupling $q^{\rm C}$, radiative cooling $\hat{G}^0$, and dissipative heating $q^{\rm v}$, the species entropies are conserved and the particles heat and cool only due to adiabatic compression and expansion. Equations 1.28–1.29 can then be evolved as conservation laws for $s_{\rm e}$ and $s_{\rm i}$ under a standard finite-difference scheme. The Coulomb coupling $q^{\rm C}$ and radiative cooling rates $\hat{G}^0$ are analytic functions of the local plasma conditions and can be applied as source terms in the usual way (Sądowski et al., 2017)

The dissipative heating $q^{\rm v}$, however, arises in accretion systems at scales far smaller than the grid-scale from physical processes which may include turbulent damping, magnetic reconnection, and shock heating. Global simulations cannot resolve these processes; they can, however, identify the total dissipative heating rate $q^{\rm v}$ numerically at the grid scale.

The *total* dissipative heating $q^{\rm v}$ is identified numerically by evolving the thermal entropies adiabatically over a proper time step $\Delta\tau$. That is, by setting the right sides of Equations 1.28 and 1.29 to zero, the code obtains the adiabatically evolved entropies $s_{\rm i,\,adiab}$ and $s_{\rm e,\,adiab}$, and the cor-

[4] In the rest frame, both Equations 1.28 and 1.29 reduce to the familiar first law of thermodynamics, $T\frac{d(n\,s)}{d\tau} = q^{+} - q^{-} =$ heating rate $-$ cooling rate.

responding adiabatically evolved energy densities, $u_{\mathrm{i,\,adiab}}$ and $u_{\mathrm{e,\,adiab}}$. To estimate the dissipative heating in the total fluid, KORAL then compares the sum of the adiabatically evolved energy densities to the separately-evolved total gas energy density u:

$$q^{\mathrm{v}} = \frac{1}{\Delta\tau}\left[u - u_{\mathrm{i,\,adiab}} - u_{\mathrm{e,\,adiab}}\right]. \tag{1.30}$$

A complication of using Equations 1.28–1.30 to adiabatically evolve the species entropies and compute the dissipative heating is that physically, entropy density is not exactly conserved on a finite grid, despite the form of Equations 1.28–1.29. Solving these equations with a finite volume method, gas parcels (fluxes at cell walls) are heated and cooled individually from compression or expansion, and then their entropies are summed. Physically, however, when two gas parcels are brought together in a cell with finite extent, the total energy density should be kept constant, increasing the entropy. That is, in solving the source-free version of Equations 1.28–1.29 with finite-volume methods, the entropy will be preserved exactly, and the final energy density will thus be underestimated. As a result, the viscous heating identified by Equation 1.30 will be systematically too large.

To avoid this problem, KORAL *mixes* the entropies from neighboring cells at constant density.[5] In the explicit evolution of Equations 1.28–1.29 the code identifies the initial values of the entropy flux on each of the cell walls, as well as the mixing fractions which they contribute to the total entropy increase in the cell. It then takes these same mixing fractions and uses them to instead add up the energy densities in the boundary entropy fluxes, keeping the fluid density fixed.

After the total viscous heating rate q^{v} is computed numerically, it must be split up between ions and electrons. The fraction of the heating that goes into the electrons, δ_{e} is determined from

[5] As noted in (Sądowski et al., 2017), mixing at constant pressure may be a more consistent procedure, but it is much more computationally intensive.

sub-grid physics as a function of the local plasma parameters.

1.4 Plasma parameters and particle heating prescriptions

This work considers simulations with two different sub-grid prescriptions for $\delta_{\rm e}$, the fraction of the local dissipative energy generated by the simulation that heats the electrons. The value of $\delta_{\rm e}$ is determined by plasma processes that occur below the grid scale. Sub-grid models for these physics depend on three local plasma physics parameters: the "plasma-beta" $\beta_{\rm i}$, the magnetization $\sigma_{\rm i}$, and the temperature ratio $T_{\rm i}/T_{\rm e}$.

The plasma-beta parameter $\beta_{\rm i}$ is the ratio of the local thermal ion pressure to the magnetic pressure:

$$\beta_{\rm i} = \frac{8\pi\, n_{\rm i} k_{\rm B} T_{\rm i}}{|B|^2}. \tag{1.31}$$

In simulations of hot accretion flows, $\beta_{\rm i} \gtrsim 10$ in the midplane, but above and below the midplane $\beta_{\rm i}$ can drop to $\beta_{\rm i} \sim 1$. In the magnetically-dominated jet region close to the axis $\beta_{\rm i} < 1$.

The magnetization $\sigma_{\rm i}$ compares the magnetic energy density to the rest-mass energy density of the ion-dominated fluid:

$$\sigma_{\rm i} = \frac{|B|^2}{4\pi\, n_{\rm i} m_{\rm i} c^2}. \tag{1.32}$$

In SANE accretion flows, $\sigma_{\rm i} < 1$ everywhere except in the innermost jet region, but in MAD flows $\sigma_{\rm i}$ can exceed unity in the disk inside $\sim$5 gravitational radii. Even in weakly magnetized flows, $\sigma_{\rm i}$ is still relatively high compared to more familiar environments such as the non-relativistic solar wind ($\sigma_{\rm i} \ll 1$).

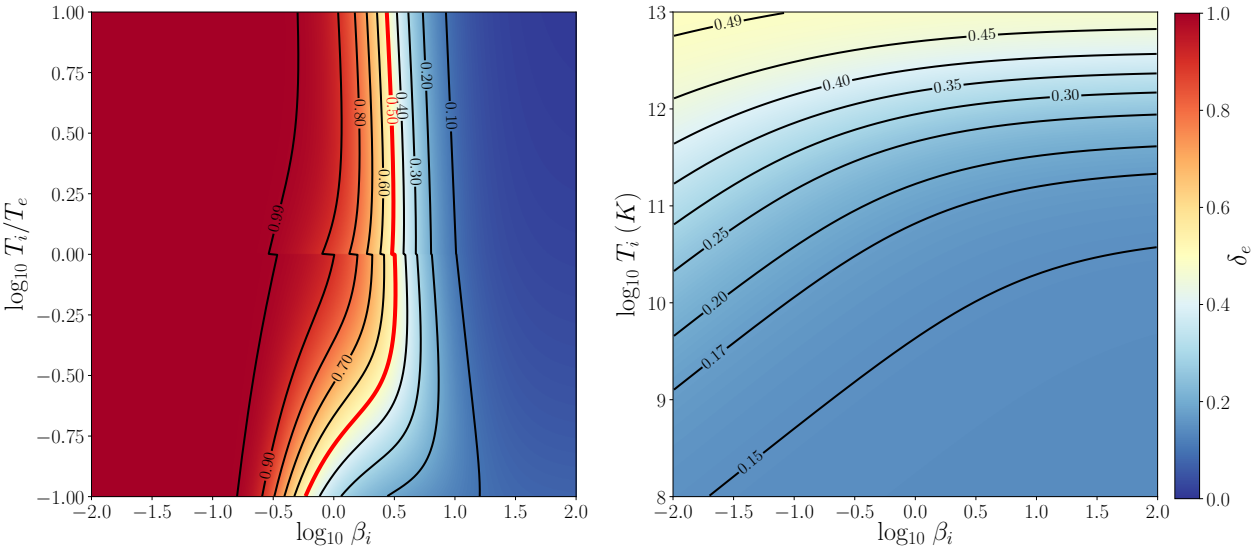


Figure 1.1: The two prescriptions for electron heating considered in this work. (Left) Turbulent cascade heating prescription (Howes, 2010). The electron heating fraction δ_e is shown as a function of the plasma-beta parameter β_i and the temperature ratio T_i/T_e. This prescription transitions rapidly from putting most of the dissipated energy into electrons at low β_i to putting most of the dissipated energy into ions at high β_i. The red contour denotes $\delta_e = 0.5$. (Right) Reconnection heating prescription (Rowan et al., 2017; Chael et al., 2018a) obtained by fitting to PIC simulation data. For clarity, this figure sets $T_i = T_e$ and plots δ_e as a function of T_i and β_i instead of the PIC simulation variables β_i and σ_w. In this prescription, δ_e never exceeds 0.5.

1.4.1 Heating from Landau-damped turbulence

The Howes (2010) prescription for δ_e is based on a model of plasma turbulence truncated by Landau damping of turbulent eddies in a weakly collisional plasma at small scales. The Howes prescription is based on nonrelativistic calculations with $\sigma_i \ll 1$ (Howes et al., 2008a,b) and while it matches solar wind measurements (Howes, 2011), it may not be well-adapted to relativistic systems like hot LLAGN accretion flows. Recently, however, Kawazura et al. (2019) performed numerical simulations of the turbulent damping process that indicate the qualitative behavior of the Howes (2010) prescription holds even in relativistic turbulent plasmas. The full Howes (2010)

fitting function is

$$\delta_{\rm e} = \frac{1}{1+f_{\rm e}}, \tag{1.33}$$

$$f_{\rm e} = c_1 \frac{c_2^2 + \beta_{\rm i}^{2-0.2\log_{10}(T_{\rm i}/T_{\rm e})}}{c_3^2 + \beta_{\rm i}^{2-0.2\log_{10}(T_{\rm i}/T_{\rm e})}} \sqrt{\frac{m_{\rm i}T_{\rm i}}{m_{\rm e}T_{\rm e}}}\ e^{-1/\beta_{\rm i}}, \tag{1.34}$$

where $c_1 = 0.92$ and $c_2 = 1.6\,T_{\rm e}/T_{\rm i}$, $c_3 = 18 + 5\log_{10}(T_{\rm i}/T_{\rm e})$ if $T_{\rm i} > T_{\rm e}$, and $c_2 = 1.2\,T_{\rm e}/T_{\rm i}$, $c_3 = 18$ if $T_{\rm i} < T_{\rm e}$.

The Howes turbulent cascade prescription is a weak function of the temperature ratio $T_{\rm e}/T_{\rm i}$ but a strong function of $\beta_{\rm i}$ (see Figure 1.1). It gives most of the turbulent heating to electrons at low $\beta_{\rm i}$, and conversely gives most of the heating to ions at high $\beta_{\rm i}$. This is a general result predicted for damped MHD turbulence (Quataert, 1998).

Since $\delta_{\rm e} \approx 1$ in regions of low $\beta_{\rm i}$, when the Howes turbulent cascade prescription is applied to accretion simulations, the resulting electron temperature will be higher in the polar region when compared to the equatorial plane. While in general radiation cools electrons to lower temperatures than ions, using this prescription can result in an electron temperature that exceeds the ion temperature in the jet region (where $\beta_{\rm i} \ll 1$).

1.4.2 Heating from magnetic reconnection

Another model for turbulent dissipation suggests that MHD turbulence may instead be truncated at small scales by magnetic reconnection (Carbone et al., 1990; Boldyrev & Loureiro, 2017; Loureiro & Boldyrev, 2017; Mallet et al., 2017; Comisso & Asenjo, 2018). Turbulent eddies become sheet-like, and they naturally fragment into plasmoids/magnetic islands via the tearing mode instability of reconnecting current sheets. One would then expect that energy dissipation in MHD turbulence is ultimately mediated by small-scale reconnection.

Rowan et al. (2017) measured electron and ion heating rates in fully kinetic particle-in-cell (PIC) simulations of trans-relativistic reconnection.[6] In these PIC simulations, the strength of the magnetic field is parametrized by the magnetization as defined with respect to the fluid enthalpy w:

$$\sigma_w = \frac{|B|^2}{4\pi\, w} = \frac{|B|^2}{4\pi\left(n_{\rm i} m_{\rm i} c^2 + \Gamma_{\rm i} u_{\rm i} + \Gamma_{\rm e} u_{\rm e}\right)}. \tag{1.35}$$

If $T_{\rm e} = T_{\rm i}$, σ_w can be expressed as

$$\sigma_w|_{T_{\rm e}=T_{\rm i}} = \frac{2}{\beta_{\rm i}\left(\Theta_{\rm i}^{-1} + \frac{\Gamma_{\rm i}}{\Gamma_{\rm i}-1} + \frac{\Gamma_{\rm e}}{\Gamma_{\rm e}-1}\right)}. \tag{1.36}$$

From Equation 1.36, it is apparent that for a given σ_w, $\beta_{\rm i}$ has a maximum value $\beta_{\rm i,max} = 1/4\sigma_w$ which is achieved in the limit when both species are highly relativistic and the thermal energy dominates the rest mass energy ($\Theta_{\rm i} \gg 1$, $\Gamma_{\rm i,e} \to 4/3$).

The magnetization σ_w represents the initial magnetic energy per electron-proton pair in units of the initial particle enthalpy. Through the reconnection of magnetic fields, some of this initial magnetic energy can be transferred from the electromagnetic field to particles, as plasma passes from the pre-reconnection 'upstream' to the post-reconnection'downstream' region. In the simulations of Rowan et al. (2017), plasma was initialized with a given temperature ratio $T_{\rm e}/T_{\rm i}$, magnetization σ_w, and plasma-beta $\beta_{\rm i}$ in the upstream region; the heating of electrons and protons was assessed by comparing the internal energy of particles in the upstream to those in the reconnection outflows.

The irreversible heating ratios $\delta_{\rm e}$ from Rowan et al. (2017) can be fit using a simple functional

[6] See also Werner et al. (2018) for a similar study. Rowan et al. (2019) studies the case where the magnetic field perpendicular to the reconnecting field lines is non-zero.

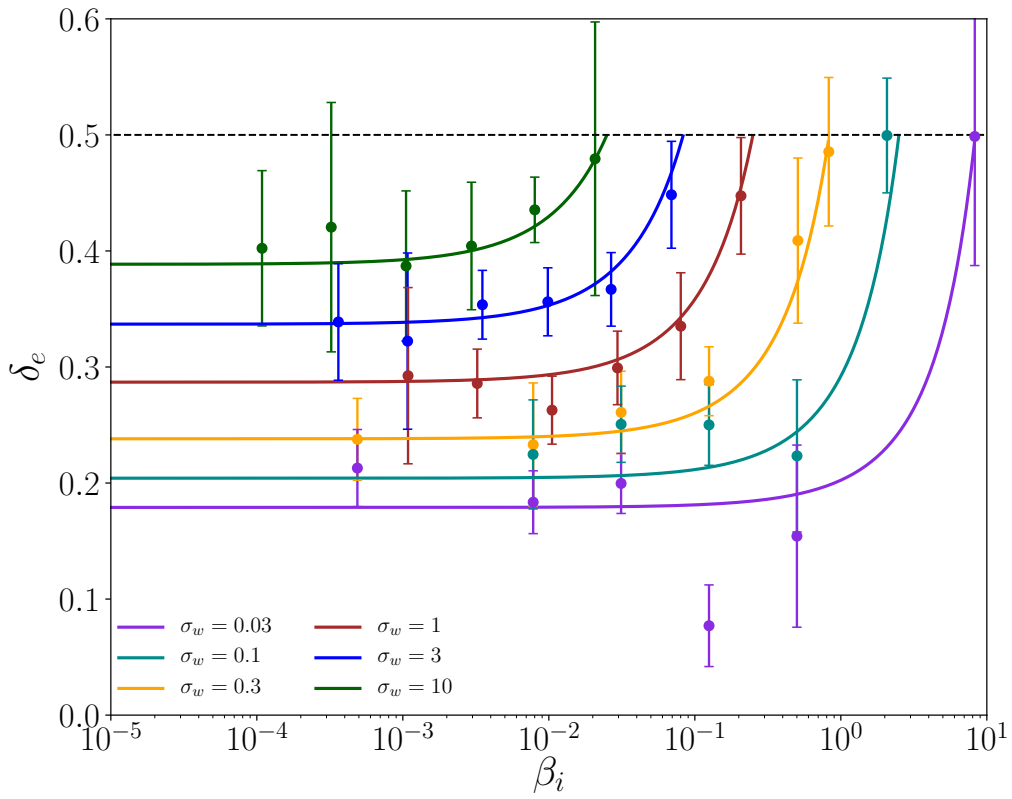


Figure 1.2: Reconnection heating prescription (Equation 1.37) fit to PIC simulation data from Rowan et al. (2017). The irreversible heating fraction $\delta_{\rm e}$ is plotted against β_i for various values of magnetization σ_w, indicated by the color. Points show the results of simulations and lines show a fit to these data. Error bars indicate, roughly, $\pm 1.0\,\sigma$ confidence intervals on the heating fractions extracted from each PIC simulation. For a given σ_w, the maximum allowed $\beta_{\rm i}$, where the ions and electrons both are ultra-relativistic, is $\beta_{\rm i,max} = 1/4\sigma_w$. The functional form constrains the heating fraction $\delta_{\rm e} \to 0.5$ as $\beta_{\rm i} \to \beta_{\rm i,max}$.

form:

$$\delta_{\rm e} = \frac{1}{2} \exp\left[\frac{-(1-\beta_{\rm i}/\beta_{\rm i,max})}{0.8+\sigma_w^{0.5}}\right], \tag{1.37}$$

where $\beta_{\rm i} < \beta_{\rm i,max} = 1/4\sigma_w$. This fit has the expected asymptotic behavior. For example, when $\beta_{\rm i} \to \beta_{\rm i,max}, \delta_{\rm e} = 0.5$ for any value of σ_w. Similarly, when $\sigma_w \gg 1$, $\delta_{\rm e} = 0.5$, independent of $\beta_{\rm i}$.

Figure 1.2 shows the irreversible heating fractions $\delta_{\rm e}$ as measured in PIC simulations from Rowan et al. (2017). The initial conditions of the upstream plasma in these simulations span the trans-relativistic regime of reconnection. In all PIC simulation data presented, the upstream temperature ratio is fixed to unity ($T_{\rm e}/T_{\rm i} = 1$) and the mass ratio is the physical $m_{\rm p}/m_{\rm e} = 1836$. For these PIC

simulations, the ratio of ion thermal to magnetic pressure spans the range $10^{-4} < \beta_{\rm i} < 10$, and the magnetization ranges from $0.03 < \sigma_w < 10$. Each curve in Figure 1.2, corresponding to a particular value of σ_w, is plotted as a function of $\beta_{\rm i}$ up to its maximum possible value $\beta_{\rm i,max} = 1/4\sigma_w$. The dashed black line at $\delta_{\rm e} = 0.5$ indicates the limit for which electrons and ions have comparable heating efficiencies, $\delta_{\rm e} \to 0.5$.

Figure 1.1 shows that Equation 1.37 has qualitatively distinct behavior from the Howes turbulent cascade heating prescription, Equation 1.33. At fixed σ_w, taking $\beta_{\rm i} \to \beta_{\rm i,\,max}$ leads to $\delta_{\rm e} \to 0.5$, while taking high magnetizations $\sigma_w \gg 1$ also leads to $\delta_{\rm e} \to 0.5$. Plotting Equation 1.37 as a function of $T_{\rm i}$ and $\beta_{\rm i}$, making the assumption that $T_{\rm i} = T_{\rm e}$ in Figure 1.1, it is apparent that in the regime of interest for ADAF simulations ($T_{\rm e} \sim 10^{10}$–10^{12} K), $\delta_{\rm e}$ is relatively low, $\delta_{\rm e} \approx 0.2$–$0.3$. In contrast to the turbulent heating model, $\delta_{\rm e}$ never exceeds 0.5, indicating that one should never expect $T_{\rm e} > T_{\rm i}$ in accretion simulations using this model. For values $\beta_{\rm i} \ll \beta_{\rm i,max}$, the irreversible heating fraction $\delta_{\rm e}$ approaches a constant value that depends only on the magnetization; this asymptotic value of $\delta_{\rm e}$ at $\beta_{\rm i} \ll \beta_{\rm i,max}$ decreases with the magnetization. In the limit of non-relativistic reconnection ($\sigma_w \ll 0.1$), Equation 1.37 yields $\delta_{\rm e} \to 0.14$, which is consistent with the expectation that the heating fraction in the nonrelativistic reconnection limit is independent of magnetization. This result is in rough agreement with recent laboratory experiments (Eastwood et al., 2013; Yamada et al., 2014) and spacecraft observations (Phan et al., 2013, 2014).

In all the PIC simulations of Rowan et al. (2017), the magnetization $\sigma_w \geq 0.03$, while the accretion simulations presented in Chapters 2 and 3 have $\sigma_w \lesssim 10^{-3}$ in the midplane of the accretion disk at radii larger than $r \gtrsim 25\, r_g$. If the behavior of $\delta_{\rm e}$ from reconnection changes in the low σ_w regime, this will have a major effect on the results of global two-temperature simulations. Future PIC studies are needed to investigate electron heating from reconnection at low magnetization, $\sigma_w < 0.03$.

In addition, the PIC simulations in Rowan et al. (2017) focused on the case of anti-parallel reconnection. This fact may explain in part the discrepancy between the heating efficiencies quoted in Rowan et al. (2017) and the conclusions of Numata & Loureiro (2015), who performed gyrokinetic simulations (implicitly assuming strong guide fields) of non-relativistic reconnection. They found an excess of electron heating at low $\beta_{\rm i}$, similar to the qualitative predictions of the Howes (2010) prescription.[7] Rowan et al. (2019) investigated the efficiency of electron heating in PIC simulations of plasmoid-dominated reconnection in the strong guide field regime. They found that the qualitative behavior of reconnection heating in this regime is more similar to the Howes (2010) prescription than the zero guide-field case considered in Rowan et al. (2017) and this work. Finally, reconnection in collisionless accretion flows itself likely occurs at the endpoint of a turbulent cascade. Considering the effects of turbulence between the grid scale and the reconnection scale may modify the results used (e.g., Shay et al., 2018).

[7]Still, the ratio of ion to electron heating efficiency that they measured at low beta ($\sim 10^{-3}$ for $\beta_{\rm i} = 0.01$) is much higher than the prediction of the Howes (2010) model.

Text in this chapter was previously published in *MNRAS* 478 (2018), 4, pp 5209-5229 (A. Chael, M. Rowan, R. Narayan, M. Johnson, and L. Sironi).

2

Electron heating in simulations of Sgr A*

The Galactic Center compact radio source Sagittarius A* (Sgr A*) hosts a supermassive black hole of mass $M = (4.1 \pm 0.03) \times 10^6\, M_\odot$ (GRAVITY Collaboration et al., 2018a). Sgr A*'s low luminosity $\sim 10^{-9}\, L_{\rm Edd}$ (Falcke et al., 1998; Genzel et al., 2003; Baganoff et al., 2003) and correspondingly low mass accretion rate $\lesssim 10^{-7}\, \dot{M}_{\rm Edd}$ (Agol, 2000; Bower et al., 2003; Marrone et al., 2007) puts it squarely in the ADAF regime of black hole accretion. The density of accreting gas around Sgr A* is estimated to be quite low ($n \sim 10^6$ cm^{-3}), so as discussed in the Introduction, the electron-ion thermalization time in Sgr A* likely exceeds the accretion time.

GRMHD simulations have been powerful tools for exploring the physics of Sgr A*'s accretion flow and the factors that contribute to its spectrum, compact emission, and rapid variability. As discussed in Chapter 1, standard GRMHD simulations do not solve for the electron temperature $T_{\rm e}$ throughout the flow. $T_{\rm e}$ is typically set manually in a radiative transfer postprocessing step, either by fixing the temperature ratio $T_{\rm e}/T_{\rm i}$ to a constant value everywhere (e.g., Mościbrodzka et al., 2009; Dexter et al., 2010), or by dividing the simulation into jet and disk regions with different temperatures (e.g., Mościbrodzka & Falcke, 2013; Chan et al., 2015a). In particular, Event Horizon

Telescope (EHT) data and images of the inner Sgr A* accretion flow at 230 GHz (e.g., Doeleman et al., 2008; Johnson et al., 2015) hold immense promise for understanding the accretion physics operating just a few Schwarzschild radii outside Sgr A*'s event horizon. For this reason, many studies of GRMHD simulations adapted to Sgr A* have focused on predictions for the optically thin 230 GHz emission surrounding the black hole shadow (e.g., Dexter et al., 2010; Shiokawa, 2013; Mościbrodzka et al., 2014; Chan et al., 2015a).

Chapter 1 developed an alternative approach to simulating systems like Sgr A* and producing spectra and images from these models. This approach, developed for the GRRMHD code KORAL by Sądowski et al. (2017), solves for T_e self-consistently during the run of the simulation using a sub-grid model for the underlying dissipation physics that sets the ratio of the electron and ion heating rates. Ressler et al. (2017) first used a similar method to successfully produce a model of Sgr A* in line with observations of the quiescent spectrum and variability properties. While a significant advance, their method did not include radiative feedback and used fixed adiabatic indices to save computational resources by solving the evolution equation for the electron entropy in post-processing. The approach in Sądowski et al. (2017) is a more general framework, where ions and electrons are evolved simultaneously along with the total gas and radiation including radiative and Coulomb couplings. Furthermore, Ressler et al. (2017) and earlier studies only used the Howes (2010) prescription for the electron heating fraction δ_e. This model was originally developed to explain non-relativistic solar wind observations, and it is unclear if it is applicable in the environment around Sgr A*.

This chapter uses the technique developed in Chapter 1 to perform four 3D two-temperature simulations of Sgr A* (originally presented in Chael et al. 2018a). These four simulations explore both the Howes (2010) Landau-damped cascade prescription for electron heating (Section 1.4.1) and a prescription for magnetic reconnection heating (Section 1.4.2) fit to particle-in-cell simulation data

from Rowan et al. (2017). The two heating prescriptions are compared in simulations with both low ($a = 0$) and high ($a = 0.9375$) black hole spins. The predicted spectra, lightcurves, and images at 230 GHz and lower frequencies are compared to the available data. While it is not possible to make conclusive statements about the plasma microphysics or accretion physics operating in Sgr A* with this limited survey of the parameter space, the analysis of this chapter is able to rule out certain parameter combinations and demonstrates the power of comparing these self-consistent simulations to data (including from the EHT) to make progress toward understanding the plasma physics in the Sgr A* accretion flow.

It is important to note that, in Sgr A*, electron-electron collisions may not be sufficient to entirely relax the electron distribution function to a thermal Maxwellian (Mahadevan & Quataert, 1997). Shocks and magnetic reconnection can accelerate a fraction of the electrons into relativistic nonthermal distributions which persist alongside the lower-energy thermal distribution. These nonthermal electrons may be responsible for Sgr A*'s continual flares across the spectrum (e.g., Marrone et al., 2008; Yusef-Zadeh et al., 2009; Eckart et al., 2012). To explore nonthermal particle evolution and emission, Chael et al. (2017) introduced an extension of the two species thermal approach (Sądowski et al., 2017) to evolve arbitrary electron distributions in accretion simulations. This method is presented in Chapter 4.

2.1 Simulations

2.1.1 Units

The accretion simulations presented in this chapter use a black hole mass fixed to a value appropriate for Sgr A*, $M = 4 \times 10^6\, M_\odot$ (Gillessen et al., 2009; Chatzopoulos et al., 2015). The gravitational radius is $r_g = GM/c^2 = 6 \times 10^{11}\,\mathrm{cm} = 0.04\,\mathrm{AU}$, and the gravitational time-scale is $t_g = r_g/c = 20\,\mathrm{s}$.

The Eddington accretion rate $\dot{M}_{\rm Edd} = 0.16\, M_\odot\ {\rm yr}^{-1}$, and the Eddington luminosity $L_{\rm Edd} = 5{\times}10^{44}$ erg s^{-1}. For this value of the Eddington accretion rate (Equation 0.3) the efficiency is set at $\eta = 0.057$, the value for a thin accretion disk with $a = 0$ (Novikov & Thorne, 1973).

2.1.2 Simulation grid and initial conditions

Chael et al. (2018a) ran four simulations in the Kerr metric using a modified Kerr-Schild coordinate grid. Two spin values were considered: $a = 0$ (zero spin case) and $a = 0.9375$ (high spin case). One simulation for each of the heating prescriptions described in Section 1.4 was run at each black hole spin value. This chapter thus considers four models: a spin zero turbulent heating prescription model, H-Lo, a spin zero model using the magnetic reconnection heating prescription, R-Lo, a spin $a = 0.935$ turbulent heating prescription model, H-Hi, and a corresponding $a = 0.9375$ model heated by magnetic reconnection, R-Hi. The simulation parameters are summarized in Table 2.1.

Each simulation was run on a 3D grid with a resolution of $256{\times}192{\times}96$ cells in radius, polar angle, and azimuth. The radial cells are distributed exponentially from a spin-dependent Boyer-Lindquist radius $r_{\rm min}$ inside the black hole horizon out to $5 \times 10^3\, r_{\rm g}$. The azimuthal cells are distributed uniformly over the range $[-\pi, \pi]$, while the polar angle cells are sampled over the interval $[0, \pi]$ using the function presented in Appendix B. To better resolve the magnetorotational instability in the disk, the grids more densely sample the regions closer to the equatorial plane.

The initial gas torii were set up using the Penna et al. (2013) equilibrium solution, with weak dipolar magnetic field loops added in the (r, θ) plane. The initial energy density in electrons was taken as one percent of the total gas energy density, with the remainder in the ions. The initial torii and simulation grids are presented in more detail in Appendix B, and are displayed in Figure B.1.

To speed up the simulations during the initial stages, the simulations were first evolved for a total time of $2 \times 10^4\, t_{\rm g}$ in 2D, suppressing the ϕ coordinate and assuming axisymmetry. This

Table 2.1: Setup of the four Sgr A* simulations.

Model	Spin	Heating	$r_{\rm min}$	$r_{\rm max}$	$N_r \times N_\theta \times N_\phi$	$[t_{\rm i}, t_{\rm f}]$ $(t_{\rm g})$
H-Lo	0	Turb. Cascade	1.5	5000	$320 \times 192 \times 96$	$[2.8 \times 10^4,\ 3.3 \times 10^4]$
R-Lo	0	Mag. Reconnection	1.5	5000	$320 \times 192 \times 96$	$[2.5 \times 10^4,\ 3.0 \times 10^4]$
H-Hi	0.9375	Turb. Cascade	1	5000	$320 \times 192 \times 96$	$[2.7 \times 10^4,\ 3.2 \times 10^4]$
R-Hi	0.9375	Mag. Reconnection	1	5000	$320 \times 192 \times 96$	$[2.8 \times 10^4,\ 3.3 \times 10^4]$

Table 2.2: Properties of the four Sgr A* simulations.

Model	Spin	Heating	$\dot{M}/\dot{M}_{\rm Edd}$	$\Phi_{\rm BH}/\sqrt{\dot{M}c\,r_{\rm g}}$
H-Lo	0	Turb. Cascade	3×10^{-7}	5
R-Lo	0	Mag. Reconnection	7×10^{-7}	4
H-Hi	0.9375	Turb. Cascade	2×10^{-7}	6
R-Hi	0.9375	Mag. Reconnection	3×10^{-7}	3

step used the mean-field dynamo presented in Sądowski et al. (2015) to prevent the decay of the axisymmetric magnetic field. After running the simulations for $2 \times 10^4\, t_{\rm g}$ in 2D, the output was regridded into 3D, rescaling the gas density by a factor of 10 and introducing 5% perturbations in the azimuthal three-velocity v_ϕ to seed departures from axisymmetry. The rescaling factor of 10 was chosen from test simulations in order to achieve the desired accretion rate ($\dot{M} \sim 10^{-7}\, \dot{M}_{\rm Edd}$; Marrone et al. 2007) in the 3D run. Since the level of angular momentum transport facilitated by the self-consistent MRI turbulence is somewhat less than that supplied by the dynamo in 2D, the accretion rate was generally lower in 3D than in 2D for the same gas density. The simulations were evolved in 3D for another $1.5 \times 10^4\, t_{\rm g}$, with the mean-field dynamo turned off.

The results presented below correspond to a $5000\, t_{\rm g}$ period for each simulation, taken from a selected range between $2.5 \times 10^4 t_{\rm g}$ and $3.5 \times 10^4 t_{\rm g}$ (see Table 2.1 for the exact range selected for each simulation).

2.1.3 Radiative transfer

Spectra and lightcurves from the four simulations were computed using the post-processing code `HEROIC`, (Zhu et al., 2015; Narayan et al., 2016). `HEROIC` solves for a spectrum and angular distribution of radiation at each grid position self-consistently using the geodesic equation and radiative transfer equation. This self-consistent solution allows for different geodesics to exchange intensities from the scattering of photons by electrons. `HEROIC` includes synchrotron, bremsstrahlung, and inverse Compton scattering in its radiative transfer calculations. At the 1.3 mm observing wavelength of the EHT, synchrotron radiation dominates the emission. To produce higher-resolution images of the accretion flow at 1.3 mm wavelength, the ray tracing and radiative transfer code `grtrans` (Dexter, 2016) was used, including only thermal synchrotron opacities.

Spectra and lightcurves were computed for several different inclination angles (10°, 20°, 40°, 60°, 80°, and 90°), measured down from the north pole. In computing lightcurves, both `HEROIC` and `grtrans` use the 'fast light' approximation, where individual images are computed using fixed lab frame time slices of the simulation output. In other words, the fluid is not allowed to evolve as photons propagate in the post-processing codes.

Before running `HEROIC` or `grtrans`, the density and magnetic field strength in the simulations were scaled (keeping the electron temperature fixed), so as to match the average observed Sgr A* 1.3 mm flux density (≈3.5 Jy: Bower et al. 2015) at an assumed inclination of 60°. The density scaling factors were different for each model, ranging from 0.06 (model H-Hi) to 1.75 (model R-Lo). Because these two-temperature GRRMHD simulations include radiation and Coulomb couplings that are not scale-free, this procedure is not physically consistent if these couplings are dynamically important. For Sgr A*, Coulomb and radiation couplings do not significantly alter the gas dynamics, so a limited amount of rescaling in post-processing should not affect the validity of these results. A

more consistent procedure would be to identify rescaling factors and then re-run the selected part of the simulation with the scaled primitives. This will be particularly important in higher accretion rate systems where radiation coupling starts to become important, such as M87 (Ryan et al. 2018; Chael et al. 2019b; 3).

Because $\sigma_{\rm i}$ can exceed unity in the jet region close to the poles, the plasma dynamics in this region are dominated by the magnetic field. Small errors in conserving total energy in the simulation can then lead to large errors in the fluid energy density, and hence the electron temperature. Furthermore, the plasma density is extremely low in this region and is often determined by a numerical floor imposed for stability. For these reasons, the innermost 4 layers of cells closest to each polar axis were not included in the radiative transfer postprocessing. Section 3.1.3 discusses these numerical floors and the reliability of the temperature evolution in high $\sigma_{\rm i}$ regions in more detail, as it is a particular problem for the magnetically dominated simulations of M87 presented in Chapter 3.

2.2 Results

2.2.1 Accretion flow properties

Figures 2.1, 2.2, and 2.3 show quantities averaged in azimuth and time over the $5000\,t_{\rm g}$ period selected for each simulation. Figure 2.1 displays the mass density ρ, the gas temperature $T_{\rm gas}$, the electron temperature $T_{\rm e}$, the ion temperature $T_{\rm i}$, the temperature ratio $T_{\rm e}/T_{\rm i}$, and the effective gas adiabatic index, $\Gamma_{\rm gas}$ (see Equation 1.26). Figure 2.2 shows the electron heating fraction $\delta_{\rm e}$, the magnetization $\sigma_{\rm i}$, the ratio of ion thermal pressure to magnetic pressure $\beta_{\rm i}$, the fluid frame radiation power per unit volume $\hat{G}^0$, and the power per unit volume produced by inverse Compton scattering $\hat{G}^0_{\rm IC}$, as calculated in KORAL using frequency-averaged quantities (Sądowski & Narayan, 2015).

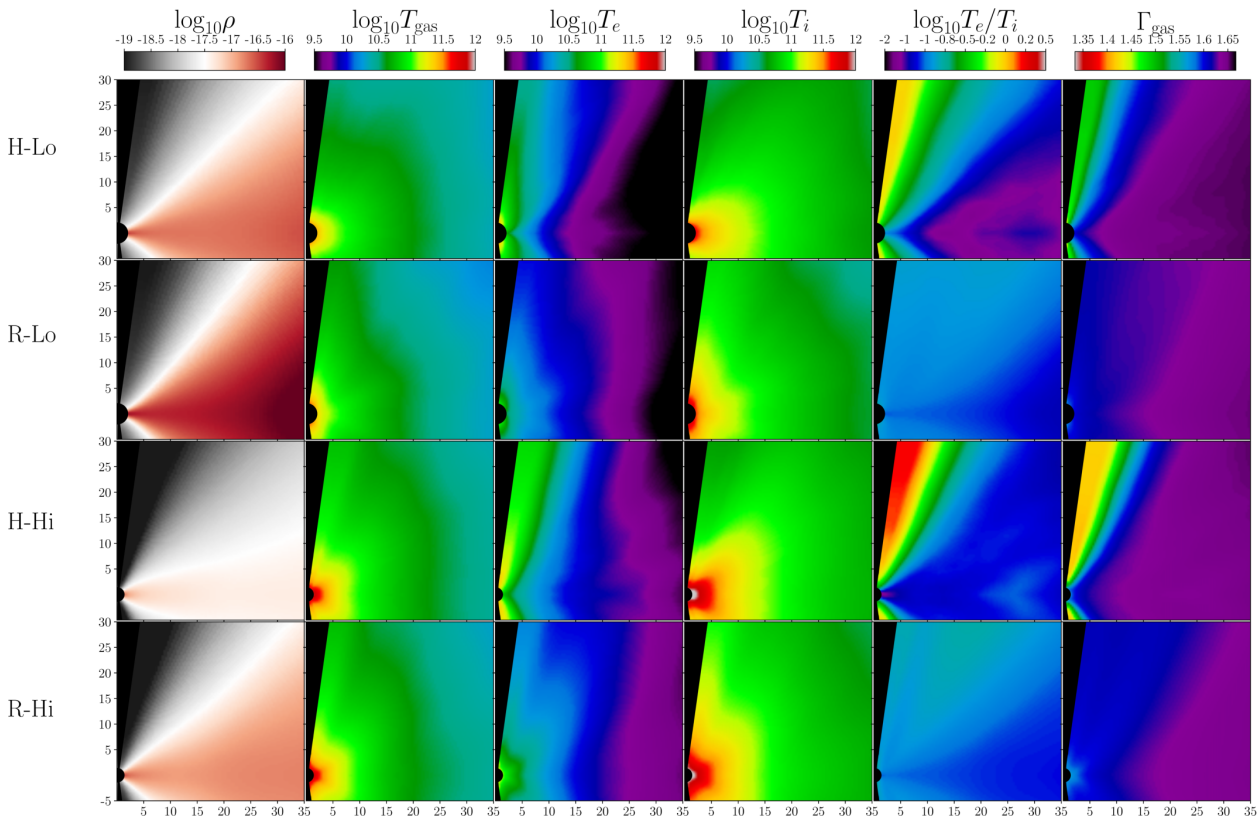


Figure 2.1: Bulk gas properties of the four Sgr A* simulations. From top to bottom, quantities are shown for the spin 0 turbulent heating model H-Lo, the spin 0 reconnection model R-Lo, the spin 0.9375 turbulent heating model H-Hi, and the spin 0.9375 reconnection model R-Hi. The fluid quantities were rescaled to produce a 230 GHz flux density around 3.5 Jy (Bower et al., 2015) when observed at 60° inclination, and they were averaged in azimuth and time for $5000\,t_{\rm g}$. The resulting averages were symmetrized over the equatorial plane. From left to right, the quantities displayed are the density ρ in g cm^{-3}, the gas temperature $T_{\rm gas}$ in K, the electron temperature $T_{\rm e}$ in K, the ion temperature $T_{\rm i}$ in K, the electron-to-ion temperature ratio $T_{\rm e}/T_{\rm i}$, and the effective gas adiabatic index $\Gamma_{\rm gas}$.

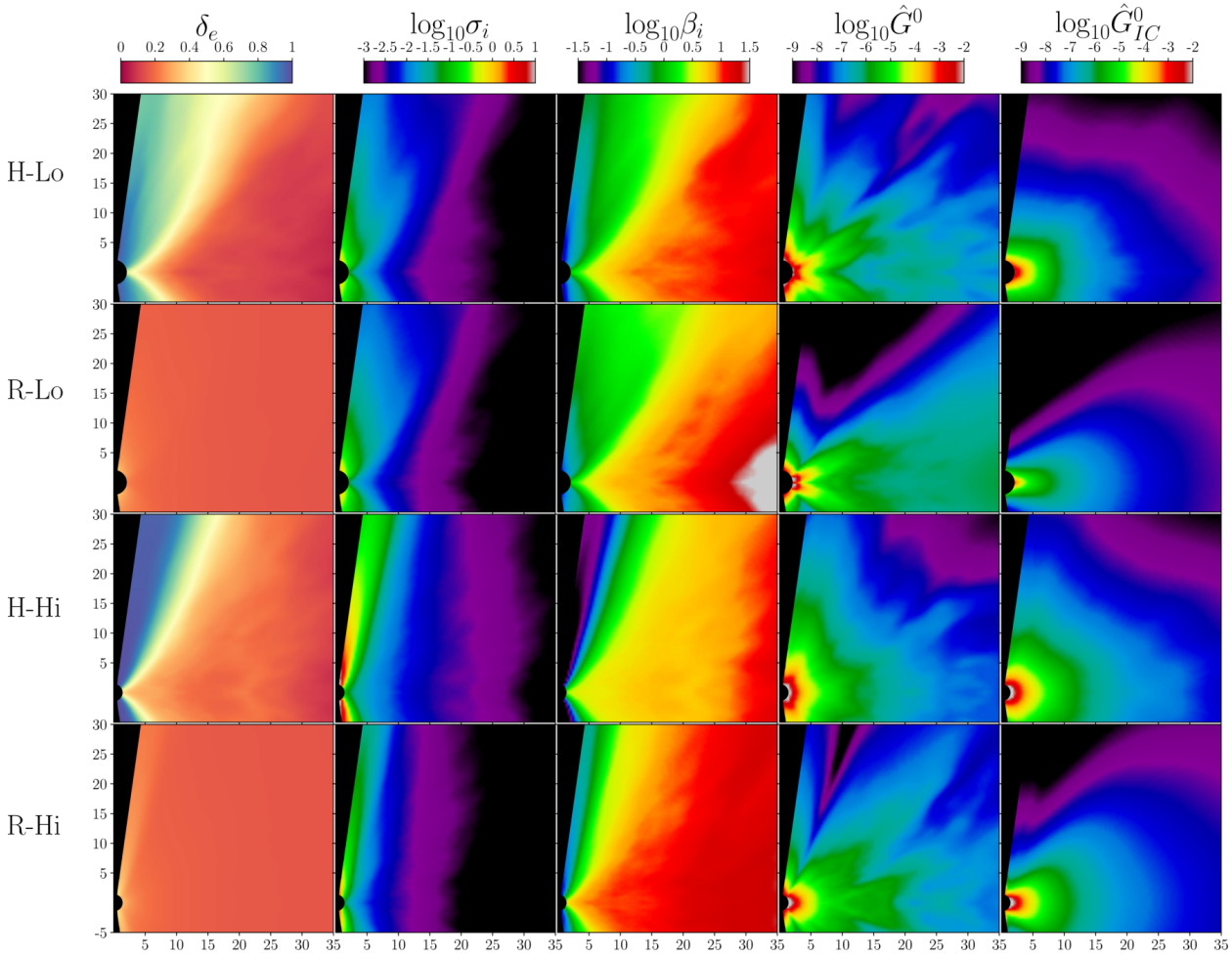


Figure 2.2: Additional azimuth and time-averaged properties of the four models of Sgr A*. From left to right, the quantities displayed are the electron heating fraction $\delta_{\rm e}$, the plasma magnetization $\sigma_{\rm i}$, the ratio of ion thermal pressure to magnetic pressure $\beta_{\rm i}$, the total rest frame radiation power $\hat{G}^0$ in erg s^{-1} cm^{-3}, and the inverse Compton radiation power $\hat{G}^0_{\rm IC}$ in the same units.

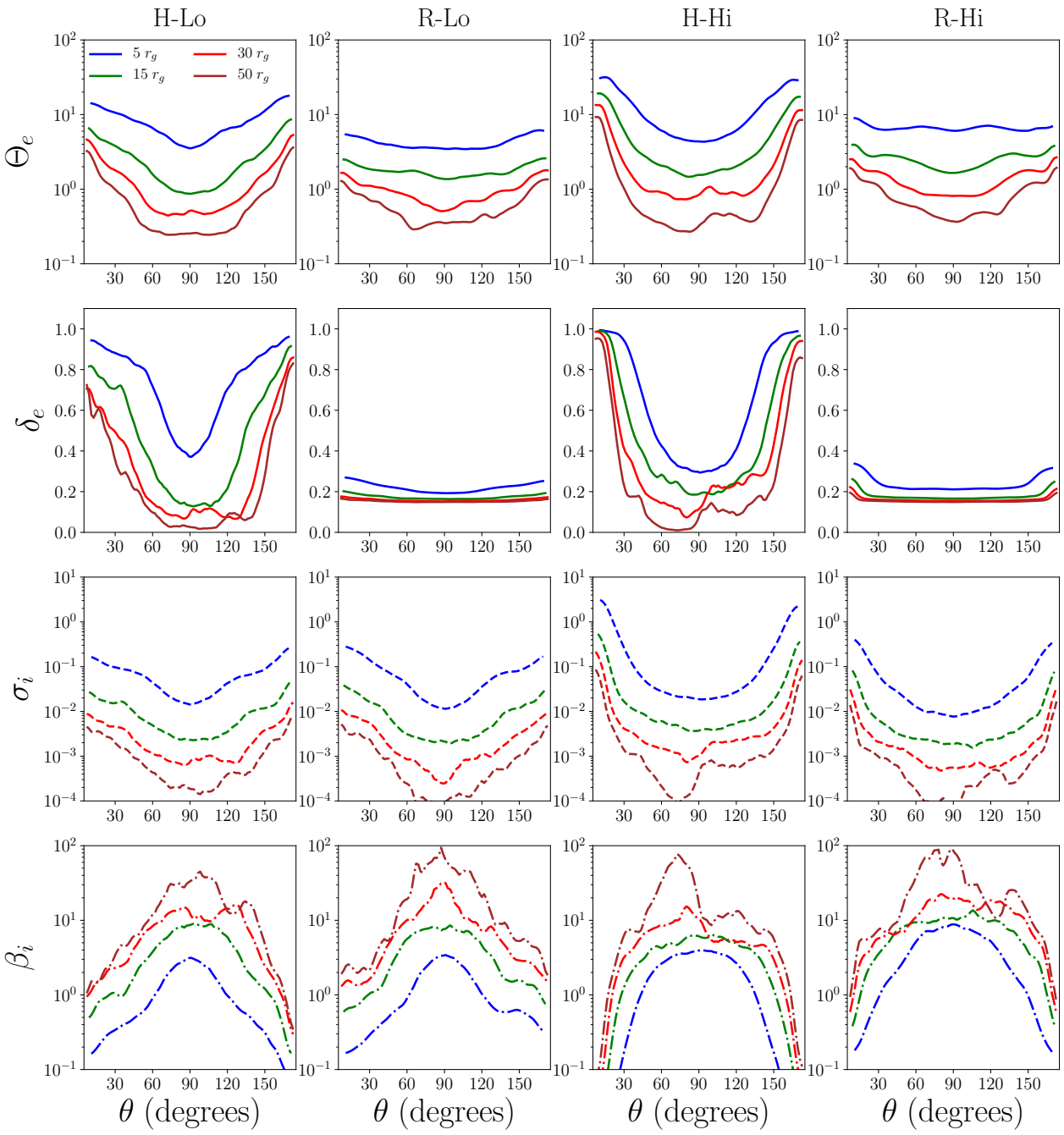


Figure 2.3: Azimuth and time-averaged fluid properties as a function of polar angle θ at four radii. Unlike in Figures 2.1 and 2.2, these data were not symmetrized over the equatorial plane. All quantities are plotted for each model at radii $5\,r_{\rm g}$ (blue), $15\,r_{\rm g}$ (green), $30\,r_{\rm g}$ (red), and $50\,r_{\rm g}$ (brown). From top to bottom, the quantities displayed are the dimensionless electron temperature $\Theta_{\rm e} = k_{\rm B}T_{\rm e}/m_{\rm e}c^2$, the electron heating fraction $\delta_{\rm e}$, the plasma magnetization $\sigma_{\rm i}$, and the ratio of ion thermal pressure to magnetic pressure $\beta_{\rm i}$.

Figure 2.3 shows angular profiles of the dimensionless electron temperature $\Theta_{\rm e} = k_{\rm B}T_{\rm e}/m_{\rm e}c^2$, the heating function $\delta_{\rm e}$, and the plasma parameters $\sigma_{\rm i}$ and $\beta_{\rm i}$, taken at three different radii. Before averaging, the primitive quantities in each simulation were scaled so as to produce an average 230 GHz flux density of approximately 3.5 Jy. Derived quantities like the species temperatures and radiation power were then recomputed from the scaled primitives. Temperatures and dimensionless ratios are unchanged by this scaling, but the density and radiation power profiles are affected.

Figure 2.1 shows that all four models produce disks that are geometrically thick. All the simulations were run from initial torii with almost identical initial density profiles, with only a slight difference between the two spins considered. However, because of the rescaling required to produce a 230 GHz flux density near the measured value for Sgr A*, the final density profiles show significant differences among the models. In particular, because the electron temperatures in the funnel and inner disk are cooler, the density in the low spin magnetic reconnection heating model R-Lo had to be scaled up to produce the right 230 GHz flux density. In contrast, the high electron temperatures in the jet/funnel region and stronger magnetic field in model H-Hi required a large downscaling in density. These differences in density are also apparent in the average accretion rates presented in Table 2.2; disk R-Lo has the highest accretion rate of $7 \times 10^{-7}\,\dot{M}_{\rm Edd}$, while disk H-Hi has the lowest overall accretion rate of $2 \times 10^{-7}\,\dot{M}_{\rm Edd}$. The accretion rates for all four models fall within the limits for Sgr A* set by Faraday rotation measurements (Marrone et al., 2007).

In addition, the $\beta_{\rm i}$ and $\sigma_{\rm i}$ distributions for the models plotted in Figures 2.2 and 2.3 show that all four models are largely in the same regime with regards to gas pressure and magnetic field strength, with some notable differences. All four models have low levels of magnetic flux, with the amount of time-averaged flux threading the black hole horizon $\Phi_{\rm BH}/(\dot{M}c)^{1/2}r_{\rm g} < 10$ in all cases. This puts these models squarely in the Standard and Normal Evolution regime of accretion (SANE: Narayan et al., 2012), well below the flux threshold for a Magnetically Arrested Disk

(MAD: Bisnovatyi-Kogan & Ruzmaikin, 1976; Narayan et al., 2003; Tchekhovskoy et al., 2011) regime, where $\Phi_{\rm BH}/(\dot{M}c)^{1/2}r_{\rm g} \approx 50$. The two low-spin models have lower field strengths, with $\sigma_{\rm i} < 0.1$ everywhere except extremely close to the black hole, and in both models $\sigma_{\rm i}$ falls to 10^{-3} past $20\,r_{\rm g}$. Both high spin models produce more field strength in the jet region, but $\sigma_{\rm i}$ always falls rapidly with radius in the equatorial plane. Of the four models, the turbulent heating prescription simulation at high spin, H-Hi, has the most magnetic flux. H-Hi achieves $\sigma_{\rm i} \sim 1$ in the jet region close to the black hole, and it has higher values of $\sigma_{\rm i}$ (and lower values of $\beta_{\rm i}$) at all radii compared to the magnetic reconnection model at the same spin. This model also launches a mildly relativistic jet, with a bulk Lorentz factor ≈2 at large radii.

Figure 2.1 shows that the total gas temperatures are the same order of magnitude in all models. The gas in the high-spin simulations reaches higher temperatures close to the black hole, but comparing the heating prescriptions at fixed spin, the choice of electron heating prescription has little effect on $T_{\rm gas}$. In contrast, the electron and ion temperatures vary dramatically with the choice of heating prescription. The distribution of the electron heating fraction $\delta_{\rm e}$ in the simulations is distinct for each heating prescription, and it shows only slight differences with spin. Because the turbulent heating prescription deposits most of the dissipated energy into electrons at low $\beta_{\rm i}$, models H-Lo and H-Hi both show higher values of $\delta_{\rm e} > 0.5$ in the funnel. The more magnetized jet in model H-Hi makes this transition sharper; it results in $\delta_{\rm e} \sim 1$ at low polar angles for all radii (most easily seen in the angular profiles in Figure 2.3), while in model H-Lo, $\delta_{\rm e}$ only approaches unity at small radii. This distribution of $\delta_{\rm e}$ produces electrons that are hotter in the jet/funnel, consistent with the simulations reported by Ressler et al. (2017) and Sądowski et al. (2017). However, the absolute temperatures seen in these models are lower than in those previous works, due to the weaker magnetic field. In all these models, electrons are relativistic in the outflow and inner disk, with $\Theta_{\rm e} > 1$ ($T_{\rm e} > 6 \times 10^9$ K), but the electron temperatures do not reach the very high values

$\Theta_e \sim 100$ seen in more magnetically dominated simulations.

Figure 2.1 also indicates that the temperature ratio T_e/T_i takes on values between 0.05 and 3 for the turbulent heating prescription, with an obvious structure with polar angle that transitions from $T_e < T_i$ in the disk near the equator to $T_e > T_i$ in the outflow and jet regions where the magnetic field strength is larger, β_i is lower, and $\delta_e \to 1$. In contrast, the magnetic reconnection prescription never heats the electrons more than ions. Figure 2.3 shows that in the weakly magnetized regime explored by the reconnection models, δ_e varies little with polar angle and does not exceed 0.3 on average. As a result, $T_e/T_i < 1$ everywhere. While the magnetic reconnection fitting function does put more heat in the electrons outside of the midplane where β_i is lower, the effect is small. T_e/T_i takes on a value around 0.1 near the equator and climbs to only around 0.15 at larger polar angles. The electron temperatures in the outer disk, relevant for free-free X-ray emission, are similar in both models (around 10^9 K), despite the slightly enhanced electron heating delivered by the reconnection model at these radii (see the $50\,r_g$ curve in Figure 2.3).

The last column of Figure 2.1 shows the effects of the electron heating on the total gas adiabatic index (Equation 1.26). Both turbulent heating models H-Lo and H-Hi show the total gas adiabatic index Γ_{gas} dropping to $\approx$1.4 in the funnel, as relativistic electrons with $\Gamma_e \approx 4/3$ start to dominate the fluid's energy budget. However, even in the models heated by magnetic reconnection, where electrons always have less than 50% of the total gas energy, the gas adiabatic index is not exactly $\Gamma_{gas} = 5/3$. Out to $\approx 20\,r_g$, the adiabatic index is closer to 1.6 than exactly 5/3, indicating the effects of relativistic electrons in lowering the total gas adiabatic index even when they do not constitute a majority of the fluid energy density.

As a result of the distinct electron temperature distributions that result from the two heating prescriptions, the distributions of radiation power in Figure 2.2 are also different. In the turbulent heating prescription at low spin, H-Lo, high electron temperatures in the outflow result in a

bolometric radiation power distribution (both in synchrotron and inverse Compton) that is roughly spherical. This spherical distribution is present both in the total radiation power $\hat{G}^0$, dominated by synchrotron emission, and in the inverse Compton power $\hat{G}^0_{\rm IC}$. In the magnetic reconnection heating model R-Lo, electrons in the outflow are not dramatically hotter than electrons in the disk, and because of their low density they make only a small contribution to the overall radiation power. As a result, the distributions of synchrotron and inverse Compton power in these models are disk-dominated. At high spin the picture remains largely the same. Hotter electron temperatures close to the black hole cause the magnetic reconnection model R-Hi to produce a more isotropic distribution of synchrotron power, but the contribution from the jet region is still significantly less than in the high spin turbulent heating model, H-Hi. At both low and high spin, the distributions of inverse Compton power show that when using the magnetic reconnection heating model (R-Lo, R-Hi), the Compton scattering is confined to the disk, whereas with the turbulent heating model (H-Lo, H-Hi), inverse Compton scattering produces significant power at all angles.

2.2.2 Spectra

Figure 2.4 displays the median Spectral Energy Distributions (SEDs) for all four models observed at 60° inclination, with the nominal $\pm1\sigma$ variability around the median denoted by the shaded region (assuming the spectral variability at each frequency is Gaussian distributed, the $\pm1\sigma$ band corresponds to the 68% interval between the 15.9th percentile and the 84.1th percentile). All models show the same characteristic features. At frequencies lower than 10^{11} Hz, the spectrum is dominated by optically thick synchrotron emission from the outer disk and outflow. The spectrum transitions to an optically thin synchrotron peak around 10^{12} Hz produced by emission in the inner disk close to the black hole. An inverse Compton hump between 10^{14} and 10^{18} Hz is produced from the Compton upscattering of NIR photons, and at X-ray frequencies the emission is dominated by

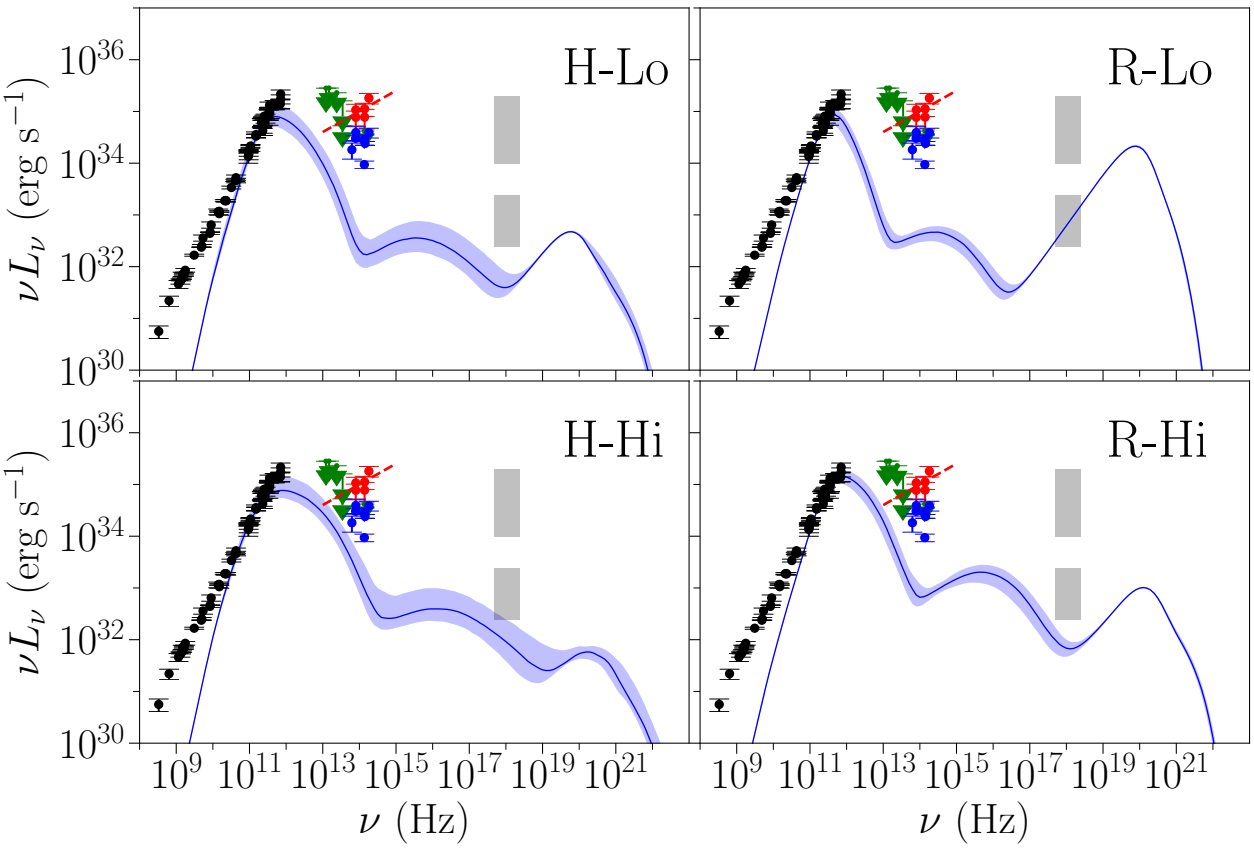


Figure 2.4: Spectral energy distributions for the four models, calculated with HEROIC for an observer at $60°$ inclination. Spectra were computed every $10\,t_g$ over a $5000\,t_g$ period. The solid blue curve shows the median spectrum for each model, and the shaded blue region shows the nominal 1σ time-variability if it is assumed that the variability distribution is Gaussian at each frequency. Data points in the radio and near-infrared are taken from references listed in Appendix A.1. Black data points show radio measurements. Green data points are near-infrared upper limits, blue data points are near-infrared quiescent measurements, and red data points are near-infrared flare measurements. The near-infrared spectral slope shown by the red line was taken from the flare measurement in Gillessen et al. (2006) as $\nu L_\nu \propto \nu^{0.4}$. The lower shaded vertical band in the X-ray represents the range of potential X-ray quiescent emission from the inner region of Sgr A*, between 10% and 100% of the total observed (Baganoff et al., 2003). The upper shaded vertical X-ray band shows the range of observed X-ray flares (Neilsen et al., 2013). From left to right, top to bottom, spectra are shown for the spin 0 turbulent heating model H-Lo, the spin 0 magnetic reconnection model R-Lo, the spin 0.9375 turbulent heating model R-Hi, and the spin 0.9375 reconnection model R-Hi.

thermal free-free emission from the hot disk at radii out to $r \sim 50\,r_{\rm g}$, peaking at 10^{20} Hz.

The inverse Compton humps of the turbulent heating models (H-Lo and H-Hi) are at higher frequencies compared to the models heated by magnetic reconnection (R-Lo and R-Hi). This trend is due to the fact that in the H models, inverse Compton emission is produced in a spherical region around the black hole, including contributions from hotter electrons in the outflow, whereas in the R models, inverse Compton scattering is confined to cooler electrons in the disk (see Figure 2.2). Consequently, the inverse Compton emission in models H-Lo and H-Hi is also more variable than in the corresponding magnetic reconnection models R-Lo and R-Hi.

All models match measurements of the high-frequency radio spectrum from optically thin synchrotron between $10^{10.5}$ and 10^{12} Hz reasonably well. However, all models under-predict the spectrum at low frequencies $\nu < 10^{10.5}$ Hz. The relatively flat low frequency slope ($L_\nu \propto \nu^{0.2}$) (Falcke et al., 1998; Herrnstein et al., 2004) could be the result of an isothermal jet or outflow not captured by the simulation (Mościbrodzka & Falcke, 2013; Ressler et al., 2017) or from emission from a population of high-energy nonthermal electrons (Özel et al., 2000; Yuan et al., 2003; Davelaar et al., 2018). Model H-Hi does the best job of fitting the low frequency spectrum down to $\sim 10^{10.5}$ Hz, whereas the other three models start failing to fit the data around 10^{11} Hz. The better performance of model H-Hi at low frequencies may be due to the increased jet emission which dominates the H-Hi model images at low frequencies (see Section 2.2.4).

X-ray emission is primarily produced by thermal bremsstrahlung at all radii out to $r \sim 50\,r_{\rm g}$, the largest radius included in the radiative transfer calculations. Because all four models have roughly similar electron temperatures $\sim 10^9$ K at this radius, the strength of the free-free peak at around 10^{20} Hz is thus primarily determined by the disk density around this radius, which is in turn set by the rescaling factor chosen to match the observed 230 GHz flux density. At each spin, because the turbulent heating models (H-Lo and H-Hi) produce higher electron temperatures in

the inner disk and outflow, they are scaled to a lower density than the corresponding magnetic reconnection heating models (R-Lo and R-Hi). Similarly, the high spin models produce hotter electron temperatures close to the black hole and therefore have density scaling factors smaller than the corresponding spin zero models, lowering their thermal free-free X-ray peaks. With the exception of model R-Lo, all of the X-ray spectra lie below the estimated quiescent luminosity from the inner disk, 10% of the total 2-10 keV X-ray luminosity measured by Baganoff et al. (2003). With the addition of free-free emission outside the maximum radius of $r = 50\,r_{\rm g}$ used in the radiative transfer, it is likely that model R-Lo would exceed the total luminosity measured by Baganoff et al. (2003).

All models substantially under-predict the observed quiescent near-infrared emission (blue data points in Figure 2.4), and the observed variability does not produce infrared flares as strong as those observed in Sgr A* (the red data points in Figure 2.4). In addition, the near-infrared spectral slope in the four thermal models is sharply negative, whereas the spectral slope measured in near-infrared flares is positive ($\nu L_\nu \propto \nu^{0.4}$, Genzel et al. 2003; Gillessen et al. 2006; Hornstein et al. 2007). The positive near-infrared spectral index suggests that flares may be produced by nonthermal electrons which are not considered in these simulations. Ponti et al. (2017) measured the spectral index of a single strong flare in the near-infrared and X-ray. In addition to confirming $\nu L_\nu \propto \nu^{0.4}$ in near-infrared, they found a difference of ≈0.5 between the X-ray and near-infrared spectral indices, suggestive of power-law synchrotron emission with a cooling break between the near-infrared and X-ray.

2.2.3 Variability

Figure 2.5 shows 230 GHz (1.3 mm) light curves for all four models observed at 60° inclination, and Figure 2.6 shows normalized near-infrared (2 μm) and X-ray (2 keV) lightcurves over the same

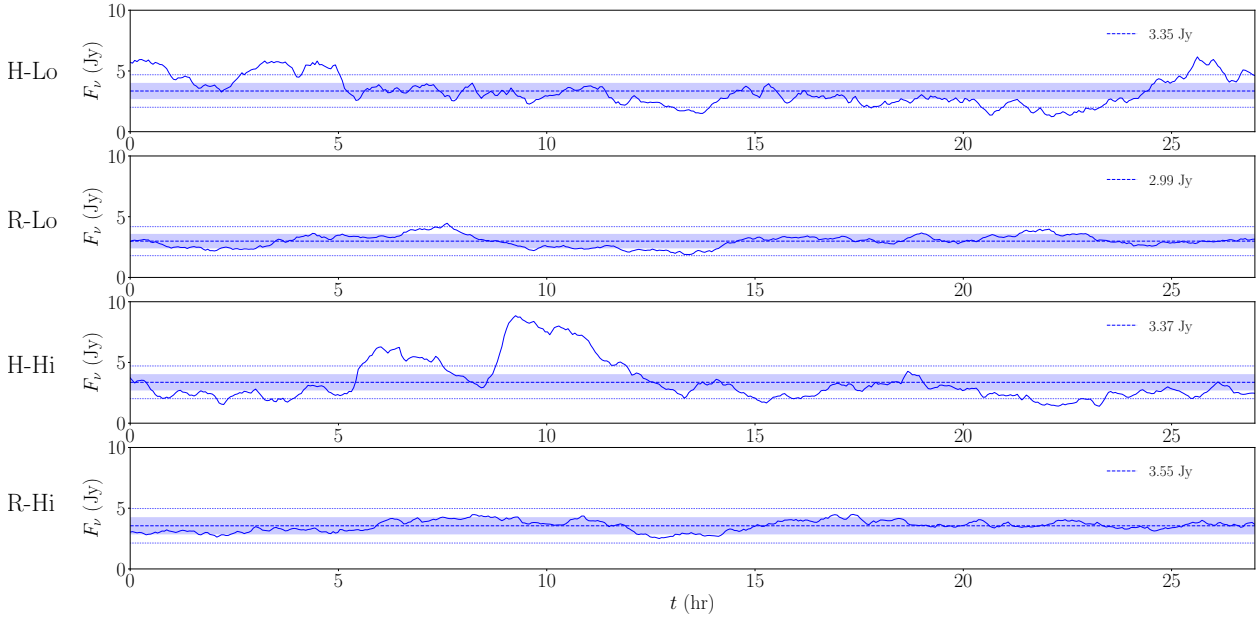


Figure 2.5: 230 GHz lightcurves of the four models at a viewing angle of 60° over intervals of $5000\,t_{\rm g}$ ($\approx$27 hr). The lightcurves were all normalized to be close to the observed average flux density of Sgr A* ($\approx$3.5 Jy, Bower et al., 2015). The dashed line shows the mean value for each lightcurve. The shaded band around the mean corresponds to 20% variability; this is roughly the root-mean-square variability level observed in Sgr A* at 230 GHz (Marrone et al., 2008). Dotted lines denote a range of 40% variability around the mean. All models show variability on hour time-scales. The models heated by magnetic reconnection have variability that falls within the observed 20% range, while the models heated by turbulent dissipation (H-Lo and H-Hi) have larger variability amplitudes. Model H-Hi shows two quasi-'flares' around 5 hr and 10 hr that produce excursions above two times the quiescent value.

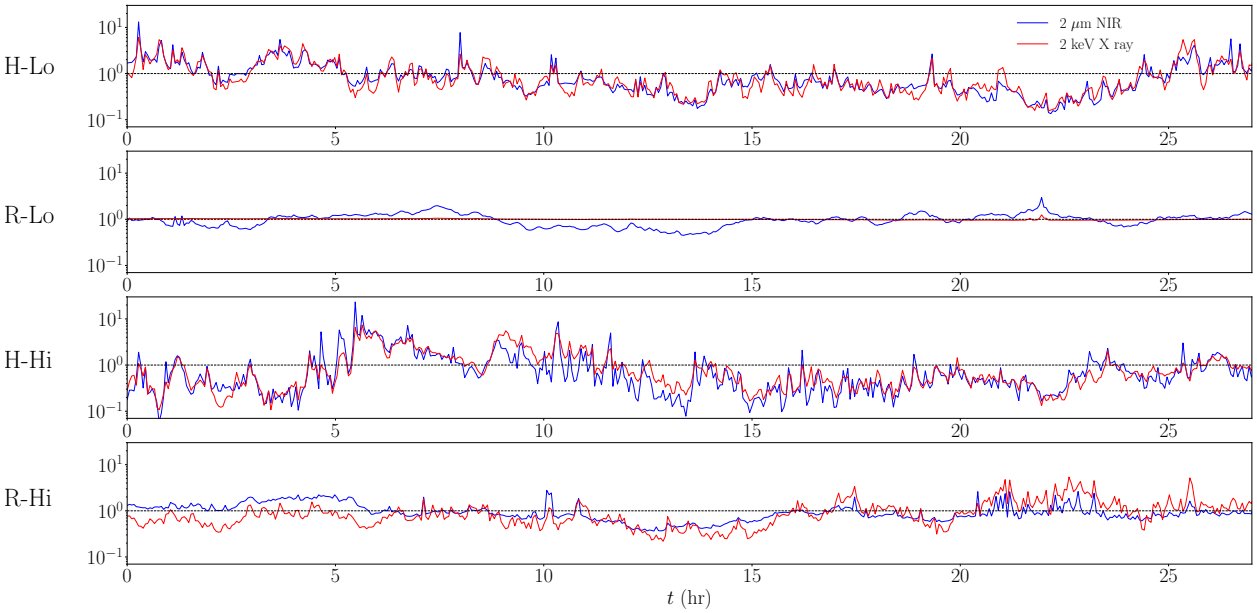


Figure 2.6: Normalized 2μm NIR and 2 keV X-ray lightcurves of the four models at a viewing angle of 60° over intervals of $5000\,t_\mathrm{g}$ ($\approx$27 hr). The curves are normalized to their mean value over the interval. Near-infrared variability arises in thermal synchrotron emission very close to the black hole, and the variability time-scale is shorter than at 230 GHz. X-ray variability results from inverse Compton scattering of near-infrared photons, and is therefore correlated with the near-infrared variability. In model R-Lo, inverse Compton scattering occurs at lower temperatures and does not upscatter enough photons to 2 keV to outshine the quiescent free-free emission from larger radii. In the other models, all X-ray flaring events have a near-infrared counterpart, but some near-infrared peaks do not get upscattered to 2 keV. These thermal simulations produce no strong X-ray flares with amplitudes >10 times quiescence.

time range. The turbulent heating prescription simulations (H-Lo, H-Hi) are more variable than their magnetic reconnection counterparts at all frequencies, since more emission in these models is produced in the high-velocity outflow region away from the equatorial plane. In contrast, emission is mostly confined to the disk in the reconnection models (Figure 2.1). In the extreme case, the spin zero reconnection model R-Lo produces practically no 2 keV X-ray variability, since Compton scattering in the cool disk does not produce enough emission at this frequency to dominate over the quiescent free-free X-ray emission from large radii (See Figure 2.4).

In all cases, the near-infrared and X-ray time variability is correlated, and the lightcurves are more rapidly varying than the millimeter lightcurve. Consistent with past studies (Chan et al., 2015b; Ressler et al., 2017), X-ray flaring events all have a near-infrared companion, whereas not all of the near-infrared flares are also seen in X-rays. This observation matches one qualitative result from observations (Yusef-Zadeh et al., 2009; Eckart et al., 2012), and it is explained in these simulations by X-ray flares being produced by local inverse Compton scattering of near-infrared photons generated by synchrotron emission. However, none of the X-ray flares is anywhere near as bright as those frequently measured from Sgr A*.

All four models fail to capture other important features of Sgr A*'s variability. Other than a few large spikes in the near-infrared in model H-Hi, there are no flares more than 10 times brighter than the average, while flares up to 30 times quiescence are observed in the near-infrared (Dodds-Eden et al., 2011; Witzel et al., 2012). These simulations furthermore do not produce any strong X-ray flares with brightness >10 times the quiescent flux as are observed on roughly 24 hour time-scales (Neilsen et al., 2013). As noted in Section 2.2.2, the spectral index of the near-infrared flares from these models is negative in νL_ν, while the measured flare spectral index is positive and seems to be stable over time. (Gillessen et al., 2006; Ponti et al., 2017). Furthermore, measurements of the near-infrared and X-ray spectral indices suggest a cooling break between these bands, indicating that

the flaring emission is synchrotron emission from a power-law nonthermal distribution (Marrone et al., 2008; Dodds-Eden et al., 2009; Ponti et al., 2017).

The 230 GHz variability is less pronounced in all cases than the corresponding near-infrared emission, with variability occurring on longer time-scales ($\gtrsim$ 1 hr). The variability amplitude is also less than at shorter wavelengths. The models heated by magnetic reconnection (R-Lo, R-Hi) produce less 230 GHz variability, as their emission is constrained to the less-active disk midplane. The variability in these models falls under the roughly ∼20% intraday root-mean-square level observed at 230 GHz (Marrone et al., 2008; Yusef-Zadeh et al., 2009; Bower et al., 2015). In contrast, both the models heated by the turbulent heating prescription (H-Lo, H-Hi) show significantly more variability than the 20% observed in Sgr A*. Model H-Hi is the most variable at 230 GHz, showing two large excursions with amplitude more than 2 times the average level; these result from sudden activity as material is ejected along the relatively powerful jet. Such dramatic intraday 230 GHz flares have not been observed from Sgr A*.

2.2.4 Images

Images of the four models at 60° inclination and 230 GHz, the observing frequency of the EHT, are presented in Figure 2.7. In all models, the emission at 230 GHz is produced by optically thin synchrotron. The linear scale images are similar among the models. They are brightest on the approaching side of the disk, where emission is relativistically beamed toward the observer. The shadow of the black hole and photon ring are visible in all four models, but slightly more emission emerges from the disk in front of the black hole in models R-Lo and R-Hi. The insensitivity of the appearance of the black hole shadow to the choice of electron heating prescription in these models is encouraging for the prospects of the EHT to measure the size of the photon ring and thus test this strong-field prediction of general relativity.

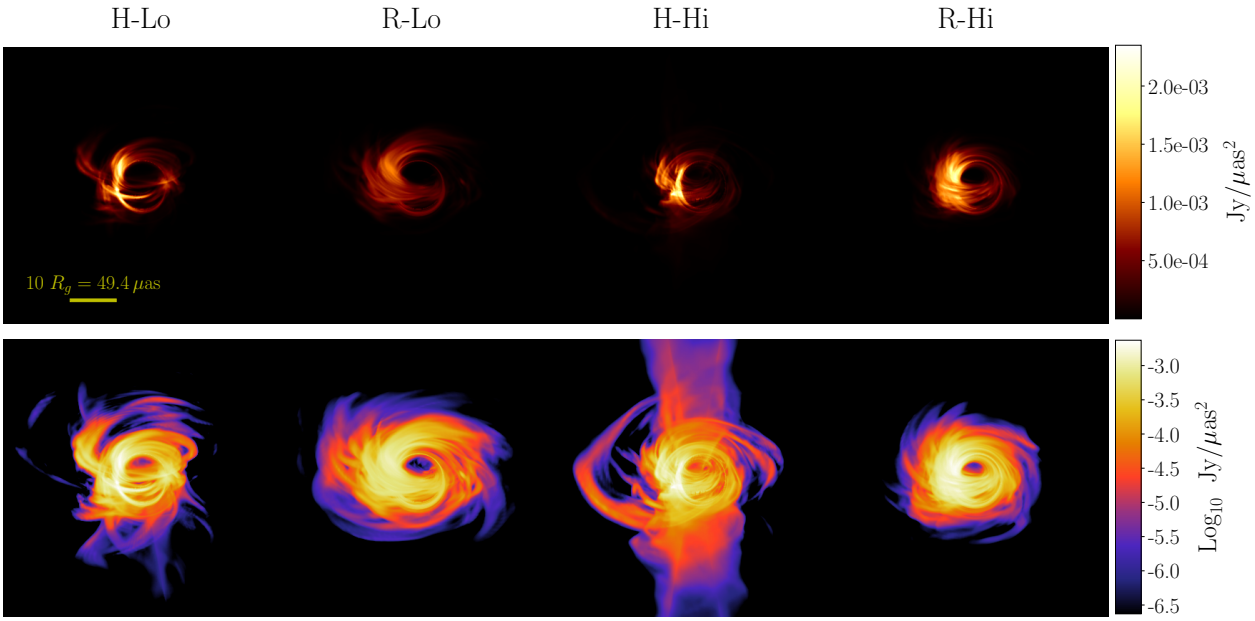


Figure 2.7: 230 GHz (observing frequency of the EHT) snapshot images at $60°$ inclination. The top row shows images with a linear scale, while the bottom row uses a log scale with a dynamic range of 10^4. The black hole shadow is apparent in all linear scale images. In log scale, the models heated by the turbulent cascade model show emission in the outflow/jet that is 100–1000 times fainter than the Doppler-boosted disk emission, while the magnetic reconnection models have all their emission confined to the disk.

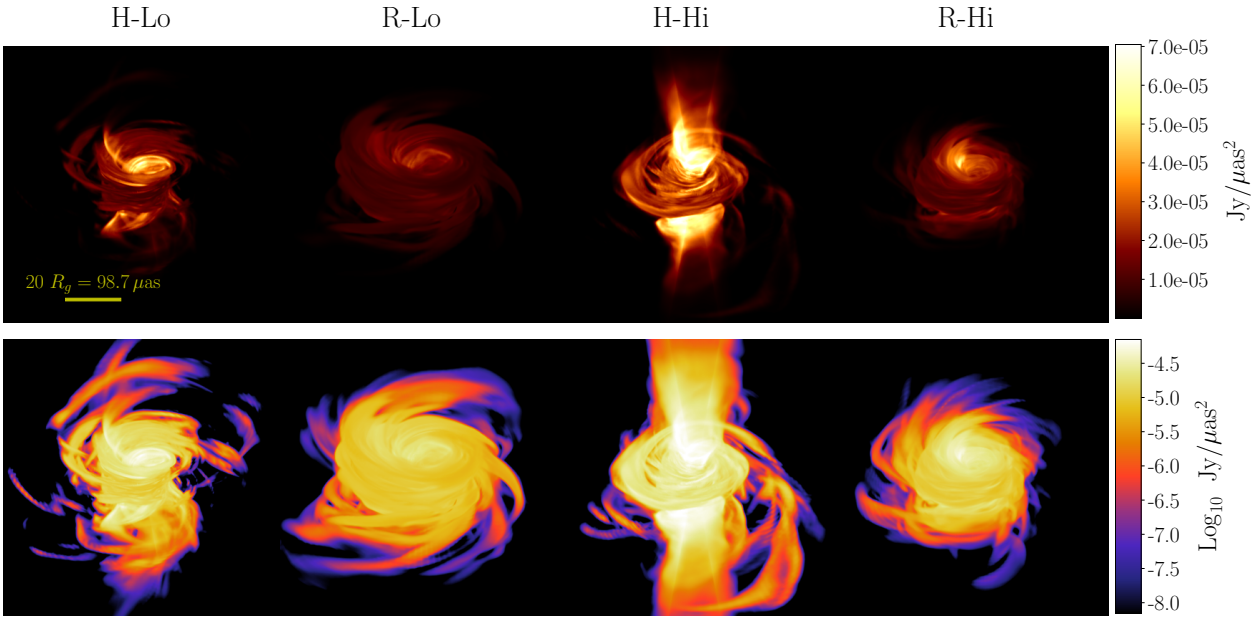


Figure 2.8: The same snapshots presented in Figure 2.7 at 43 GHz and $60°$ inclination. The top row shows images with a linear scale and the bottom row with a log scale. At this lower frequency, the models heated by the damped turbulent cascade prescription show a pronounced polar outflow, particularly the high spin model H-Hi, which launches a mildly relativistic jet. The magnetic reconnection heated disks produce larger, dimmer images at this wavelength, with emission produced only in the thick disk.

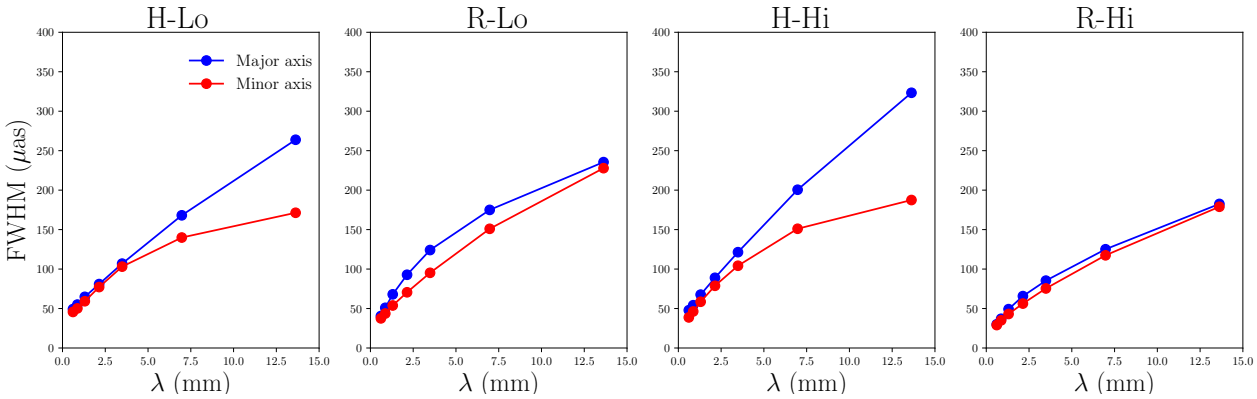


Figure 2.9: Average image sizes as a function of wavelength for the four models at $60°$ inclination. Images were averaged over $5000\,t_g$ and the image sizes along the major and minor axis were calculated by fitting an elliptical Gaussian to the image Fourier transform. The major axis FWHM data are plotted in blue and the minor axis data are plotted in red. From left to right, sizes are presented for models H-Lo, R-Lo, H-Hi, and R-Hi. The image size grows with wavelength in all cases. In the optically thin regime at wavelengths shorter than 1.3 mm, all the models behave similarly, with similar image sizes growing linearly with wavelength. At longer wavelengths, the models using the turbulent heating prescription show a large anisotropy as the jet/polar outflow begins to dominate the emission, while the models heated by magnetic reconnection remain isotropic.

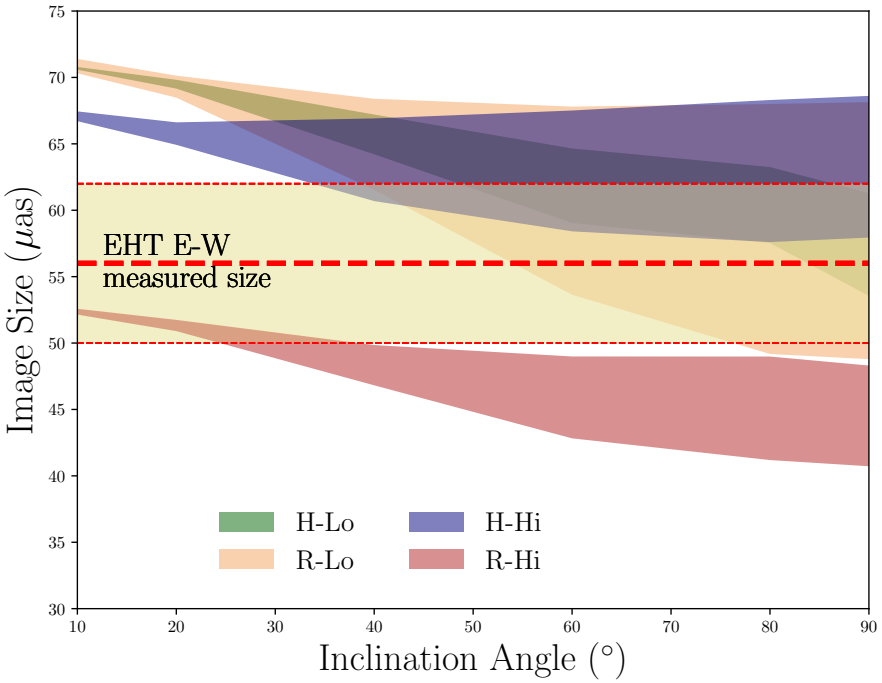


Figure 2.10: Average image sizes at 230 GHz for the four models as a function of inclination angle. Images were averaged over $5000\,t_g$ and the image sizes in the major and minor axis were calculated by fitting an elliptical Gaussian to the image Fourier transform. The image sizes are plotted as bands marking the range of values between the fitted major axis FWHM and minor axis FWHM at each inclination angle. The range of values for the East–West 230 GHz image size measured by the EHT ($56 \pm 6\,\mu$as; Johnson et al., 2018) is plotted as a yellow band. In all cases, the image size grows with decreasing inclination angle as Doppler beaming becomes less significant. All four models satisfy the EHT constraint at a given inclination, but model R-Hi only falls within the measured range when viewed nearly face-on.

While the linear scale 230 GHz images are similar, the log-scale images reveal significant differences. The models heated by reconnection produce more extended disk emission, and the turbulent cascade heated models produce faint emission in a jet at 230 GHz. Figure 2.8 shows images of the same snapshots at 43 GHz. At this frequency, the emission is from optically thick synchrotron and the black hole shadow is obscured in all models. The models heated by the turbulent cascade prescription produce most of their 43 GHz emission in the jet or outflow region at large polar angle. The high spin model H-Hi has a strong jet collimated around the polar axis, while the polar outflow in the spin zero model is less collimated but still produces an image elongated along the axis perpendicular to the disk.

Motivated by interferometric observations of Sgr A*, a representative 'size' for each image at various wavelengths can be computed to compare to the interferometric data from the VLBA and EHT. Specifically, each baseline joining two sites of an interferometric array samples a visibility $\tilde{I}(\boldsymbol{u})$, given by the Fourier transform of the image $I(\boldsymbol{x})$:

$$\tilde{I}(\boldsymbol{u}) = \int d^2\boldsymbol{x}\, I(\boldsymbol{x})\, \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{u}\cdot\boldsymbol{x}}. \tag{2.1}$$

In this expression, $\boldsymbol{x}$ is an angular coordinate on the image measured in radians, and $\boldsymbol{u}$ is the baseline vector measured in units of the observing wavelength. On a short baseline that only partially resolves the source,

$$\tilde{I}(\boldsymbol{u}) \approx \int d^2\boldsymbol{x}\, I(\boldsymbol{x}) \left[1 - 2\pi \mathrm{i} \boldsymbol{u}\cdot\boldsymbol{x} - 2\pi^2 \left(\boldsymbol{u}\cdot\boldsymbol{x}\right)^2\right]. \tag{2.2}$$

The term linear in $\boldsymbol{u}$ gives a visibility phase slope with baseline length that is proportional to the position of the image centroid. Millimeter VLBI usually lacks absolute phase referencing, so the image centroid can be redefined to be at the origin, eliminating the linear term. Short baselines

will then see a quadratic fall in the visibility amplitude $\left|\tilde{I}(\boldsymbol{u})\right|$ with increasing baseline length. The quadratic coefficient is proportional to the second moment of the image projected along the baseline vector direction. Thus, the characteristic size along a specified direction corresponds to this second moment. In general, the size is anisotropic and will be defined by a quantity analogous to the image moment of inertia tensor. The image major and minor axis sizes are reported in terms of the equivalent Gaussian major axis full width at half maximum (FWHM), minor axis FWHM, and position angle (measured east of north). For example,

$$\theta_{\rm maj} = \sqrt{-\frac{2\ln(2)}{\pi^2 I_0}\,\nabla^2_{\hat{\boldsymbol{u}}_{\rm maj}} \tilde{I}(\boldsymbol{u})\Big|_{\boldsymbol{u}=0}}, \tag{2.3}$$

where $\theta_{\rm maj}$ is the characteristic FWHM of the major axis, $I_0 \equiv \tilde{I}(\mathbf{0})$ is the total flux density of the image, and $\nabla^2_{\hat{\boldsymbol{u}}_{\rm maj}}$ is the second directional derivative along the direction of the major axis.

For each model, synchrotron images were generated with `grtrans` at 22, 43, 86, 240, 230, 345, and 490 GHz, time averaged over the $5000\,t_{\rm g}$ range considered for each simulation. While none of the simulations reproduce the flat low-frequency spectrum, all the frequencies considered here are high enough to approximately match the measured spectrum of Sgr A* (see Figure 2.4). The image size and orientation were then computed according to the definition above (Equation 2.3).

Figure 2.9 shows the resulting FWHMs of the major and minor axes for the four models observed at 60° inclination. In the optically thin regime at wavelengths shorter than 1.3 mm, all models produce approximately isotropic images that grow linearly with wavelength. At longer wavelengths in the optically thick regime, the models heated via the turbulent cascade prescription (H-Lo, H-Hi) show a large anisotropy as the jet or polar outflow begins to dominate the emission. In contrast, the models heated by magnetic reconnection (R-Lo, R-Hi) remain isotropic.

The transition to jet-dominated emission at low frequencies in the turbulent heating models

results from the way the Howes (2010) prescription puts most of the dissipated energy into the electrons in the polar regions (Figures 2.1 and 2.3). This 'disk-jet' structure is consistent with the results of Ressler et al. (2017), who used the same heating prescription (in a more magnetized system). The 'disk-jet' morphology has been explored in previous phenomenological models (e.g., Falcke & Biermann, 1995; Yuan et al., 2002) and has been applied to GRMHD simulations by setting the electron temperature in post-processing (e.g., Mościbrodzka & Falcke, 2013; Mościbrodzka et al., 2014). This structure is *not* present in the models heated by the magnetic reconnection prescription, which does not deposit enough heat in the polar regions to allow the low-density fluid there to make a substantial contribution to the emission. Instead, the 43 GHz emission in these models is confined to the disk and the image is roughly circular when observed at 60°. Unlike the emergence of the black hole shadow in the optically thin synchrotron emission around 230 GHz, the 'disk-jet' structure at low frequencies is not a generic prediction of all GRMHD models; it is dependent on the choice of heating prescription.

One can compare the image size predictions from these models with interferometric measurements of Sgr A* made over the same frequency range. However, these comparisons are complicated by the effects of strong interstellar scattering for the line of sight to Sgr A*, with angular broadening from scattering dominating over intrinsic structure at wavelengths longer than a few millimeters. While many authors have inferred the intrinsic size of Sgr A* at millimeter and centimeter wavelengths by deconvolving these scattering effects, uncertainties in the scattering kernel render these estimates highly uncertain (see e.g., Psaltis et al., 2015). The most secure image size estimates are those made with the EHT at 1.3 mm, where the scattering effects are minimal, but these suffer from the additional limitation of extremely sparse baseline coverage. While early estimates of the source size found a FWHM of approximately 40 μas from a direct Gaussian fit to the data, more recent data have found visibility amplitudes on shorter baselines that are discrepant from this Gaussian fit

(Johnson et al., 2015; Lu et al., 2018). The appropriate representative image size at 1.3 mm should be estimated by taking the second moment of models that fit the short- and intermediate-baseline data. Johnson et al. (2018) gives $50 - 62\mu$as as a representative range of image sizes along the East-West direction, as constrained by current EHT data.

Figure 2.10 shows average image sizes fit to the time-averaged images at 230 GHz for the four models as a function of inclination angle. The image sizes are plotted as bands marking the range of values between the fitted major axis FWHM and minor axis FWHM at each inclination angle. In all model images at 1.3 mm, the image size grows with decreasing inclination angle as Doppler beaming becomes less significant. All four models satisfy the EHT constraint for at least one inclination angle. While models R-Lo, H-Lo, and H-Hi produce sizes consistent with observations at inclinations higher than 45°, model R-Hi produces the smallest images and only falls within the measurements at nearly face-on inclination.

2.3 Discussion

2.3.1 Comparison to Ressler et al. 2017

Ressler et al. (2017) presented the first 3D GRMHD simulation of Sgr A* with two-temperature electron-ion thermodynamics. They used a black hole spin of $a = 0.5$ and the Howes (2010) turbulent cascade prescription to heat the electrons. KORAL's simulation method used in this chapter differs from that of Ressler et al. (2017) in some notable ways. Their work includes the anisotropic conduction of heat along magnetic field lines, which KORAL ignores, although they report this conduction has little effect on the spectrum and image of Sgr A*. On the other hand, Ressler et al. (2017) ignore the radiative cooling of electrons and Coulomb coupling of electrons to ions, while KORAL includes both. Again, these effects are mostly unimportant for very low accretion rate sys-

tems like Sgr A*. Note that radiative cooling in particular will become significant in systems with higher accretion rates $\gtrsim 10^{-6}\,\dot{M}_{\rm Edd}$ like M87 (Ryan et al. 2018; Chael et al. 2019b; 3).

Notably, Ressler et al. (2017) used a fixed adiabatic index $\Gamma_{\rm gas} = 5/3$ in evolving the total gas, from which the dissipation is identified. In a separate post-processing step they set $\Gamma_{\rm e} = 4/3$ in evolving the electrons and estimating their temperature. In the trans-relativistic regime of the accretion flow in Sgr A*, electrons transition from non-relativistic ($\Theta_{\rm e} < 1, \Gamma_{\rm e} \approx 5/3$) at large radii to relativistic ($\Theta_{\rm e} > 1, \Gamma_{\rm e} \approx 4/3$) at radii close to the black hole and in the outflow. As a result, the effective adiabatic index of the total gas (Equation 1.26) will not be fixed at 5/3, even if electrons are cooler than ions or have less than 50% of the thermal energy (see Figure 2.2). Changes in the effective adiabatic indices in different regions of the simulation will affect the thermodynamics and the amount of dissipation identified in the numerical evolution. For instance, in the simple analytic shock test presented in Ressler et al. (2015), a gas with an adiabatic index $\Gamma_{\rm gas} < 5/3$ produces more dissipation and heats electrons to higher temperatures than if $\Gamma_{\rm gas}$ is fixed to 5/3. This difference could be important, especially in the jet region where $\Gamma_{\rm gas}$ is well below 5/3. The different treatment of the species and total gas adiabatic indices presented in Ressler et al. (2015, 2017) versus Sądowski et al. (2017) and the present work, and the different effects on the amount of dissipation identified between the treatments, deserves further study, particularly in more magnetized systems (see Section 2.3.2).

Despite the various differences in approach as described above, the picture from the low and high spin turbulent cascade models (H-Lo, H-Hi) in the present work and the model presented in Ressler et al. (2017) is largely consistent. All three models produce similar correlated near-infrared and X-ray variability from synchrotron self-Compton, and all obtain a spectrum at millimeter wavelengths that is consistent with observations. All these models show more variability at 230 GHz than the approximately 20% observed. In the 230 GHz images, both Ressler et al. (2017) and this chapter's

turbulent cascade models show a pronounced photon ring and a 'disk-jet' structure, where lower frequency emission is dominated by a jet or polar outflow. In this outflow, electrons are heated to high temperatures due to the strong $\beta_{\rm i}$-dependence of the Howes (2010) turbulent heating prescription.

Ressler et al. (2017) note that using the Howes (2010) turbulent heating prescription self-consistently produces the 'disk-jet' morphology that had been invoked in previous studies (e.g., Falcke & Biermann, 1995; Yuan et al., 2002; Mościbrodzka et al., 2014; Chan et al., 2015a) to explain the low-frequency Sgr A* spectrum. The main strength of these disk-jet phenomenological models is in reproducing the radio spectrum with an isothermal jet. Earlier works have produced a 'disk-jet' structure by setting the electron temperature manually in post-processing. For instance, both Mościbrodzka et al. (2014) and Chan et al. (2015a) identify 'jet' and 'disk' regions in their single temperature GRMHD simulations based on some criteria and then apply a constant $T_{\rm e}$ in the jet and a constant ratio $T_{\rm e}/T_{\rm i}$ elsewhere. Although a jet is visible in images at frequencies < 230 GHz in models heated by the Howes (2010) turbulent heating prescription, none of the self-consistent thermodynamic models presented in Ressler et al. (2017) or the present chapter reproduce an *isothermal* outflow.

This work shows that under a different physically motivated heating prescription, the disk-jet structure vanishes, at least in the thermal emission. Thus, a 'disk-jet' morphology is not a guaranteed outcome of simulations of Sgr A* with self-consistent electron heating. The form of the heating is important in determining the image shape and evolution with wavelength. Even when the 'disk-jet' structure is present, it remains unclear how to provide the jet with the additional heating needed for it to remain isothermal and reproduce the observed flat radio spectrum at $\nu < 10^{11}$ Hz.

2.3.2 Disk magnetization

A key difference between the spectra of the simulations presented here which use the Howes (2010) turbulent heating prescription and the spectrum presented by Ressler et al. (2017) is that their model produces substantially more near-infrared synchrotron emission, and meets (or even exceeds) measurements of the quiescent near-infrared and X-ray emission. However, they consider a disk that is substantially more magnetized. The dimensionless magnetic flux $\Phi_{\rm BH}/(\dot{M}c)^{1/2}r_{\rm g} \approx 40$ in their model, close to the MAD saturation value of ≈50 (Tchekhovskoy et al., 2011). In contrast, this chapter's models all have $\Phi_{\rm BH}/(\dot{M}c)^{1/2}r_{\rm g} < 10$ (see Table 2.2).

As a result, when compared to models H-Lo and H-Hi, Ressler et al. (2017)'s simulation has much lower $\beta_{\rm i}$ and a much higher $\sigma_{\rm i}$ in the outflow and close to the black hole. Whereas $\sigma_{\rm i} > 1$ regions exist at only the innermost radii in the high spin model H-Hi, Ressler et al. (2017) find large $\sigma_{\rm i} > 1$ in a substantial part of the outflow close to the axis (though they exclude this region from their radiative transfer). In their model, $\beta_{\rm i}$ averaged over the inner $25\,r_{\rm g}$ drops below 0.1 at polar angles $< 30^\circ$ and $> 150^\circ$, while in this work there is similar behavior at $5\,r_{\rm g}$ only in the spin 0.9375 model. Consequently, the heating rate $\delta_{\rm e}$ in their model is greater at a given radius and polar angle than in simulations H-Lo and H-Hi. Temperatures in their model reach $\Theta_{\rm e} \approx 100$, whereas the maximum $\Theta_{\rm e}$ in this work is 30, in model H-Hi (see Figure 2.3).

This combination of hot electrons and strong magnetic fields in the inner disk and outflow combine to produce more near-infrared synchrotron in the Ressler et al. (2017) simulation, and the median spectrum presented in their work goes though the measured quiescent values from Sgr A*. However, their simulation also fails to reproduce both the measured near-infrared flare spectral slope ($\nu L_\nu \propto \nu^{0.3}$) and the large observed flare amplitudes in both the near-infrared and X-ray. In fact, the normalized variability in their models is quite similar to model H-Lo in the

millimeter, near-infrared, and X-ray. In the 230 GHz emission, both model H-Lo and their model show variability amplitudes on the order of 40% relative to the mean, which is significantly larger than the root-mean-square range of of 20% reported by Marrone et al. (2008). In the near-infrared and X-ray, all excursions are contained within one order of magnitude from the mean and no strong flares are generated.

When the near-infrared quiescent emission in Ressler et al. (2017)'s simulation is inverse Compton upscattered to X-ray frequencies, it results in more quiescent X-ray emission than in any of this chapter's models, at the upper limit of the quiescent range of 10 – 100% of the Baganoff et al. (2003) value. This is despite the fact that Ressler et al. (2017) do not include bremsstrahlung emission in their radiative transfer. It seems likely that if bremsstrahlung were included, their model would overpredict the total measured Sgr A* quiescent X-ray emission. As the turbulent heating prescription puts nearly 100% of the energy into electrons in the jet and close to the black hole, at higher disk magnetizations the gas adiabatic index $\Gamma_{\rm gas}$ will become closer to 4/3 than 5/3 in a substantial part of the accretion flow. In this regime, the self-consistent treatment of the adiabatic index $\Gamma_{\rm gas}$ used in `KORAL` could become important and lead to different results for the jet luminosity and spectrum from those reported in Ressler et al. (2017).

2.3.3 The need for a nonthermal population

The four models presented in this chapter all produce spectra that match observations of Sgr A* at frequencies near the synchrotron peak around 10^{11}–10^{12} Hz (Figure 2.4). In addition, they all produce 230 GHz images consistent with the size measured by the EHT over some range of inclination angle (Figure 2.10). However, none of these models reproduce the characteristic large-amplitude X-ray flares observed ∼daily from Sgr A*, and they all underpredict the quiescent near-infrared emission. In addition, they do not show bright infrared flares with hard spectra, and

they fail to reproduce the low-frequency radio spectral slope.

While an isothermal jet or outflow can fit the low-frequency radio data (Yuan et al., 2002; Mościbrodzka et al., 2014; Chan et al., 2015a), a high-energy nonthermal electron population is another potential solution (Özel et al., 2000; Yuan et al., 2003). Recently, Davelaar et al. (2018) have applied a hybrid nonthermal-thermal κ distribution function in the jet in postprocessing Sgr A* GRMHD simulations. They show that nonthermal electrons in the jet can match both the relatively flat low frequency spectrum and the measured near-infrared spectral index. They effectively recover the 'disk-jet' model, but light up the jet with nonthermal particles instead of hot electrons at a constant temperature.

No thermal-only model has successfully reproduced the observed infrared variability or X-ray flares from Sgr A*. Chan et al. (2015b), Ressler et al. (2017), and the present work all reproduce the observed qualitative behavior whereby spikes in the X-ray always correspond to a near-infrared event. This behavior is a natural result of synchrotron self-Compton, whereby the X-ray flares are generated by upscattering near-infrared synchrotron photons. However, neither this chapter nor previous works have successfully reproduced the large flare amplitudes observed in the X-ray and near-infrared, nor the positive νL_ν power-law slope measured in the near-infrared (Genzel et al., 2003; Gillessen et al., 2006; Hornstein et al., 2007). The positive spectral index is a particularly important clue pointing toward nonthermal electrons as the source of Sgr A* flares, as no thermal synchrotron model that peaks in the submillimeter can produce a positive spectral index in the near-infrared. Furthermore, Marrone et al. (2008), Dodds-Eden et al. (2009), and Ponti et al. (2017) report a spectral index difference of ≈ 0.5 between the X-ray and near-infrared, suggestive of a synchrotron cooling break between the near-infrared and X-ray.

The large amplitudes of the observed near-infrared and X-ray flares again point to nonthermal electrons. Ball et al. (2016) demonstrated that inserting localized patches of nonthermal electrons

in post-processing can produce strong X-ray flares of greater than 10 times the quiescent value. Li et al. (2017) used an analytic MHD model to show that magnetic reconnection of flux ropes powering the acceleration of nonthermal electrons can reproduce the main features of near-infrared and X-ray flares from nonthermal synchrotron radiation. Cooling nonthermal electrons in a strong magnetic field also provides an alternative explanation to synchrotron self-Compton for both the observed correlations between X-ray and near-infrared flares and the shorter lifetimes of the X-ray flares (Kusunose & Takahara, 2011). To properly explore the signatures of nonthermal emission, one should thus include nonthermal particle acceleration and self-consistent evolution in the GRMHD simulation. The method of Chael et al. (2017) developed in Chapter 4 is well-suited to this end.

2.4 Summary and conclusions

In the four simulations considered in this chapter, the underlying heating prescription that models the plasma microphysics around Sgr A* has major effects on the properties of the accretion flow, as well as on the resulting simulated spectra and images. Under the turbulent cascade heating prescription, even though the simulations all have a relatively weak magnetic field, electrons are heated to very high temperatures in the funnel and are cooler in the disk. In contrast, the reconnection heating prescription heats electrons by nearly the same fraction everywhere (Figure 2.3). Energy is mostly radiated from the disk in the two simulations using the reconnection heating prescription, whereas with turbulent heating a significant amount of the radiation comes from the jet and outflow. This is particularly true in the high spin model H-Hi, which launches a mildly relativistic jet (Figure 2.2).

Once normalized to the 230 GHz flux density observed for Sgr A*, the spectra of all the four models match observations and are similar over the range 10^{11}–10^{12} Hz (Figure 2.4). However, none

of these thermal models can reproduce the low frequency radio spectrum nor the near-infrared flux density and spectral index. While the variability from synchrotron self-Compton produces a correlation between the near-infrared and X-ray that is qualitatively similar to the observed behavior, there are no large near-infrared or X-ray flares (Figure 2.6). Because more of their emission comes from the outflow and jet, the models heated by the turbulent cascade prescription are highly variable, and exceed the 20% level of root-mean-square variability measured for Sgr A*. The models heated by reconnection, on the other hand, all lie within the 20% variability bands at 230 GHz (Figure 2.5).

All four models produce 230 GHz images with distinct shadows and photon rings, and all models produce average 230 GHz images that are consistent with the size measured by the EHT over some range of inclination. Consistent with past studies, the turbulent heating prescription simulations produce images that are dominated by an outflow or jet at frequencies lower than 230 GHz. In contrast, neither simulation using the magnetic reconnection heating prescription produces a jet in the image at lower frequencies (Figures 2.8 and 2.9). Thus, while the transition of the synchrotron emission from optically thick to optically thin and the emergence of the black hole shadow around 230 GHz is a universal feature in all models of Sgr A*, a 'disk-jet' structure is not. It is sensitive to the choice of thermal electron heating prescription.

This chapter explores only weakly magnetized disks, and further simulations must be performed to compare different heating mechanisms in disks at or near the MAD limit. However, while more magnetized simulations may produce higher near-infrared and X-ray quiescent flux density, simply taking these thermal two-temperature simulation to greater magnetizations is unlikely to produce either the correct radio or near-infrared spectral indices or strong X-ray flares. Recent work in adding nonthermal electron distributions to GRMHD simulations in postprocessing (Ball et al., 2016; Davelaar et al., 2018) has supported earlier analytic work (Özel et al., 2000; Yuan

et al., 2003; Kusunose & Takahara, 2011) indicating that high-energy nonthermal populations are necessary to solve these remaining problems in modelling Sgr A*'s spectrum and variability. Electron acceleration to nonthermal energies is observed in particle-in-cell simulations of trans-relativistic reconnection (Werner et al., 2018; Ball et al., 2018). A future work will couple the self-consistent nonthermal electron evolution method developed in Chapter 4 (Chael et al., 2017) with physical models of relativistic, nonthermal electron acceleration to investigate the origin of Sgr A*'s variability and flares.

Page intentionally left blank

Text in this chapter was previously published in *MNRAS* 486 (2019), 2, pp 2873-2895 (A. Chael, R. Narayan, and M. Johnson).

3

Electron heating in MAD simulations of M87

Like Sgr A* (Chapter 2), the core of M87 is a LLAGN with a luminosity many orders of magnitude below the Eddington limit. Unlike Sgr A*, the SMBH in M87 launches a relativistic jet out to many kiloparsecs. Jets like that from M87 are likely powered by the black hole's rotational energy, which is extracted by ordered magnetic fields threading the black hole event horizon (Blandford & Znajek, 1977; Tchekhovskoy et al., 2011; Zamaninasab et al., 2014). These jets have been extensively investigated in GRMHD simulations. These simulations have demonstrated that jets powered by the black hole spin can be launched from thick disks and accelerated to high Lorentz factors (McKinney, 2006; Komissarov et al., 2007; McKinney & Blandford, 2009). For the specific case of M87, Dexter et al. (2012), Mościbrodzka et al. (2016a), and Mościbrodzka et al. (2017) have investigated the spectra and 230 GHz images predicted from various GRMHD simulations.

Recently, Ryan et al. (2018) carried out axisymmetric simulations of M87 with two-temperature evolution and frequency-dependent radiative transport. They found that, because M87 is more radiatively efficient than Sgr A*, including radiation in the simulation along with the temperature evolution of the electrons is critical. They performed simulations at both low ($3.3 \times 10^9\, M_\odot$; Walsh

et al., 2013) and high ($6.2\times10^9\,M_\odot$; Gebhardt et al., 2011) values of the SMBH mass and found that the accretion flow in the high mass, high spin case produced a 230 GHz image consistent with EHT observations published prior to 2019 (Doeleman et al., 2012; Akiyama et al., 2015). However, their simulations were performed with weak values of magnetic flux threading the horizon. As a result, the jets in their simulations had a narrower opening angle than that observed in VLBI images of M87, and the jet power was lower than the measured value by several orders of magnitude.

Magnetically Arrested Disks (MADs; Bisnovatyi-Kogan & Ruzmaikin, 1976; Narayan et al., 2003) are accretion flows choked by magnetic pressure near the black hole. In GRMHD simulations (e.g., McKinney et al., 2012; Narayan et al., 2012; Sądowski et al., 2013b), MADs are seen to launch jets with wide opening angles and large jet powers. Both of these features of MAD jets are observed in M87; the jet power is large ($\sim10^{43}-10^{44}$ erg s^{-1} ; Reynolds et al., 1996; Owen et al., 2000; Stawarz et al., 2006; de Gasperin et al., 2012), and the jet is launched with a wide opening angle ($\sim55^\circ$ at 43 GHz ; Walker et al., 2018). A MAD model of M87 is thus an attractive target for simulations with electron-ion thermodynamics for comparison with observational data.

This chapter presents the results of two fully 3D, two-temperature GRRMHD simulations of Magnetically Arrested Disks around the black hole in M87 performed using the code KORAL, again comparing the Landau-damped turbulent cascade heating prescription from Howes (2010) with the magnetic reconnection prescription from Rowan et al. (2017). These simulations (originally presented in Chael et al. 2019b) are the first Magnetically Arrested Disks evolved with two-temperature electron-ion thermodynamics.

3.1 Simulations

3.1.1 Units

In both simulations presented in this chapter, the distance to M87 is fixed to $D = 16.7$ Mpc (Mei et al., 2007), the black hole mass is $6.2 \times 10^9 M_\odot$ (Gebhardt et al., 2011, ,when scaled for this distance). The dimensionless black hole spin in both simulations is set to $a = 0.9375$.

For this mass, the gravitational length scale of M87 is $r_g = GM/c^2 = 9.2 \times 10^{14}\,\mathrm{cm} = 61\,\mathrm{AU}$. The corresponding angular scale is $r_g/D = 3.7\,\mu\mathrm{as}$. The gravitational time-scale is $t_g = r_g/c = 3 \times 10^4\,\mathrm{s} = 8.5\,\mathrm{hr}$.

M87's Eddington luminosity is $L_{\mathrm{Edd}} = 7.8 \times 10^{47}$ erg s^{-1}. The Eddington accretion rate (Equation 0.3) is $\dot{M}_{\mathrm{Edd}} = L_{\mathrm{Edd}}/\eta c^2 = 77\,M_\odot\ \mathrm{yr}^{-1}$. For these simulations, the efficiency $\eta = 0.18$, as expected for a thin accretion disk with $a = 0.9375$ (Novikov & Thorne, 1973).

3.1.2 Simulation setup

The simulations of M87 in this chapter were performed in the Kerr metric using a modified Kerr-Schild coordinate grid that is exponential in radius and concentrates grid cells near the equator (see Appendix B). The resolution was $288 \times 224 \times 128$ cells in the r, θ, and ϕ directions, respectively, which well-resolves the magnetorotational instability (MRI) that enables accretion. To capture the evolution of the jet at large radii, the outer boundary of the simulation box is at $10^5\,r_g$.

As in the simulations presented in Chapter 2, the initial equilibrium gas torii used the model of Penna et al. (2013). To build up magnetic field to the point where the disk reaches the saturation value of magnetic flux and becomes magnetically arrested, the initial torus was threaded with a single weak ($\beta_{\mathrm{min}} = 100$) magnetic field loop centered around $r \approx 50\,r_g$. The initial energy in

electrons was set at one percent of the total gas energy, with the remainder in ions.

The simulations used outflowing boundary conditions at the inner and outer radial boundaries, and reflecting boundary conditions were imposed at the the polar axes. In the nearest two cells to the polar axis, the simulation controls numerical instability from fluid flow across the poles by replacing the value of u^θ with the value from the third cell at the end of each timestep.

In the jet region, high fluid velocities rapidly evacuate the funnel and cause the fluid density to drop without bound. In order to ensure the numerical stability of the simulations in this region, KORAL puts a global ceiling on the magnetization $\sigma_{\rm i}$, as measured in the zero angular momentum observer (ZAMO) frame (McKinney et al., 2012). In this frame, the fluid density is increased to bring the magnetization back to the chosen limit, $\sigma_{\rm i,max} = 100$.

One simulation (H10) used the Howes (2010) prescription for dividing viscous dissipation between electrons and ions (Equation 1.33). The other simulation (R17) used the magnetic reconnection prescription of Rowan et al. (2017) and Chael et al. (2018a) (Equation 1.37). The simulations first ran for $10^4\,t_{\rm g}$ in 3D. During this this time, both simulations formed a thick disk at small radii and accumulated magnetic flux on the black hole horizon that exceeds the MAD threshold of $\approx 50\sqrt{\dot{M}c}\,r_{\rm g}$ (Tchekhovskoy et al., 2011; McKinney et al., 2012). At this point, to ensure the 230 GHz flux density from the models matches the 0.98 ± 0.04 Jy of compact emission in M87 measured by the EHT in 2009 and 2012 (Doeleman et al., 2012; Akiyama et al., 2015), the gas density was rescaled by a factor of 1/4 and magnetic field by 1/2 (keeping the temperatures and magnetization fixed).

At this point, the simulations were run from the rescaling point for another $1000\,t_{\rm g}$ to allow the models to settle into a new equilibrium. The results in the following Sections were taken from the final $5000\,t_{\rm g}$, from $t = 11,000\,t_{\rm g}$ to $t = 16,000\,t_{\rm g}$. The absence of large secular evolution in the accretion rate and 230 GHz light curve as a function of time in both models (Figure 3.4) indicates

the system has likely settled into a new equilibrium after the rescaling by $t = 11,000\,t_{\rm g}$.

In both simulations, the region of inflow equilibrium inside which the fluid quantities should be reasonably converged is defined by finding where the characteristic accretion time $t_{\rm acc} = 5000\,t_{\rm g}$. At a given radius r, the accretion time-scale is (Narayan et al., 2012):

$$t_{\rm acc} = \frac{r}{|v_r|}, \tag{3.1}$$

where $v_r = u^r/u^t$ is the Boyer-Lindquist radial three-velocity.

Over the final $5000\,t_{\rm g}$ period, the inflow equilibrium region in the disk extends to $\approx 30\,r_{\rm g}$ (Figure 3.2). In the fast moving jet, the region of outflow equilibrium extends to $\approx 4700\,r_{\rm g}$ (Figure 3.5).

3.1.3 Reliability of emission from high magnetization regions

In highly magnetized regions ($\sigma_{\rm i} > 1$), the plasma dynamics and thermodynamics in GRMHD simulations become increasingly suspect. Because the magnetic field dominates the energy budget in these regions, small errors in the total energy evolution can induce large changes in the internal energy and plasma temperature. At high $\sigma_{\rm i} \gtrsim 100$, these errors typically lead the code to crash, as the implicit solver fails to converge on a solution for the internal energy density from the conserved quantities.

As discussed in Section 3.1.2, KORAL imposes a ceiling on the magnetization $\sigma_{\rm i,max} = 100$ to ensure numerical stability. This ceiling results in a constant injection of gas density in the innermost jet regions. Figure 3.1 illustrates this ceiling on $\sigma_{\rm i}$ by showing profiles of ρ and $\sigma_{\rm i}$ versus polar angle θ in the time- and azimuth-averaged simulation data. At each radius, the density levels off at a floor value inside the polar angle θ where $\sigma_{\rm i}$ hits the simulation ceiling $\sigma_{\rm i,max} = 100$. This leveling off is a numerical artifact, and therefore, radiation from these regions will be artificially intense. One

must not include these regions where floors are active in spectra and images generated from the simulations.

Even in regions where the magnetization is not so strong as to lead to numerical instabilities and the imposition of the density floor, however, it remains a worrying possibility that errors in the thermodynamic evolution may still build up to bias the simulation results. The degree of unreliability of the radiation from the plasma temperature in these regions ($100 > \sigma_{\rm i} > 1$) is more difficult to assess than in the regions where the density is obviously artificially high (Figure 3.1). This potential unreliability is a problem in all GRMHD simulations (not just two-temperature ones) and in nearly all disk configurations (not just MADs), but it is a particularly significant concern for the MAD simulations in this chapter.

In producing spectra and images from GRMHD simulations, it is standard practice to only consider regions less magnetized than some cutoff value, $\sigma_{\rm i} < \sigma_{\rm cut}$. In most cases, $\sigma_{\rm cut} = 1$ is chosen as a conservative cutoff, eliminating all radiation from all magnetically dominated regions. For non-MAD simulations (e.g., Ressler et al., 2015, 2017; Chael et al., 2018a; Ryan et al., 2018), this choice is unlikely to substantially affect the results, as the emission from regions $\sigma_{\rm i} > 1$ is not a significant component of the spectrum (see e.g., Ressler et al., 2017, Appendix C).

In the MAD simulations considered in this chapter, however, the primary features of interest – the jet and near-horizon region – are highly magnetized. Including at least some emission from the $\sigma_{\rm i} > 1$ regions may be necessary to compare these MAD simulations to observations. Consequently, this chapter sets $\sigma_{\rm cut} = 25$, a factor of four lower than the $\sigma_{\rm i} = 100$ ceiling where density floors are imposed. This choice eliminates radiation from the density floor regions where the density is not set by mass loading from the disk, and where the thermodynamic evolution is definitely unreliable and unstable (Figure 3.1). Section 3.2.5 explores the effects of this choice in detail. The spectra and images from these simulations are sensitive to the choice of $\sigma_{\rm cut}$, indicating that it

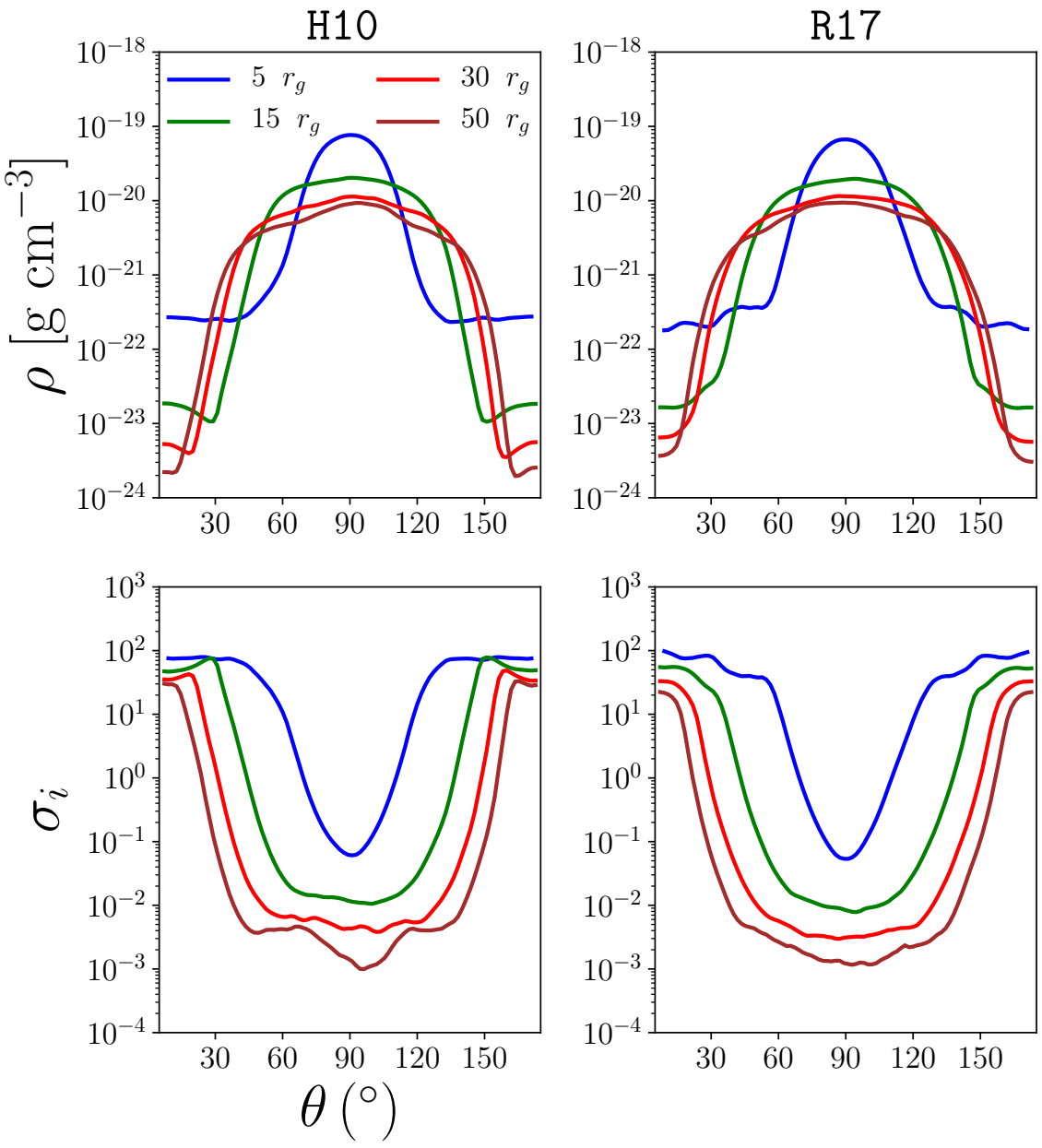


Figure 3.1: Azimuth and time-averaged density (top) and magnetization (bottom) as a function of polar angle θ for the two simulations at four radii: $r = 5\,r_{\rm g}$ (blue), $r = 15\,r_{\rm g}$ (green), $r = 30\,r_{\rm g}$ (red), and $r = 50\,r_{\rm g}$ (brown). Snapshot quantities were averaged in azimuth and then time-averaged from $11,000 - 16,000\,t_{\rm g}$. These data were not symmetrized over the equatorial plane. The ceiling on the magnetization $\sigma_{\rm i,max} = 100$ (imposed in the ZAMO frame) imprints itself as a floor on the density that takes effect at the same polar angle θ. Because the radiation produced in this region is unreliable, regions where $\sigma_{\rm i} > 25$ are excluded in the radiative transfer computations.

is an important free parameter in considering emission from MAD simulations. Future work on identifying the regions from where emission is reliable in highly-magnetized flows (as well as on more robust methods for evolving thermodynamics in these flows) will be critical in making firm conclusions in comparing images and spectra to data from these sources.

Finally, note that the potential unreliability of the thermodynamics outside the bulk of the disk is not even entirely confined to high $\sigma_{\rm i}$ regions. Along the jet wall, even in regions with $\sigma_{\rm i} < 1$, the density gradient is large and there is an effective contact discontinuity between the funnel/"corona" region and the disk. At this interface, the large entropy and density gradients are difficult for the Riemann solver to handle without substantial diffusion, which leads to non-negligible, time-averaged negative heating rates from Equation 1.30 (Ressler et al., 2017). This problem may be tractable with extremely high resolution simulations that can resolve this interface (now potentially feasible with GPUs, Liska et al. 2018), and with more advanced Riemann solvers than the Lax-Friedrichs solver typically used in GRMHD codes (e.g., the Harten-Lax-van-Leer-Discontinuities solver used in White et al., 2016).

3.1.4 Radiative transfer

As in the simulations of Sgr A* in Chapter 2, the spectra, images, and lightcurves from the M87 simulations in this chapter come from two post-processing codes, `HEROIC` and `grtrans`. Both codes used the value $\sigma_{\rm cut} = 25$ throughout (Section 3.1.3). Unlike in the Sgr A* simulations of Chapter 2, no additional density/magnetic field rescaling was imposed in postprocessing.

The jet inclination angle of M87 is constrained from observed "super-luminal" motion of jet components in VLBI images (Heinz & Begelman, 1997). This chapter assumes an inclination angle of 17° (Mertens et al., 2016; Walker et al., 2018). This angle is measured up from the lower pole, so that the sense of rotation of the accretion disk and black hole spin is clockwise on the sky. This

is the preferred orientation of the jet angular momentum vector as determined by the differential brightening and pattern velocities of the jet limbs in VLBI images (Walker et al., 2018). To match the orientation of the M87 jet on the sky at $-72°$ east of north (Reid et al., 1982), the computed images are rotated 108° counterclockwise.

3.2 Results

3.2.1 Accretion flow properties

Figures 3.2 and 3.3 show quantities averaged in azimuth and time over the time period $t = 11,000 - 16,000\, t_g$ after rescaling the density, internal energy, and magnetic field to match the 230 GHz flux density measured by the EHT in 2009 and 2012.

Figure 3.2 shows properties related to the thermodynamics of the accretion flow: the electron heating fraction δ_e, the gas temperature T_{gas}, the electron temperature T_e, the ion temperature T_i, and the temperature ratio T_e/T_i. Figure 3.3 displays the mass density ρ, the bulk Lorentz factor $\gamma = u^0/\sqrt{-g^{00}}$, the magnetization σ_i, the ratio of ion thermal pressure to magnetic pressure β_i, and the ratio of radiation pressure to gas pressure in the fluid frame $\beta_R = \hat{E}/3p$.

In each profile in Figures 3.2 and 3.3, the solid white contour shows the $\sigma_i = 1$ surface, while the dotted black contour shows the surface where the Bernoulli number Be $= 0.05$. Expressing $T^{\mu}{}_{\nu}$ in Boyer-Lindquist coordinates, the Bernoulli number is (Narayan et al., 2012; Sądowski et al., 2013b)

$$\mathrm{Be} = -\frac{T^t{}_t + R^t{}_t}{\rho u^t} - 1. \tag{3.2}$$

For a cold unmagnetized fluid, Be $= 0.05$ corresponds to a flow velocity of $\approx 0.3c$ at infinity.

From the leftmost panels of Figure 3.2, the different asymptotic behaviors of the two heating

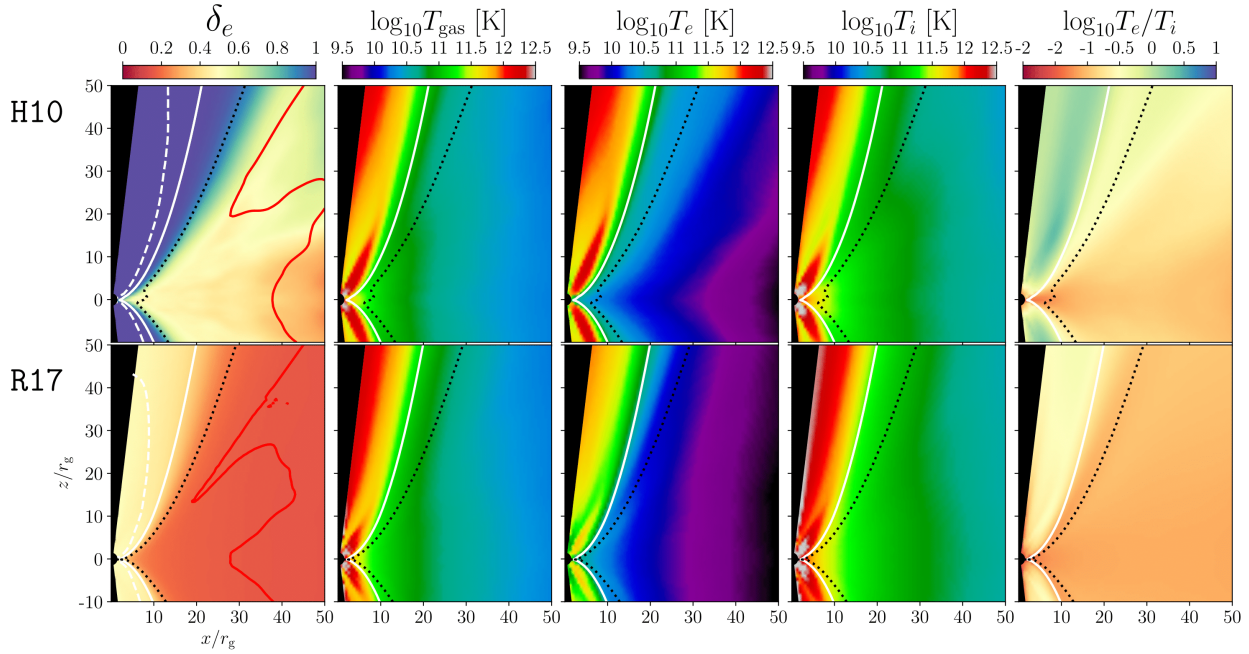


Figure 3.2: Time- and azimuth-averaged thermodynamic quantities from simulations `H10` (top) and `R17` bottom over the period $t = 11,000 - 16,000\,t_{\rm g}$. From left to right, the quantities shown are the electron heating fraction $\delta_{\rm e}$, the combined gas temperature $T_{\rm gas}$ in K, the electron temperature $T_{\rm e}$, the ion temperature $T_{\rm i}$, and the electron-to-ion temperature ratio $T_{\rm e}/T_{\rm i}$. The solid white contour in each panel denotes the surface where $\sigma_{\rm i}$=1, and the dashed black contour shows the surface where the Bernoulli parameter (Equation 3.2) $\mathrm{Be} = 0.05$, which this chapter takes as the definition of the jet-disk boundary. The solid red contour in the first column indicates the boundary of the inflow equilibrium region, defined such that $t_{\rm acc} = 5000\,t_{\rm g}$ (Equation 3.1). The dashed white contour in the first panel shows the $\sigma_{\rm i}$=25 surface; this is the maximum $\sigma_{\rm i}$ included in the radiative transfer (see Section 3.1.4).

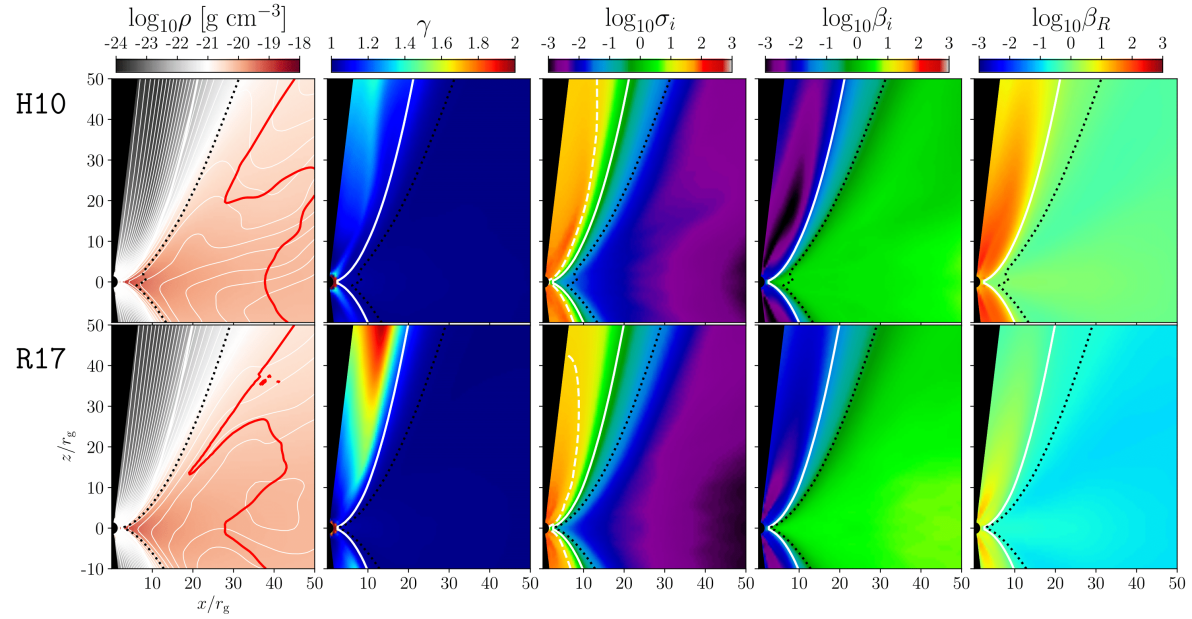


Figure 3.3: Additional time- and azimuth-averaged properties of the two simulations. From left to right, the quantities displayed are the density ρ in g cm^{-3}, the bulk Lorentz factor γ, the plasma magnetization $\sigma_{\rm i}$, the ratio of ion thermal pressure to magnetic pressure $\beta_{\rm i}$, and the ratio of radiation pressure to thermal pressure β_R. In the first column, white contours show the poloidal magnetic field in the averaged data.

prescriptions introduced in Sec. 1.4 are evident in the average values of δ_e in the jet region. In model `H10`, $\delta_e \approx 1$ everywhere inside the jet region defined by the Be $= 0.05$ contour. Outside the highly magnetized funnel, δ_e drops to nearly zero in the outer regions of the disk, $r \gtrsim 40\,r_g$. In contrast, in model `R17`, δ_e reaches its limit of equipartition of thermal energy ($\delta_e = 0.5$) inside the $\sigma_i = 1$ contour. In the less magnetized disk outside $r \gtrsim 15\,r_g$, δ_e falls to a small but nonzero value $\delta_e \approx 0.2$.

While the gas temperature distribution is similar in the two models (the second column of Figure 3.2), the different heating prescriptions result in different electron temperatures and temperature ratios in the inner disk and jet. In model `H10`, the deposit of nearly all thermal energy into electrons inside the jet results in high electron temperatures $T_e \sim 10^{11}$ K near the black hole that climb to $T_e \sim 10^{12}$ K in the jet around 50 r_g. While the temperature ratio T_e/T_i is less than unity in the regions closest to the black hole, it rises above unity by $r \approx 20\,r_g$ along the jet.

In contrast, in the magnetic reconnection heated model `R17`, $T_e/T_i < 1$ everywhere. In the jet around 30 r_g from the black hole, $T_e/T_i \approx 0.3$, and the ratio increases with radius, reaching 0.75 around 1000 r_g. In the disk, while the value of δ_e is higher than in the turbulent cascade heating simulation `H10`, the disk temperature ratio is not substantially different, with an average value of $\sim$0.1 around $30\,r_g$. This was not the case in the Sgr A* simulations presented in Chapter 2 (Chael et al., 2018a), where the turbulent cascade heated simulations had lower electron temperatures in the disk than the simulations heated by magnetic reconnection. The similarity of the outer disk electron temperatures in the present models may arise from the increased importance of Coulomb coupling in the denser regions of these higher accretion rate simulations (Ryan et al., 2018).

Figure 3.2 shows that, as in the simulations of Sgr A* in Chapter 2, the choice of electron heating prescription has a noticeable effect on the electron-ion thermodynamics of the system. Unlike in the low-magnetic-flux simulations considered in that work, however, the two MAD simulations

Table 3.1: Time-averaged quantities for both M87 simulations.

Model	$\langle \dot{M}/\dot{M}_{\rm Edd} \rangle$	$\langle \Phi_{\rm BH}/(\dot{M}c)^{1/2} r_{\rm g} \rangle$	$\langle P_J(100) \rangle$ [erg s^{-1}]	ϵ_J	$\langle P_{J,\rm rad}(100) \rangle$	$\epsilon_{J,\rm rad}$
H10	3.6×10^{-6}	55	6.6×10^{42}	0.5	8.8×10^{42}	0.7
R17	2.3×10^{-6}	63	1.3×10^{43}	1.6	1.4×10^{43}	1.6

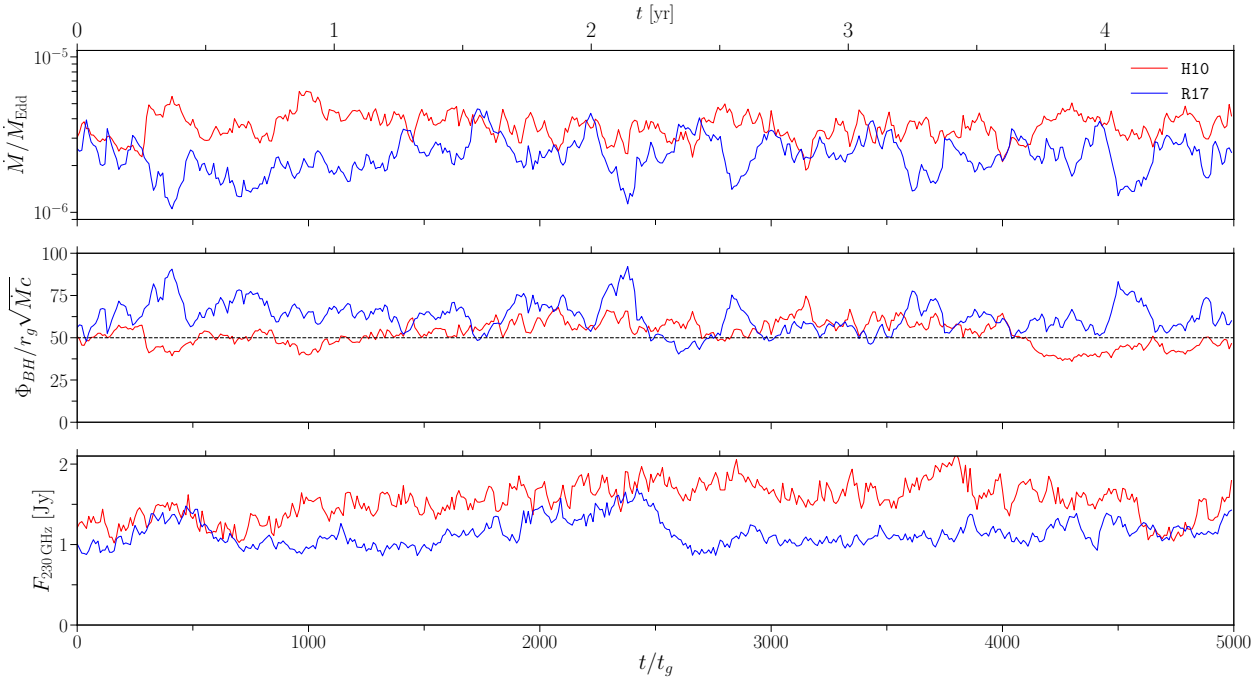


Figure 3.4: Time variability of the two MAD simulations H10 (red) and R17 (blue) plotted over the 5000 $t_{\rm g} = 4.84$ yr period from $11,000\,t_{\rm g}$ to $16,000\,t_{\rm g}$. (Top) Mass accretion rate $\dot{M}/\dot{M}_{\rm Edd}$. Both simulations show strong variability in the accretion rate as fluid parcels slip through the magnetic-pressure-dominated region close to the black hole horizon. (Middle) The dimensionless MAD parameter representing the amount of magnetic flux threading the black hole. Both models are well in the MAD regime, $\Phi_{\rm BH}/\sqrt{\dot{M}c r_{\rm g}} \gtrsim 50$ (McKinney et al., 2012), but the reconnection heated model R17 has a systematically higher magnetic flux on the horizon for most of the time considered and a correspondingly lower accretion rate, which is suppressed by the additional magnetic pressure. (Bottom) 230 GHz light curves computed from high-resolution grtrans images of the two simulations.

considered here have notably different gas and radiation kinematics as well, arising from the choice of heating prescription, even though both models produce thick (scale height-to-radius ratio $h/r \approx 0.41$ at $10\,r_{\rm g}$), highly magnetized disks.

Table 3.1 summarizes important time-averaged quantities from the simulations, and Figure 3.4 shows the accretion rate, magnetic flux through the horizon, and 230 GHz flux density as a function of time. Both simulations have a low accretion rate $\sim 10^{-6} \dot{M}_{\rm Edd}$. Both simulations also reach the MAD state, with the magnetic flux threading the black hole $> 50\sqrt{\dot{M}c r_{\rm g}}$, though R17 is slightly

more magnetized. In both simulations, while the averaged value of the accretion rate is stable, there are larger excursions with time than in the low-magnetic-flux simulations of Chapter 2 due to the interaction of the accretion flow and the magnetic flux. This is particularly apparent in simulation `R17`, which is more magnetized.

The high jet electron temperatures arising from the turbulent cascade heated model in simulation `H10` produce a much more intense radiation field from synchrotron and inverse Compton scattering in the jet region than in `R17`. This is easily seen in the last column of Figure 3.3, which shows the ratio of the radiation pressure to gas pressure as measured in `KORAL`. While this quantity is $\lesssim$1 almost everywhere in the `R17` simulation, it approaches values $\gtrsim$100 in the jet in model `H10`.

While `H10` produces more powerful synchrotron and inverse Compton radiation than `R17`, the optical depth to Compton scattering in both disks is low, with $\tau_{\rm IC} \sim 10^{-2}$. For the less magnetized models of Ryan et al. (2018), in contrast, the higher densities and temperatures needed to match the M87 spectrum with weaker magnetic fields made inverse Compton scattering more efficient, with an optical depth $\tau_{\rm IC} \sim 0.1 - 1$.

The conversion of much of model `H10`'s energy and momentum to radiation in the inner jet has a significant impact on its mechanical properties. While both simulations launch relativistic jets, `H10`'s jet is weaker, with Lorentz factor $\gamma \approx 1.5$ at $50\,r_{\rm g}$ compared to $\gamma \approx 2$ for `R17` (Figure 3.1, second column). Figure 3.5 shows the Lorentz factors of the two jets at large scales. By the jet equilibrium radius around $4700\,r_{\rm g}$, `H10` reaches a Lorentz factor of $\gamma \approx 3$, while `R17` reaches $\gamma \approx 5$. The conversion of fluid and magnetic energy into radiation also likely accounts for `H10`'s smaller mean value of horizon flux $\langle \Phi_{\rm BH} \rangle$ and correspondingly larger mass accretion rate $\dot{M}$ (Table 3.1).

At large radii, $r > 1000\,r_{\rm g}$, the simulations resolve the jet with $\approx$6 cells in polar angle out to the $\sigma_{\rm i} = 1$ contour (white lines in Figure 3.5). Higher resolution simulations will be necessary to test whether the observed jet opening angle in simulated images from these simulations is affected by

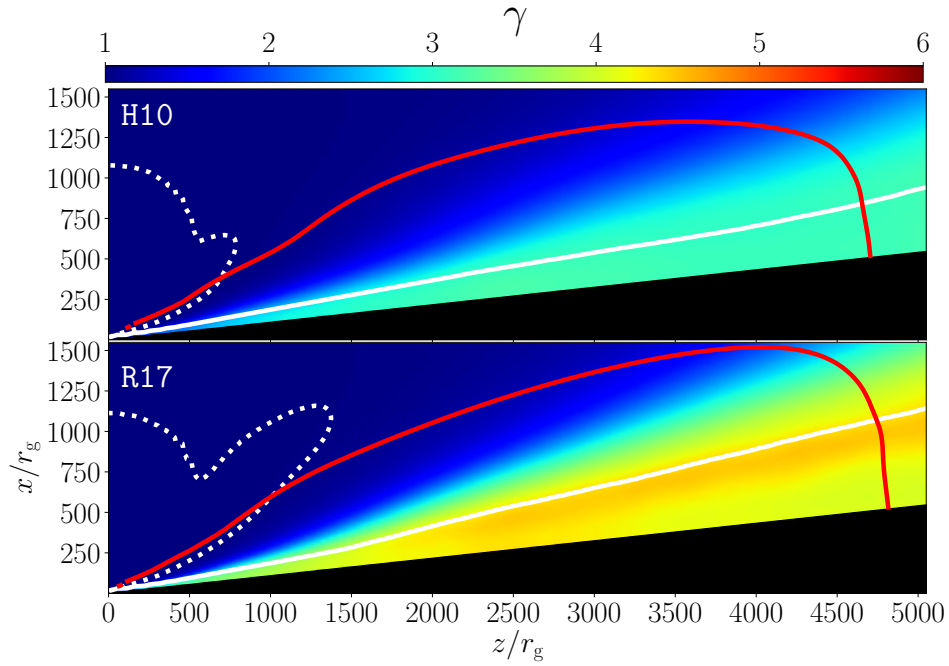


Figure 3.5: Large-scale jet Lorentz factors γ from the simulations `H10` (top) and `R17` (bottom). The solid white contour shows the surface where σ_i=1, and the dashed white contour shows where the Bernoulli parameter $\mathrm{Be} = 0.05$. The red contour indicates the extent of the region of inflow/outflow equilibrium, where $t_{\rm acc} = 5000\,t_g$ (Equation 3.1). The black region along the jet axis indicates the two nearest cells to the polar axis, which are removed from all analysis due to their boundary conditions (Section 3.1.2).

the simulation resolution. Furthermore, using reflective boundary conditions along the polar axis (Section 3.1.2) may enlarge the jet width; tests of similar MAD jets in Cartesian coordinates (e.g., Porth et al., 2017) or using a misaligned grid (e.g., Liska et al., 2018) are necessary to assess the dependence of apparent jet width on the numerical grid.

The jet power (including thermal, magnetic, and jet mechanical contributions) is (Tchekhovskoy et al., 2011; Ryan et al., 2018)

$$P_J = -\int \left(T^r{}_t + \rho u^r\right) \sqrt{-g}\,\mathrm{d}\theta\mathrm{d}\phi, \tag{3.3}$$

where the integral is at a fixed r is over the jet cap, defined by the criterion $\mathrm{Be} > 0.05$ (Narayan et al., 2012; Sądowski et al., 2013b). The time-averaged jet power measured by Equation 3.3 is roughly constant with radius from around $r = 10\,r_g$ out to $r = 1000\,r_g$. The average jet powers at

$100\,r_{\rm g}$ from the averaged data are $6.6 \times 10^{42}\,{\rm erg\ s^{-1}}$ for model `H10` and $1.3 \times 10^{43}\,{\rm erg\ s^{-1}}$ for `R17` (Table 3.1).

While the jet powers obtained from the two simulations agree to within a factor of two, the value obtained for model `R17` is more consistent with the measured values for M87 of $\sim 10^{43} - 10^{44} {\rm erg\ s^{-1}}$ (Reynolds et al., 1996; Stawarz et al., 2006). Comparing the jet power to the accretion rate gives a jet efficiency $\epsilon_J = P_J/\dot{M}c^2$ of 1.6 and 0.5 for `R17` and `H10`, respectively, indicating that spin energy is being extracted from the black hole, especially in model `R17`, which has greater than 100% efficiency (Tchekhovskoy et al., 2011).

Because much of `H10`'s energy and momentum is converted to radiation in the jet, it has a correspondingly lower mechanical jet power. Including radiation in the jet power measurement gives

$$P_{J,\rm rad} = -\int \left(T^r{}_t + R^r{}_t - \rho u^r\right) \sqrt{-g}\,{\rm d}\theta\,{\rm d}\phi. \tag{3.4}$$

Including radiation increases the measured jet powers to $P_{J,\rm rad} = 8.8 \times 10^{42}$ erg s^{-1} for `H10` and $P_{J,\rm rad} = 1.4 \times 10^{43}$ erg s^{-1} for `R17`; it increases the jet efficiencies in the two models to 0.7 and 1.6, respectively.

Much of the intense radiation in `H10` is produced from regions near the $\sigma_{\rm i,max} = 100$ ceiling. Because mass is constantly being injected in these regions, energy and momentum are added to the simulation and are then efficiently converted into radiation. As discussed in Section 3.1.4, one should not trust the radiation produced in this region, and these regions are excluded in the post-processing computation of spectra and images in the following sections. Figure 3.6 shows a comparison of the average radiation field from the `KORAL` output for model `R17` with the average radiation fields computed from `HEROIC`, both with no $\sigma_{\rm cut}$ imposed and using $\sigma_{\rm cut} = 25$. With no $\sigma_{\rm cut}$, both codes produce consistent results, with radial radiation streamlines and extremely high

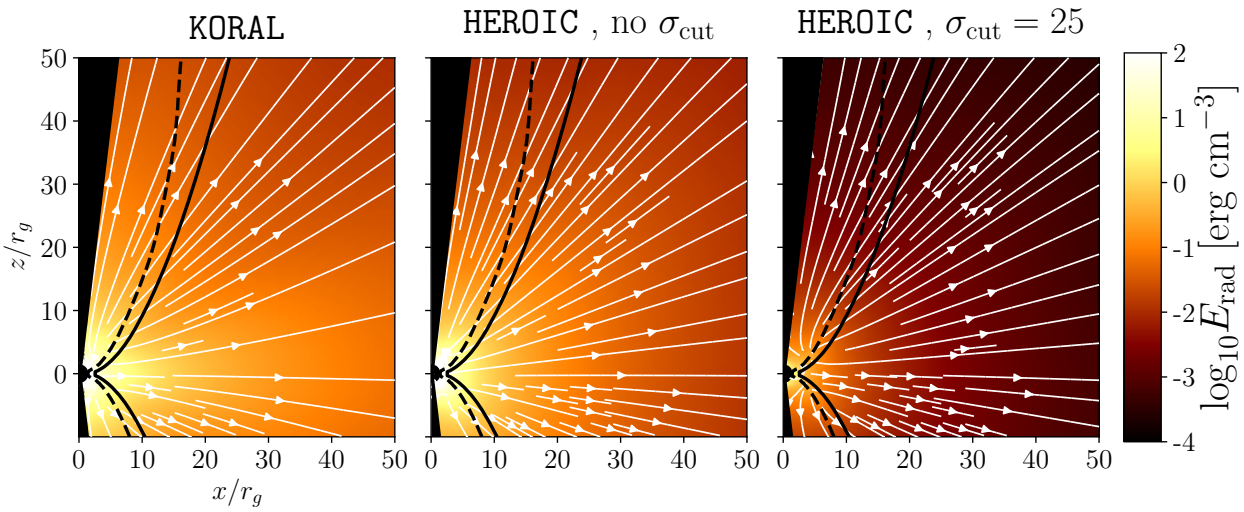


Figure 3.6: (Left) Lab frame radiation energy density and radiation flux streamlines from time- and azimuth- averaged data from simulation `H10`. The solid contour indicates the $\sigma_{\rm i} = 1$ surface, and the dashed contour indicates $\sigma_{\rm i} = \sigma_{\rm cut} = 25$. (Center) the same quantities computed in the postprocessing code `HEROIC`, with no cut on $\sigma_{\rm i}$ imposed. (Right) the radiation energy density and flux streamlines from `HEROIC` after zeroing emissivities from the regions $\sigma_{\rm i} > \sigma_{\rm cut}$. The `KORAL` and `HEROIC` results are in good agreement when no cut on radiation from $\sigma_{\rm i} > 25$ regions is included; both codes produce approximately radial radiation streamlines and extremely high luminosities originating from near the black hole. When the $\sigma_{\rm i}$ cut is imposed in the `HEROIC` post-processing, the average energy density in radiation drops by more than a factor of 10.

energy radiation energy densities. However, when emissivities from regions with $\sigma_{\rm i} > 25$ are zeroed out in `HEROIC` (as is done to produce the spectra and images in Sections 3.2.2, 3.2.3, and 3.2.4) the energy density of the radiation field everywhere drops by a factor of $\approx$50.

In the (optically thin) GRRMHD simulation itself, because the frequency-averaged radiation field produced in $\sigma_{\rm i} > 25$ regions spreads through the simulation volume at nearly the speed of light, it is difficult to extract meaningful radiation quantities that are unaffected by the density floors. For this reason, when interpreting the jet power and other quantities from these simulations, it is important emphasize again that the results may be strongly dependent on the specific choice of density floors and $\sigma_{\rm cut}$. More work is needed to better understand the impact of radiation from high $\sigma_{\rm i}$ regions on the global structure of optically thin accretion flows; future two-temperature MAD simulations should consider not including radiation from $\sigma_{\rm i} > \sigma_{\rm cut}$ regions during the simulation run itself.

3.2.2 Spectra

Figure 3.7 shows the spectral energy distributions (SEDs) from both simulations. These spectra were obtained from postprocessing with `HEROIC` over the full $5000\,t_g$ duration of the simulation from $11,000$-$16,000\,t_g$, beginning $1000\,t_g$ after rescaling the density to approximately match the compact 230 GHz flux density measured by Doeleman et al. (2012) and Akiyama et al. (2015). Synchrotron, free-free, and inverse Compton emission are all included in the `HEROIC` calculations, although free-free emission does not contribute significantly even in the X-ray. These spectra do not include radiation produced from regions with $\sigma_i > 25$, and consequently the total luminosity from the postprocessing spectra is less than that produced in the simulation bolometric $R^{\mu}{}_{\nu}$. In computing the spectra, `HEROIC` used data from the simulation out to $1000\,r_g$; diffuse emission from a good fraction of the jet is included, but the outermost regions of the jet are ignored.

SEDs from both models are largely consistent with the radio spectrum data up to the synchrotron peak at 230 GHz; this flat spectrum is also produced by analytic models of relativistic jets (Blandford & Königl, 1979; Falcke & Biermann, 1995). The model spectra underpredict the total measured flux density at frequencies $< 10^{10}$ Hz; at these low frequencies, the jet on scales larger than $1000\,r_g \approx 3600\mu$as likely makes substantial contributions to the total emission.

Neither SED matches the measured flux density at infrared through ultraviolet frequencies, although the hot jet electrons in the `H10` model do extend the thermal synchrotron spectrum to the $\approx 3 \times 10^{13}$ Hz measurements by Perlman et al. (2001) and Whysong & Antonucci (2004). The observed emission from the near-infrared to ultraviolet may be explained by the addition of a high-energy nonthermal electron population in the disk (Broderick & Loeb, 2009) or the jet (Dexter et al., 2012; Prieto et al., 2016).

The hotter electrons in the `H10` simulation produce more inverse Compton power at X-ray fre-

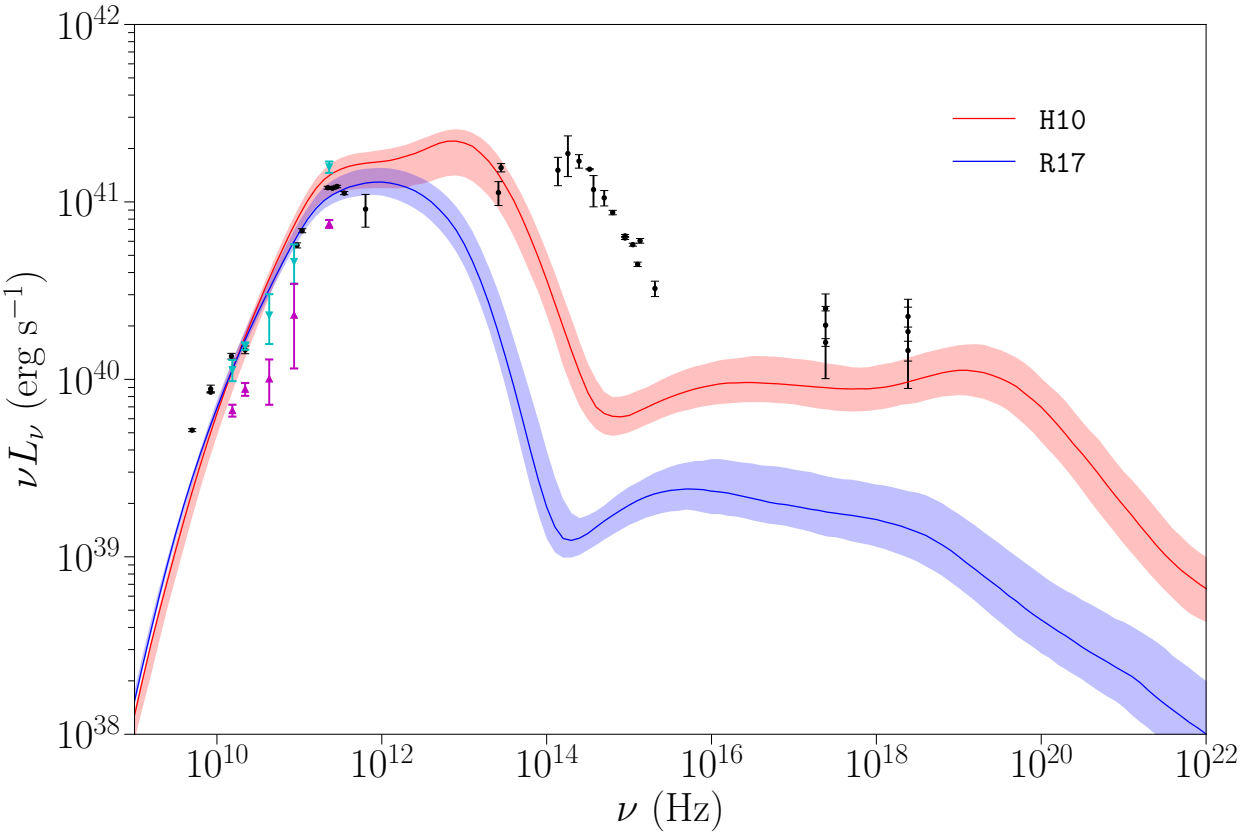


Figure 3.7: SEDs for the two models calculated with HEROIC for an observer at 17° inclination (Walker et al., 2018) measured up from the simulation south pole. Spectra were computed from 3D simulation snapshots every $10\,t_{\rm g}$ over the $5000\,t_{\rm g}$ period from $t = 11,000\,t_{\rm g}$ to $t = 16,000\,t_{\rm g}$ after rescaling the density to approximately match the 0.98 Jy flux density of compact emission at 230 GHz (Doeleman et al., 2012; Akiyama et al., 2015). The solid curve shows the median spectrum for each model, and the shaded region shows the nominal 1σ time-variability. Data points are taken from Table 1 of Prieto et al. (2016) (black) and Table A.1. Measurements of the total flux density at radio wavelengths from Table A.1 are displayed in cyan, while measurements of the compact flux density of the core are displayed in magenta.

quencies. Both models produce flat SEDs from Comptonization at X-ray frequencies, which roughly match the slope of the *Chandra* measurements of the core of M87 (Di Matteo et al., 2003), but both simulations underpredict the total flux density in the X-ray. However, the shape of the spectrum at frequencies $> 10^{12}$ Hz depends strongly on the choice of $\sigma_{\rm cut}$ (see Section 3.2.5). It is possible that a two-temperature MAD simulation that can be trusted up to higher values of $\sigma_{\rm i}$ could fit all of the spectral data up to the X-ray. Neither model's SED extends significantly into the γ-ray, where the observed emission may be dominated by the jet knots, such as HST-1 (Abramowski et al., 2012).

3.2.3 43 and 86 GHz images and core-shift

To compare the models against existing images and data from VLBI observations, `grtrans` images were computed at 15, 22, 43, 86, 230, and 345 GHz. These images only include emission from the jet out to $3000\,r_{\rm g}$. With the small inclination angle of 17° (Mertens et al., 2016; Walker et al., 2018), this maximum radius corresponds to a maximum *projected* jet length of $850\,r_{\rm g}$, or ≈3 mas. In contrast, jet emission in the 43 GHz VLBI image extends out to at least 20 mas (Walker et al., 2018). For each model, a representative snapshot was chosen where the core flux density at 230 GHz was close to the measured value of 0.98 Jy from the EHT in 2009 and 2012 (Doeleman et al., 2012; Akiyama et al., 2015).

Figure 3.8 shows log-scale images of both models at 43, 86, and 230 GHz, each with a dynamic range of 10^4. The jet structure is similar in both simulations, with a wide apparent opening angle that increases to $> 90°$ at the jet base in the 230 GHz image. The jets in both models show filamentary magnetic field structure close to the black hole that rotates clockwise as viewed from the selected orientation. The spiral filaments are more prominent in the `H10` snapshot. In general, emission in the `H10` model comes from the high-temperature, high-magnetic field inner jet, and magnetic filaments in the jet dominate over disk emission at 230 GHz (Figure 3.14). At longer

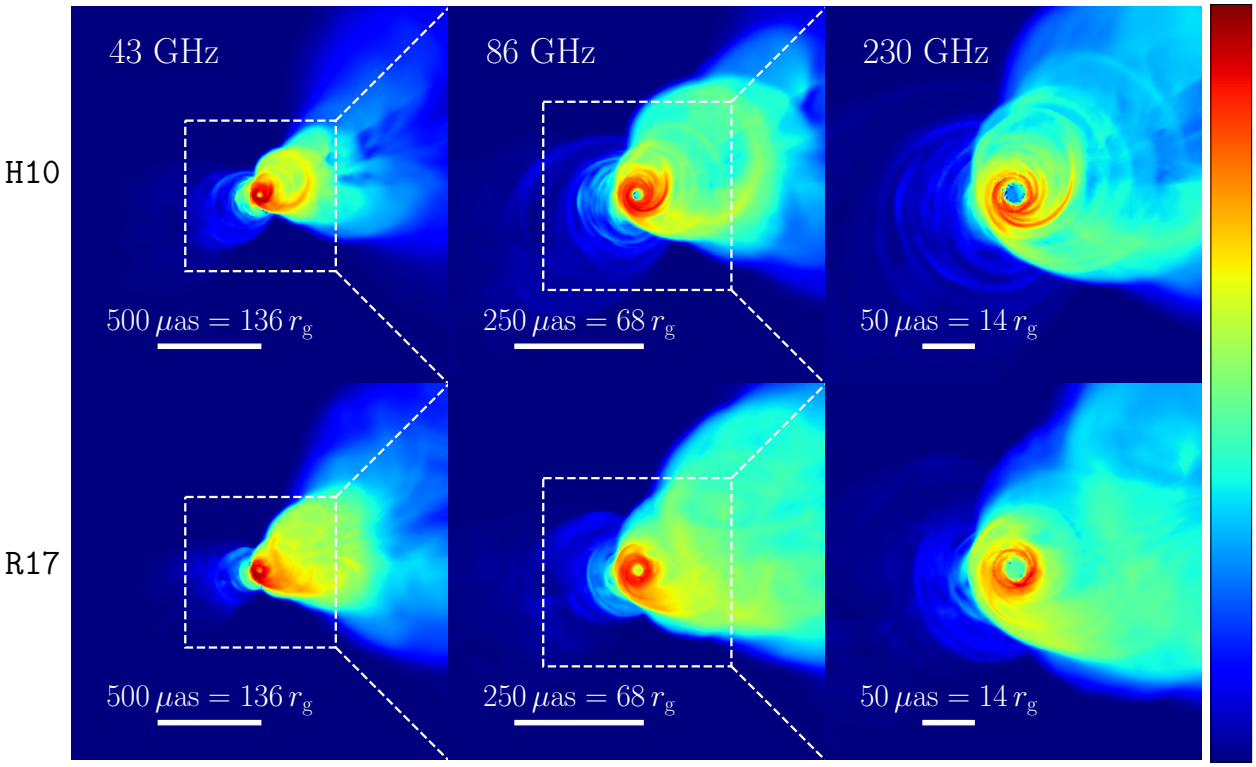


Figure 3.8: Log scale images of simulation snapshots of the two models at 43 GHz (left), 86 GHz (middle) and 230 GHz (right). Snapshots were observed at an inclination angle of 17° up from the simulation south pole and rotated 108° counterclockwise to match the observed jet orientation. The intensity scale is different at each frequency, but for each frequency the scale displays a dynamic range of 10^4 and is the same for both the image from the H10 simulation (top) and R17 simulation (bottom). The image length scale changes with frequency; dotted boxes on the 43 and 86 GHz images show the fields-of-view of the 86 and 230 GHz images, respectively. The jet structure is qualitatively similar in the two simulations, with wide apparent opening angles which narrow with distance from the SMBH at lower frequencies. Images at all frequencies show a faint counterjet, and the black hole shadow is evident in both models even down to 43 GHz.

wavelengths, the jet images produced from the two models are qualitatively similar. The higher jet power in the R17 model leads to brighter emission at larger distances from the black hole at all frequencies, while the core is consistently brighter in the H10 images.

VLBI images show that the jet of M87 is wide, with an apparent opening angle that decreases with radius. The apparent opening angle is ≈55° at 43 GHz (Walker et al., 2018), increasing up to ≈127° in the innermost regions ($\sim 50 r_g$) of 86 GHz images (Kim et al., 2018). Figure 3.9 compares the simulation images of the inner 2 mas at 43 GHz with an image of 2007 VLBA data (Walker et al., 2018) reconstructed with a new closure phase and amplitude imaging method implemented

in the `eht-imaging` software library (see Chapter 4 and Figure 9 of Chael et al. 2018b). The top row shows the snapshot images from both simulations, and the bottom row shows the VLBI image and the snapshot simulation images convolved with a beam that has a size half of the nominal value reported in Walker et al. (2018). Lines indicating the 55° apparent opening angle measured by ridge line analysis in Walker et al. (2018) are overlaid on all images. Previous simulations of M87 using weakly magnetized disks produced narrow jets (Mościbrodzka et al., 2016a; Ryan et al., 2018), with narrow opening angles $\lesssim$30°. In contrast, the observed wide apparent opening angle is naturally produced in these MAD simulations. The simulation images blurred to the same resolution as the VLBI image also show limb brightening in the jet, though the contrast between intensity on the jet edges and along the axis is in general less prominent than in the VLBI image. The counterjet is faint but visible at the edge of the dynamic range in both model images, but it is more prominent in the VLBI image at this epoch.

Figure 3.10 compares the simulation images at 86 GHz with the 2014 image reconstructed from GMVA observations in Kim et al. (2018) (their figure 3, panel d). Again, both the `H10` and `R17` simulations produce wide opening angle jets at 86 GHz consistent with the Kim et al. (2018) image. At these frequencies, the opening angle of `R17` is slightly larger than that of `H10`, likely due to the fact that this simulation is even more MAD (Table 3.1). Like the GMVA image, the 86 GHz model images also show noticeable counterjet emission. However, the limb brighting in the blurred simulation images is somewhat less than in the VLBI image.

VLBI observations with absolute phase referencing allow for estimates of the relative image location at different frequencies. Because the M87 jet is optically thick at wavelengths longer than a few millimeters, the image centroid of the bright, compact core emission moves with frequency, giving rise to the so-called "core-shift" effect (Blandford & Königl, 1979).

Hada et al. (2011) conducted measurements of the core-shift of M87 at frequencies from 2.3 to

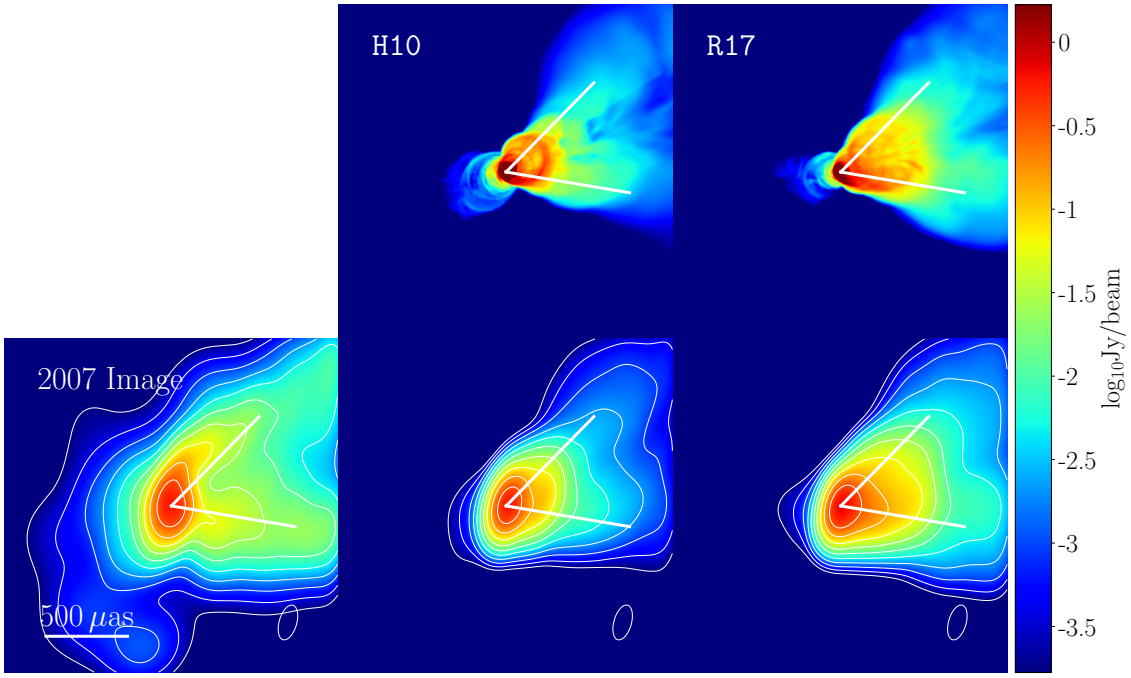


Figure 3.9: Log scale 43 GHz images of snapshots from the two models overlaid with the measured $55°$ apparent opening angle (Walker et al., 2018). The leftmost panel shows an image reconstructed from 2007 VLBA data (Walker et al., 2018) using the `eht-imaging` library (Figure 9 of Chael et al., 2018b). The image is convolved with a Gaussian beam half the size of the nominal beam reported in Walker et al. (2018). The top row shows the high-resolution `grtrans` images of the two simulations at 43 GHz, and the bottom row shows the simulation images convolved with the same beam as the 2007 reconstruction.

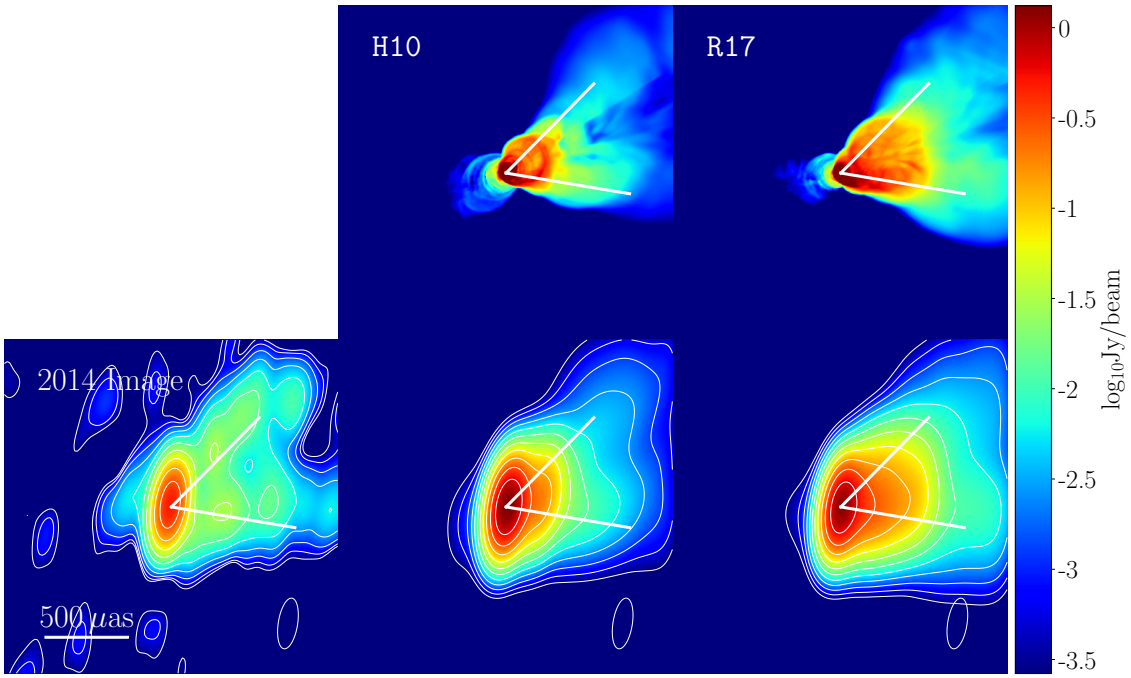


Figure 3.10: Log scale images of 86 GHz snapshots of the two models overlaid with the measured $55°$ apparent opening angle. The leftmost panel shows an image from GMVA observations reported in Kim et al. (2018). The top row shows the high-resolution `grtrans` images of the two simulations at 86 GHz The bottom row shows the simulation images convolved with the Gaussian beam reported in Kim et al. (2018).

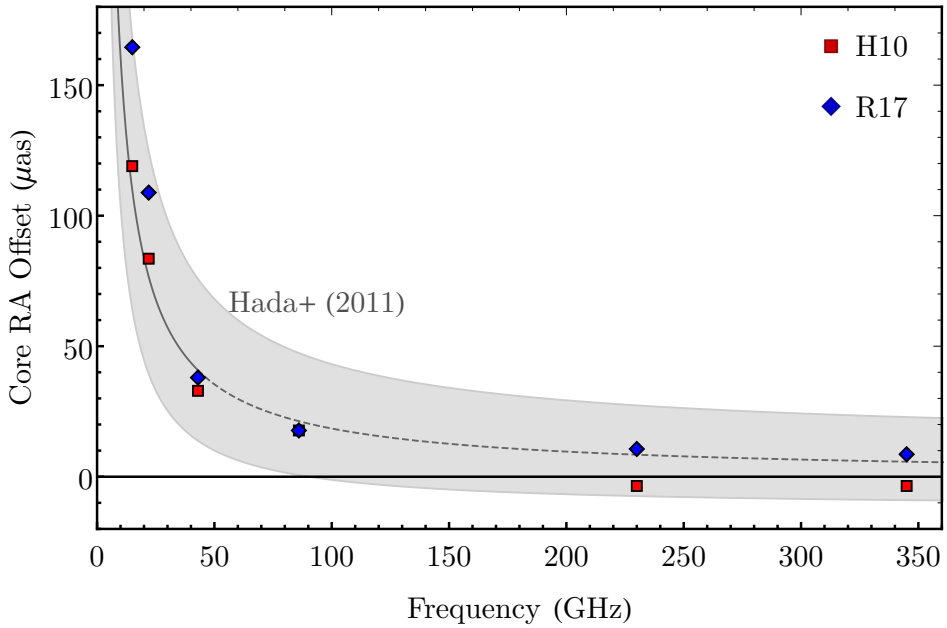


Figure 3.11: Frequency dependent core-shift for simulations `H10` and `R17` (see Section 3.2.3). The gray line and shaded region show the best-fit model and 1σ confidence region from Hada et al. (2011) using measurements from 2 - 43 GHz (here, the central model is re-referenced to a core-shift of zero at $\nu \to \infty$). All points shown are compatible with the VLBI measurements over this frequency range. The extrapolated model and corresponding simulation estimates at higher frequencies must be interpreted with caution because the images are no longer dominated by a bright, optically-thick core.

43 GHz, finding that the millimeter core is coincident with the SMBH and disk that launch the jet. They estimated that the radio core has a right ascension displacement (relative to the 43 GHz core) given by $\Delta\mathrm{RA} \approx A\lambda_{\mathrm{GHz}}^{-\alpha} + B$, where $A = (1.40 \pm 0.16)\,\mathrm{mas}$, $B = (-0.041 \pm 0.012)\,\mathrm{mas}$, and $\alpha = 0.94 \pm 0.09$. The analogous core-shifts from the simulated images in this chapter were computed by first convolving the images with the wavelength-dependent observing beam of Hada et al. (2011) and then measuring the location of the peak in the resulting image.

Figure 3.11 shows the results of this analysis. Both `H10` and `R17` produce images with core-shifts that are compatible with the results of Hada et al. (2011) at frequencies as low as 15 GHz. Even though VLBI constrains the core-shift at yet lower frequencies, estimating core-shifts at these frequencies is difficult, because the image sizes and observing beams below 15 GHz are comparable to the raytracing domain used in `grtrans`.

Radio-jet core-shifts can be used to measure the jet magnetic field. Recently, Zamaninasab et al.

(2014) and Zdziarski et al. (2015) used core-shifts to measure the jet magnetic fields of several LLAGN sources (including M87) on ∼pc scales; they found their values of magnetic field and jet powers were consistent with jets launched by MADs. This is consistent with the findings of this chapter for M87.

The magnetic field strength in the core of M87 can also be estimated using the measured core sizes from VLBI. Kino et al. (2014) used this method with a model of the optically thick 43 GHz synchrotron emission to estimate a field strength 1 G $\lesssim |B| \lesssim$ 15 G at an angular scale of 100 μas. Assuming an inclination angle of 17°, an angular scale of 100μas corresponds to a de-projected distance $r \approx 100 r_{\rm g}$. From the time- and azimuth-averaged simulation data, both models H10 and R17 have 1 G $\lesssim |B|_{100\,r_{\rm g}} \lesssim$ 2 G, within the measured range. Closer to the black hole, Kino et al. (2015) used EHT data to estimate a field strength 58 G $\lesssim |B| \lesssim$ 127 G on scales $\sim 10\mu$as, corresponding to a de-projected distance of $\approx 10\, r_{\rm g}$. From the averaged simulation data, $|B|_{10 r_{\rm g}} \approx$ 20 G at this radius in the jet. However, Kino et al. (2015) obtain their estimate by assuming a spherical 230 GHz emission region that is optically thick to synchrotron self-absorption, which is not observed at 230 GHz in these simulations.

3.2.4 230 GHz images

Prior to 2019, the EHT had already constrained the 1.3 mm emission size at the core of M87 to be on the order of ∼40 μas (Doeleman et al., 2012; Akiyama et al., 2015), approximately the size of the lensed black hole shadow for $M = 6.2 \times 10^9\, M_\odot$ (Gebhardt et al., 2011) and $D = 16.7$ Mpc (Mei et al., 2007). After the initial publication of this work in (Chael et al., 2019b), the EHT produced the first image of the black hole shadow in M87 from observations conducted with a full array in 2017 (The Event Horizon Telescope Collaboration et al., 2019d).

Figure 3.12 compares three linear scale images at 230 GHz showing time evolution of the source

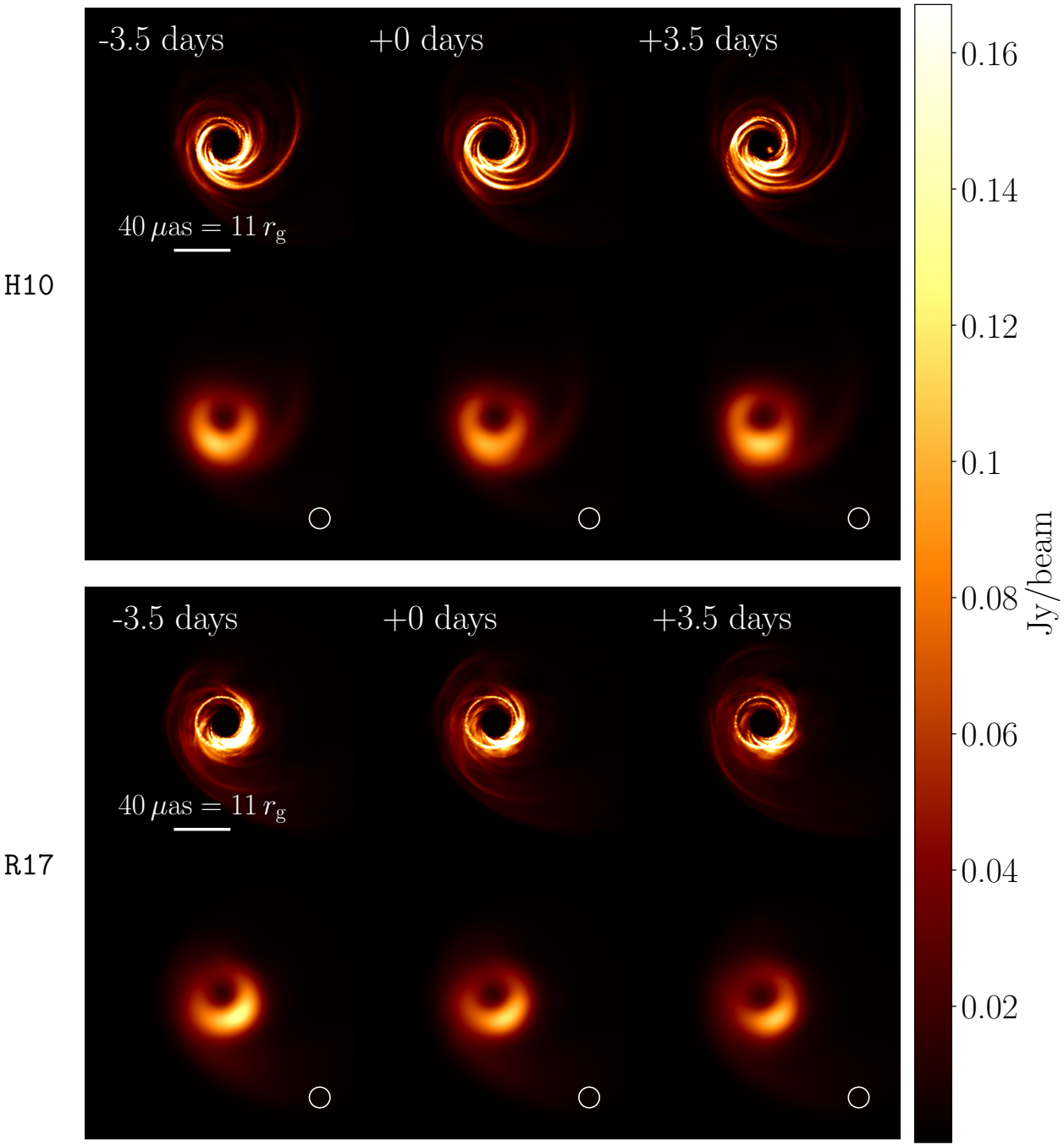


Figure 3.12: Linear scale images of sequential snapshots at 230 GHz from the two models showing time evolution over $20\,t_{\rm g} \approx 7\,{\rm days}$, the usual length of an EHT observing campaign. The top two rows show images from simulation `H10`, and the bottom two rows show images from simulation `R17`. In each set of images, the top row shows high-resolution images from `grtrans`, and the bottom row shows the same images blurred with a circular Gaussian beam with a $15\,\mu$as FWHM, approximately representing the imaging resolution of the EHT.

over roughly 1 week, the usual length of observation campaigns by the EHT. In each set, the top row shows high-resolution images from `grtrans`, and the bottom row shows the same images blurred with a Gaussian beam with a 15 μas full width at half maximum (FWHM). This beam size is approximately the imaging resolution of the EHT (Chael et al., 2018b).

The 230 GHz images from both models show a distinct black hole shadow that is large enough to be imaged by the EHT. The diameter of the shadow is approximately $9.7\,r_{\rm g} \approx 36\,\mu$as (slightly less than the diameter for a Schwarzschild black hole, $2\sqrt{27}r_{\rm g} \approx 38\,\mu$as), given the assumed mass and distance. Because the accretion flow is not directly face-on, the bottom half of the ring structure is brighter due to Doppler boosting of the jet and disk emission. Based on the differential limb brightening and velocities in the large-scale jet (Walker et al., 2018), the simulations were oriented so the jet rotates clockwise in the plane of the sky. Given the sense of rotation determined by Walker et al. (2018) and the direction of the projected jet on the sky, the 230 GHz ring is expected to be brighter on the bottom from Doppler boosting. This north-south asymmetry was confirmed in the images from the full EHT in The Event Horizon Telescope Collaboration et al. (2019d).

Bright ridges tracing the rotating helical magnetic field are visible in both simulations. These extended structures are fainter than the bright ring, but they are visible even in the images blurred to the EHT's resolution. The brightest spot in the 230 GHz image moves with the rotation of the magnetic field lines, particularly in the `H10` model, which lacks bright disk emission. When blurred to the EHT's resolution, this evolution will generally follow the clockwise sense of rotation of the disk and jet. In the specific frames selected from the `H10` simulation, however, shifts in the relative brightness of two filaments as they rotate produce an apparent evolution in the blurred frames that is slightly counterclockwise. The 2017 EHT observations showed time-variability in the data and images on similar timescales, but with only four days of observation it is impossible to directly tie the motion in EHT images of the M87 shadow (The Event Horizon Telescope Collaboration et al.,

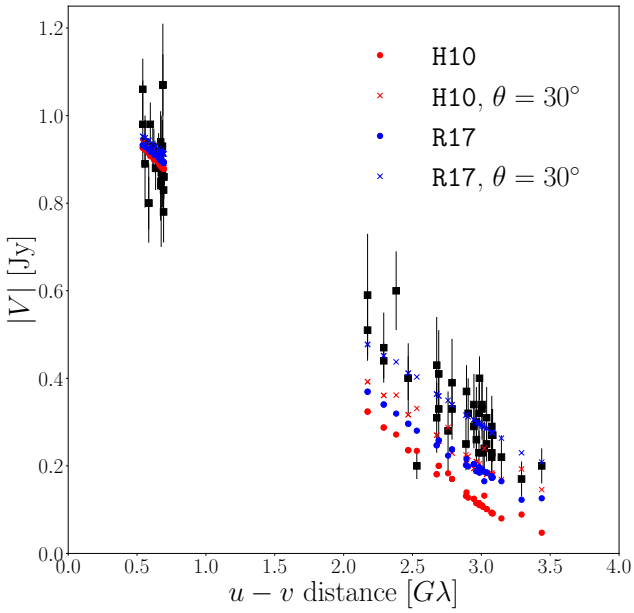


Figure 3.13: Comparison of visibility amplitudes from the EHT in 2009 (Doeleman et al., 2012) and 2012 (Akiyama et al., 2015). The corresponding visibilities obtained from the fiducial 230 GHz snapshot images (Figure 3.12) for model H10 and R17 are displayed by red and blue circles, respectively. At the chosen values of inclination $\theta = 17°$ and distance to the black hole $D = 16.7$ Mpc, the images from the simulations are somewhat too large and, therefore, underpredict the measured visibility amplitudes on the Hawai'i-California and Hawai'i-Arizona baselines. At a larger inclination angle $\theta = 30°$, sampled visibility amplitudes from the simulated images (denoted by red and blue xs) better match the observations.

2019d) with rotating material in the accretion flow.

The precise shadow diameter and shape is sensitive to the inclination and black hole spin (Bardeen et al., 1972; Chandrasekhar, 1983). Even for these simulated images that have a prominent shadow, any one EHT image reconstruction would leave substantial uncertainty in the shadow size due to the limited resolution and contributions to the source structure from the foreground jet. It may be possible, however, to make a more precise measurement of the shadow size with multi-epoch imaging. While the shadow is a persistent feature set only by the mass and spin of the black hole, the foreground jet at 230 GHz rotates quickly, completing a full revolution on a time-scale of weeks to months.

Figure 3.13 compares visibility amplitudes extracted from the fiducial 230 GHz images in Figure 3.12 with observations from the EHT with stations in Hawai'i, California, and Arizona in 2009

(Doeleman et al., 2012) and 2012 (Akiyama et al., 2015). The compact flux density measured in these two years was ≈0.98 Jy. At the chosen values of inclination, $\theta = 17°$ up from the south pole, and distance to the black hole $D = 16.7$ Mpc, the snapshot images from the simulations are somewhat too large and underpredict the measured visibility amplitudes on the Hawai'i-California and Hawai'i-Arizona baselines. Compared to the extremely compact images obtained from the less magnetized 2D simulations in Ryan et al. (2018), the wide-opening-angle jets produce extended emission that increases the overall image size, though the bright $\sim 40\mu$as photon ring remains the most prominent feature.

It is important to note that the image size in the simulations is highly sensitive to the assumed viewing inclination. At larger values of inclination angle than the $\theta = 17°$ taken from Walker et al. (2018), the image size decreases. While the jet inclination angle is constrained to $\lesssim 20°$ at distances $\sim$100pc from the black hole from apparent superluminal velocities measured near the HST-1 knot (Giroletti et al., 2012), the inclination is not as definitively constrained on scales closer to the black hole. In their conservative estimate, Mertens et al. (2016) give an upper limit $\theta \lesssim 27°$. At $\theta = 30°$, the visibilities from `R17` nearly match the observations, and the simulated visibilities from `H10` are much less discrepant than at 17° (Figure 3.13). Furthermore, at this larger value of the inclination angle, the limb-brightening in the 43 and 86 GHz images from both simulations more closely matches the VLBI maps (left panels of Figures 3.9 and 3.10), and the counterjet is more prominent at both frequencies.

The image size in the simulations at a fixed inclination angle is also highly sensitive to the choice of $\sigma_{\rm cut} = 25$ determined to avoid including emission from the regions with densities set to the floor value. In nature, these highly magnetized regions will produce emission which may be significant contributors to the 230 GHz flux density. Including radiation from $\sigma_{\rm i} > 25$ regions makes the image more compact (see Section 3.2.5). Thus, it remains possible that different prescriptions for

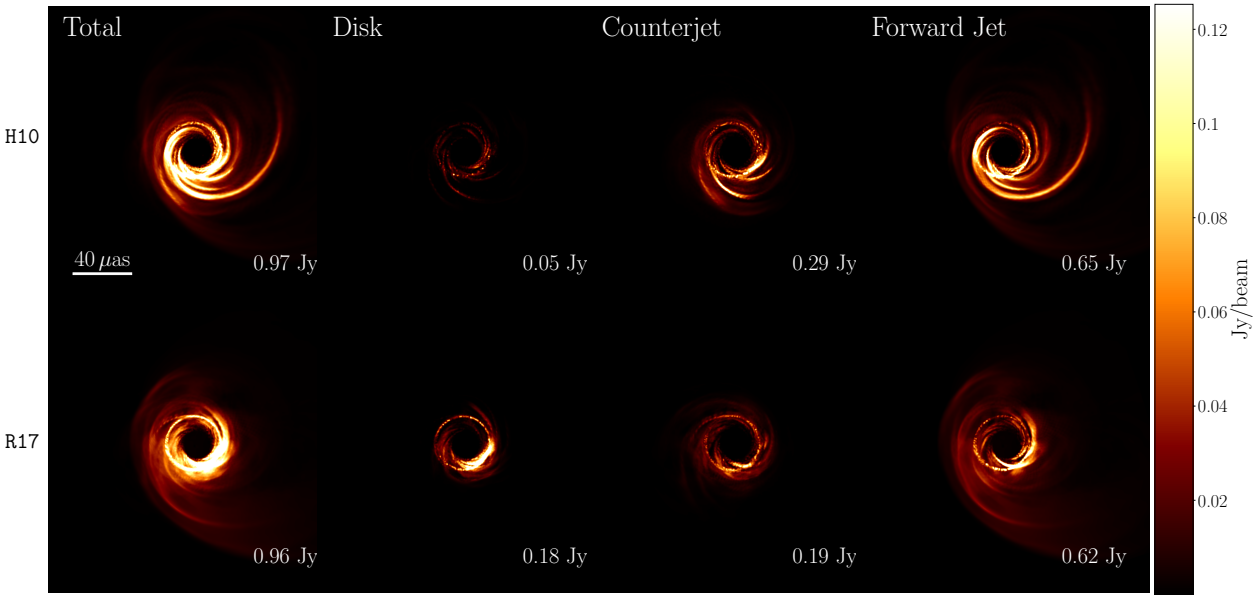


Figure 3.14: Snapshot 230 GHz images of the two simulations generated using `grtrans` to zero out emissivities from selected regions, highlighting the emission from different components of the accretion flow. The leftmost column shows the image produced including all regions in the radiative transfer with $\sigma_{\rm i} < \sigma_{\rm cut}$; these are the same images as in the middle row of Figure 3.12. The second column from the left shows the image generated from the disk, setting emissivities in the jet regions (defined as Be > 0.05) to zero. The third column shows emission only from the counterjet (Be > 0.05 and polar angle $\theta < \pi/2$). The fourth column shows emission from only the forward jet (Be > 0.05 and polar angle $\theta > \pi/2$). Since the accretion flow is optically thin at 230 GHz, the total flux densities of the component images nearly add up to the total flux density in the image generated from the entire emissivity distribution. Both simulations have images that are dominated by emission originating in the accretion flow and forward jet. `H10` has more counterjet emission, and `R17` has more disk emission.

including radiation in post-processing from highly magnetized regions may produce images from MAD simulations that are consistent with the 2009 and 2012 EHT size measurements.

Figure 3.14 considers the same 230 GHz snapshots as in the central column of Figure 3.12 and decomposes the emission into three component parts: disk, counterjet, and forward jet. As above, regions with Be > 0.05 are defined to be in the jets, and Be ≤ 0.05 regions are in the disk. These component images were produced by zeroing out the emissivities outside the selected regions when doing the radiative transfer in `grtrans`. All the images in Figure 3.14 maintain the global cut on emissivities from regions with $\sigma_{\rm i} > 25$.

Because the entire M87 accretion flow is optically thin at 230 GHz, the total flux densities of the component images in Figure 3.14 nearly add up to the total flux density in the image

generated from the entire emissivity distribution. In both simulations, the majority (≈60%) of the emission comes from the forward jet. The forward jet emission in both cases consists of a spiral structure surrounding the black hole where emission traces magnetic field lines, as well as a persistent component from the photon ring. Note that a substantial fraction of the jet emission comes from between the Be = 0.05 and $\sigma_{\rm i} = 1$ surfaces, so adopting $\sigma_{\rm i} > 1$ as the definition of the jet region results in assigning more of what is currently classified as forward jet and counterjet emission to the disk.

In the `H10` model, nearly all of the emission not from the forward jet is produced from the counterjet, which adds to the prominent photon ring. Because the electrons in this model are so much hotter at the base of the jet than in the disk, the emission from these regions dominates the total, and the disk emission is negligible. In contrast, the reconnection heating model `R17` has approximately equal contributions from the counterjet and disk. The disk emission shows a persistent bright spot from the Doppler-boosted accretion flow. Unlike the bright spots produced from rotating magnetic field lines, this spot remains constant in position and does not rotate around the photon ring with time. This feature points to the possibility of using multi-epoch imaging with the EHT to disentangle the source structure and identify whether or not the accretion disk emission makes a substantial contribution to the total image flux density.

3.2.5 The effects of $\sigma_{\rm cut}$

This section explores the effects of different choices for $\sigma_{\rm cut}$ in the results presented in the previous sections. As discussed in Section 3.1.3, a choice of $\sigma_{\rm cut}$ is necessary in radiative transfer in order to exclude emission from the regions that most evidently suffer from errors in the gas-dynamical evolution. Although the ratio of electron-to-ion temperatures behaves as expected from the two heating prescriptions in this region (Figure 3.2), the overall temperature scale from the gas evolution

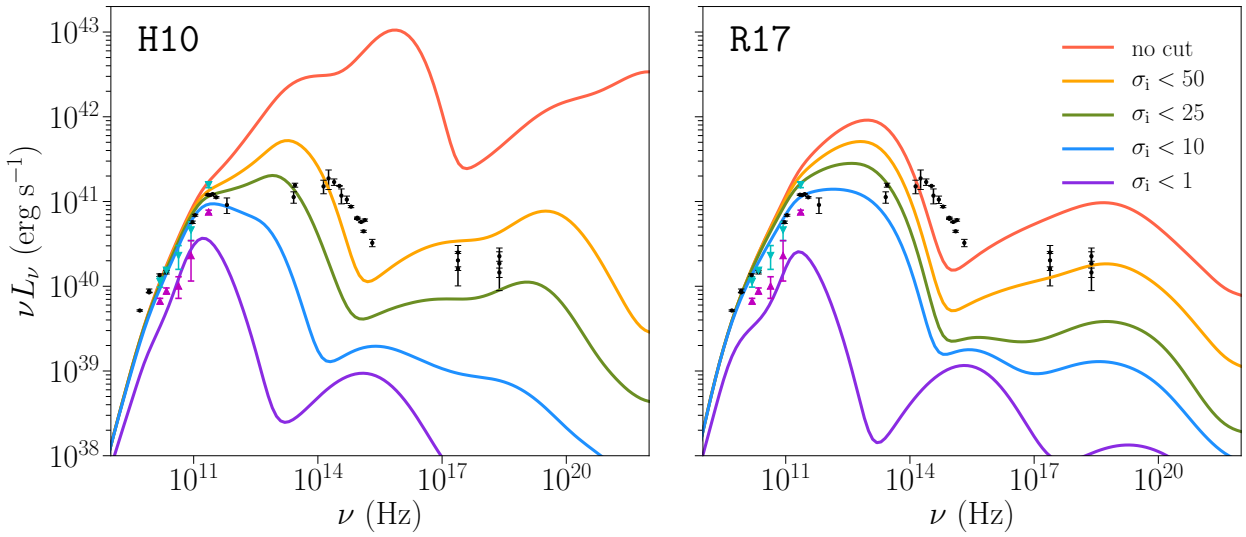


Figure 3.15: Snapshot spectra from the two simulations generated with different values of $\sigma_{\rm cut}$ in the radiative transfer. Spectra were generated with `HEROIC` by zeroing out emissivities from all regions with fluid frame magnetization $\sigma_{\rm i} \geq \sigma_{\rm cut}$, with $\sigma_{\rm cut}$ =1, 10, 25 (the fiducial value), 50, and with no ceiling. Because the images have different total flux densities, the intensity scale at each $\sigma_{\rm cut}$ is different. In both simulations, any choice of $\sigma_{\rm cut}$ above unity has little effect on the radio spectrum up to 230 GHz. Most emission in this part of the spectrum comes from less-magnetized regions farther from the black hole. The choice of $\sigma_{\rm cut}$ has a drastic effect on the spectrum at higher frequencies as direct synchrotron emission and Compton scattering in the most magnetized, high-temperature regions close to the black hole is added, increasing the radiative power. When no $\sigma_{\rm i}$ ceiling is imposed, model `H17` has an extreme total luminosity $> 10^{43}$ erg s^{-1}.

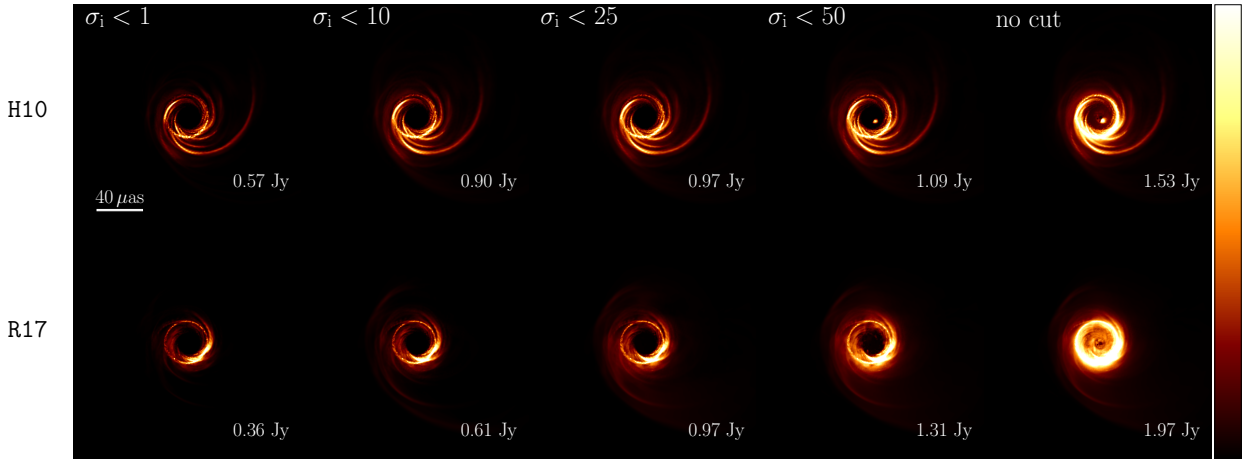


Figure 3.16: Snapshot images from the two simulations generated with different values of $\sigma_{\rm cut}$ in the radiative transfer. From left to right, images were generated using $\sigma_{\rm cut}$ =1, 10, 25 (the fiducial value), 50, and with no ceiling. In both simulations, the overall image structure is similar at all cuts up to $\sigma_{\rm cut} = 50$. Because $\sigma_{\rm i}$ increases rapidly with decreasing polar angle in the jet region (Figure 3.1), including regions of higher and higher magnetization does not open up very different regions of the accretion flow to the radiative transfer. In contrast, including the entire interior of the jet (the rightmost images) produces substantial new emission at and in front of the photon ring.

may suffer from errors in any region with $\sigma_{\rm i} > 1$. On the other hand, a value of $\sigma_{\rm i} > 1$ may be necessary to investigate jet emission from these simulations and compare to data.

Figure 3.15 shows spectra generated with five different values $\sigma_{\rm cut}$ =1, 10, 25 (the fiducial value), 50, and no ceiling. In both simulations, any choice of $\sigma_{\rm cut} > 1$ has relatively little effect on the radio spectrum up to 230 GHz. Most emission in this part of the spectrum comes from less-magnetized material with $\sigma_{\rm i} < 25$ farther up from the black hole. As a result, the predictions of both models at frequencies < 230 GHz should be relatively insensitive to the choice of $\sigma_{\rm cut}$.

For $\sigma_{\rm cut} = 1$, both simulations are under-luminous across the radio spectrum. This change in the luminosity with decreasing $\sigma_{\rm cut}$ from 10 to 1 indicates that the majority of the radio flux originates in regions in the jet with $1 < \sigma_{\rm i} < 10$. However, this inference is dependent on the overall normalization of the simulations chosen to match the observed 230 GHz flux density with $\sigma_{\rm cut} = 25$.

For $\nu \gtrsim 230$ GHz, direct synchrotron emission and Compton scattering in the most magnetized, high-temperature regions close to the black hole makes a substantial contribution to the radiative power. The predictions of both models at these frequencies are thus strongly dependent on the choice of the $\sigma_{\rm cut}$, and should be viewed with caution. Although the imposition of a global ceiling on $\sigma_{\rm i} < \sigma_{\rm i,max}$ for stability when running the simulation makes it impossible to conclusively predict the emission from the models at these higher frequencies, some trends in the spectra are with increasing $\sigma_{\rm cut}$ in Figure 3.15 are apparent that should be explored in future work. In model `R17`, the overall luminosity, peak synchrotron frequency, and Compton power all increase when adding in increasingly magnetized regions of the simulation, but the overall spectral shape does not drastically change. It seems possible that future simulations with a higher, more reliable absolute bound on $\sigma_{\rm i,max}$ could extend the ceiling on $\sigma_{\rm cut}$ imposed in the radiative transfer and produce a model that better matches the higher frequency measurements.

In contrast, when no $\sigma_{\rm cut}$ is imposed, model H17 has an extremely large total luminosity $> 10^{43}$ erg s^{-1}, dramatically exceeding measurements at all frequencies above 230 GHz. As noted in Section 3.2.1, this extreme luminosity affects the dynamics of the fluid, most notably by reducing the mechanical jet power in this model relative to R17. This extreme luminosity ultimately results from the high electron temperatures produced by $\delta_{\rm e} \to 1$ in the most highly magnetized regions near the black hole. However, because the density in these regions is set by the $\sigma_{\rm i,max} = 100$ ceiling, it remains possible that a simulation using the Howes (2010) turbulent cascade heating prescription with a higher value of $\sigma_{\rm i,max}$ and a correspondingly more evacuated jet would produce more reasonable total radiative power.

Figure 3.16 presents 230 GHz images from the two simulations using different choices of $\sigma_{\rm cut}$. In both simulations, the overall image structure is similar at all cuts up to $\sigma_{\rm cut} = 50$. Since $\sigma_{\rm i}$ increases rapidly with polar angle in the jet region (Figure 3.1), including regions of higher and higher magnetization does not open up very different regions of the accretion flow to radiative transfer as long as $\sigma_{\rm cut}$ remains below the overall simulation ceiling.

The rightmost images in Figure 3.16 were produced with no $\sigma_{\rm i}$ imposed in the radiative transfer, opening up the entire interior of the jet. Taking $\sigma_{\rm cut} \to \infty$ produces substantial new emission from both simulations that originates from close to the black hole, dramatically increasing the brightness and compactness of the images. In the H10 model, the forward jet compact emission is concentrated in a bright spot in the middle of the ring from the very-high-temperature electrons at the jet base, while in the R17 model, the highest-$\sigma_{\rm i}$ emission forms a more diffuse haze in the middle of the jet in front of the photon ring. Furthermore, in both models, the addition of strong emission close to the black hole dramatically increases the prominence of the lensed counterjet emission in the photon ring, to the point where the forward jet and counterjet contributions to the 230 GHz image become approximately equal. The strength of the emission at the base of the jet/counterjet sets the

relative brightness of the photon ring to the surrounding disk and jet (Dexter et al., 2012), but this emission is unfortunately unconstrained in the simulations because of the uncertainties associated with the choice of $\sigma_{\rm cut}$.

3.3 Discussion

The two MAD simulations presented in this paper produce spectra and images that are broadly consistent with observations of the M87 jet at centimeter and millimeter wavelengths. Model `R17` produces a jet power that is in line with observations, while model `H10` produces a jet power lower by a factor of two. Both models reproduce the wide jet opening angle observed at 43 GHz, and they are both consistent with core-shift observations. However, at a viewing angle of 17°, the 230 GHz images from both models are too large to match EHT observations from 2009 and 2012.

M87 has been simulated before using GRMHD and GRRMHD codes, though not as frequently as Sgr A*. The model of Mościbrodzka et al. (2016a) is a representative state-of-the-art combination of single-fluid GRMHD and radiative transfer. The authors used disks with relatively weak magnetization from Shiokawa (2013) and added thermal electrons in post-processing that were assumed to be hot in the jet and cool in the disk (Mościbrodzka et al., 2014). They produce a model with a jet power in the correct range, a flat radio spectrum, and an limb-brightened jet image at 43 GHz. The jets in their simulations show substantial variability and apparent superluminal motion from field lines along the funnel wall. However, the jets in their simulation have apparent opening angles at 43 GHz that are smaller than the observed ∼55°.

As in the MAD models in this chapter, 230 GHz images from Mościbrodzka et al. (2016a) show spiral structures from helical field lines in the jet; this is a common prediction of both weakly magnetized and MAD models. At 230 GHz, their images are dominated by the counterjet (see also

Dexter et al., 2012). Unlike in the MAD models presented here, their 230 GHz images satisfy the constraints on the image size from the EHT at 20° inclination.

The authors have also investigated the polarized emission from their models (Mościbrodzka et al., 2017). They have shown that in their counterjet-dominated models it is possible to produce images with rotation measure and polarization fraction in line with observations, through the depolarization of the counterjet emission as it passes through the cooler disc. The emission in both MAD models at 230 GHz is dominated by the forward jet, although counterjet emission is still significant. It is possible that, with less opportunity for depolarization through the disk, the forward-jet-dominated images might produce a net polarized flux that exceeds the observed value (∼1%; Kuo et al. 2014). Expanding the analysis of these simulations to polarized images is an important direction for future work.

Recently, Ryan et al. (2018) preformed the first two-temperature simulation of M87 using the code `ebhlight` in axisymmetry. Unlike `KORAL`, which considers only the frequency-integrated radiation field, `ebhlight` uses a Monte Carlo method where photons with distinct frequencies are emitted and absorbed on the simulation grid. Consequently, they obtain spectra as a natural product of their simulations without having to perform radiative transfer in post-processing.

Ryan et al. (2018) considered disks that were far less magnetized than those explored here. Consequently, to match the observed 230 GHz flux density they required higher accretion rates than in these simulations. In their best-fitting model, the accretion rate is $\dot{M}/\dot{M}_{\rm Edd} \sim 10^{-6}$, about three times the value from the lower-accretion-rate model `H10`. In all their models, they found that Coulomb heating of electrons becomes important in the outer disk. As in this chapter, they see that radiation plays a significant role in the inner disk. Notably, they explore simulations with both high (Gebhardt et al., 2011) and low (Walsh et al., 2013) black hole mass and consider two values of the black hole spin ($a = 0.5$ and $a = 0.9375$). They find their high-spin, high mass

model produces both a spectrum and 230 GHz image consistent with the available data – this chapter has adopted these preferred parameter values in this study. Unlike in the present chapter, at $M = 6.2 \times 10^9 M_\odot$ and $a = 0.9375$, they obtain a compact, counterjet-dominated 230 GHz image that is consistent with past EHT measurements of the overall image size. However, the jet powers produced in their simulations are several orders of magnitude too low, and their weakly magnetized disks also produce jet opening angles that are too small when compared against VLBI observations. These problems are not present in this chapter's MAD models, which match the observed spectral and image characteristics of the M87 jet well at all frequencies between 15 and 86 GHz.

Simulating the jet interior remains a problem in all simulations, not just in the MAD regime, and all simulations must impose some sort of density floor in the magnetized, evacuated jet to ensure numerical stability. As discussed in Section 3.2.5, this problem is particularly important in MAD models where much of the emission may come from these highly magnetized regions. The matter content of the jet is still unknown; it may be filled with plasma loaded from the disk or populated by a pair plasma of electrons and positrons (Mościbrodzka et al., 2011; Broderick & Tchekhovskoy, 2015). Furthermore, it is likely that nonthermal electrons in the disk (Broderick & Loeb, 2009) or jet (Dexter et al., 2012) contribute to the emission at 230 GHz and into the infrared, optical, and ultraviolet. Further work with these simulations using additional postprocessing prescriptions for the jet matter content and electron distribution may be able to provide constraints on models for the jet interior, while still relying on predictions from self-consistent temperature evolution in the jet wall and disk regions of the simulation.

3.4 Summary and conclusions

This chapter presents the first Magnetically Arrested disk simulations performed with two-temperature, radiative, general relativistic magnetohydrodynamics, previously published in Chael et al. (2019b). After normalizing to the observed (2009 and 2012) flux density at 230 GHz, the two simulations `H10` and `R17` produce spectra consistent with that of M87 in the radio, millimeter, and submillimeter. They both produce powerful jets with jet powers consistent with or close to the measured range. The jets both have a wide opening angle consistent with the wide opening angle of the M87 jet observed with VLBI at 43 and 86 GHz. The simulated images at centimeter and millimeter wavelengths exhibit a core-shift which reproduces that reported in Hada et al. (2011). The 230 GHz images from both simulations clearly show the lensed photon ring, i.e., the black hole shadow, indicating that even in forward-jet-dominated MAD images, the full EHT should be able to image this feature. In addition, the images are dynamic on time-scales of months to years. If these images are reflective of M87, then repeated EHT observations should be able to detect the motion of rotating magnetic fields that are driven by the spin of the black hole.

Two sub-grid electron heating mechanisms were considered in this chapter, and they produce jet outflows with somewhat distinct properties. The magnetic reconnection heating model of Rowan et al. (2017) used in simulation `R17` launches a jet powered by the black hole spin with a mechanical jet power $\sim 10^{43}$ erg s^{-1}, in the correct range for M87 (Reynolds et al., 1996; Stawarz et al., 2006). In contrast, the jet heated by the turbulent cascade model of Howes (2010) in simulation `R17` produces intense radiation at the jet base; this radiation saps the jet of mechanical energy, resulting in a mechanical jet power a factor of two less than in `R17`.

Despite the differences in their kinematics, the spectra and images of the two models are quite similar at centimeter, millimeter, and submillimeter wavelengths. The simulations did not run

long enough to produce the full extent of the jet observed with VLBI, but the images from the inner milliarcsecond show similar structure to VLBA and GMVA images at the corresponding wavelengths. Notably, both models reproduce the measured ≈55° opening angle (Walker et al., 2018), and both show emission from the counterjet, though the models produce less counterjet emission than observed in some 43 GHz VLBI images.

At 230 GHz, the simulations produce images that are larger than the size measured by the EHT in 2009 and 2012 (Doeleman et al., 2012; Akiyama et al., 2015). However, the image size is strongly dependent on the viewing inclination, and it becomes consistent with EHT observations around $\theta = 30°$, the upper end of the most conservative range established by VLBI observations (Mertens et al., 2016). The image size also shrinks when the extremely magnetized on-axis regions are included in the radiative transfer. In this chapter, emission from regions with a magnetization greater than $\sigma_{\rm cut} = 25$ is excluded from the radiative transfer. A different treatment of these regions from a simulation with a higher overall ceiling on the magnetization could potentially bring the size of the 230 GHz images in line with observations.

While these simulations under-produce the measured flux in the optical, ultraviolet, and X-ray, it seems clear that the spectrum at these high frequencies is dominated by the hottest, most magnetized regions closest to the black hole. A different treatment of the magnetization ceilings imposed in the simulation and radiative transfer may bring the simulation spectra from these regions more in line with observations. At the other end of the spectrum, investigating the jet on large scales and at frequencies < 15 GHz requires a longer simulation time than considered in this chapter.

The 230 GHz images from these simulations all show distinct black hole shadows, consistent with the EHT's landmark first image of M87's horizon-scale structure in 2019 (The Event Horizon Telescope Collaboration et al., 2019a)which are primarily illuminated by emission originating in the

forward jet. Future investigations of polarized images from the simulations will investigate whether these forward-jet-dominated models satisfy constraints on the 230 GHz polarization fraction; these constraints are naturally satisfied by Faraday depolarization of counterjet emission (Mościbrodzka et al., 2017). Furthermore, the rotation of the accretion disk and jet makes the 230 GHz images dynamic on time-scales of weeks to months. It may be possible to distinguish between models with polarimetric EHT images and by tracking stationary and moving features through repeated observations.

Page intentionally left blank

Text in this chapter was previously published in *MNRAS* 470 (2017), 2, pp 2367-2386 (A. Chael, R. Narayan, and A. Sądowski).

4

Nonthermal particle evolution in accretion simulations

Chapter 1 presented a method for evolving separate thermal populations of electrons and ions in global accretion flow simulations; Chapters 2 and 3 applied this method to simulations of Sgr A* and M87 and compared the results with observational data from both systems. However, at the lowest densities in the accretion flows around Sgr A* and M87, collisions will not completely relax the electron distribution function to a local Maxwellian (Mahadevan & Quataert, 1997). Even if the bulk of the electron distribution function is thermal, processes like shocks and magnetic reconnection (e.g., Sironi & Spitkovsky, 2011, 2014) can accelerate a small fraction of the electrons into a relativistic nonthermal distribution, which will persist for a long time because of the lack of collisions. In the specific case of Sgr A*, while traditional ADAF models emitting via thermal synchrotron radiation can describe the bulk of the emission from the 'submm bump' around 10^{12} Hz (Narayan & Yi, 1995b; Narayan et al., 1998; Quataert & Narayan, 1999), the quiescent infrared and unresolved X-ray emission in the spectrum are most easily explained with hybrid models that include a small population of high-energy, power-law electrons (Özel et al., 2000; Yuan et al., 2003, 2004; Broderick & Loeb, 2006). In addition, Sgr A* flares continually in the millimeter,

near-infrared (NIR), and X-ray (Genzel et al., 2003; Yusef-Zadeh et al., 2006; Neilsen et al., 2013; Zhang et al., 2017). The infrared and X-ray flares are correlated, and the spectrum of strong flares shows a stable spectral index and a cooling break indicative of nonthermal synchrotron radiation (Gillessen et al., 2006; Ponti et al., 2017). In 2018, spatially resolved observations of NIR flares by the GRAVITY interferometer revealed circular motion (GRAVITY Collaboration et al., 2018b), indicating that the compact flares may originate in "hot spots" of plasma orbiting near the black hole's innermost stable circular orbit (Broderick & Loeb, 2006).

Most GRMHD simulations are post-processed assuming emission only originates from thermal electrons, but some recent investigations have added a population of electrons with power-law distributions during post-processing to investigate the effect on the quiescent spectrum and sub-mm image size of Sgr A* (Mao et al., 2016; Davelaar et al., 2018). Adding nonthermal electrons in post-processing to single-fluid GRMHD simulations, Ball et al. (2016) successfully reproduced strong flare amplitudes, suggesting that the rapid and localized injection of broad nonthermal, power-law electron distributions can efficiently generate flares in these systems.

This chapter outlines the next, logical step from the thermal, two-temperature simulations considered in Chapters 1– 3. Namely, in addition to evolving thermal ions, thermal electrons and radiation in a GRRMHD simulation, this method evolves a locally isotropic population of nonthermal electrons. The exchange of energy and momentum among the various particle populations and the radiation field is accounted for at each time step during the global evolution. The nonthermal electrons are heated by a prescribed fraction of the total viscous heating rate; they further gain and lose energy by adiabatic compression and expansion, Coulomb coupling, inverse Compton scattering, and radiative cooling. Like the thermal, two temperature method used in Chapters 1– 3, this new algorithm (originally presented in Chael et al. 2019b) is implemented in the GRRMHD code KORAL.

Section 4.1 presents the additional equations and physics added to the simulation procedure described in Chapter 1 to evolve a nonthermal population of electrons in the presence of radiation, and section 4.2 describes the numerical algorithm used in KORAL. Section 4.3 discusses a number of simple test problems used to validate the code, and Section 4.4 presents initial results of a 2D toy simulation of a Sgr A*-like accretion flow including nonthermal electrons. Section 4.5 discusses the next steps necessary to use the algorithm developed in this Chapter in realistic, 3D simulations of Sgr A* to investigate the physical origin of its variability and high-energy flares.

4.1 Physics

4.1.1 Fluid populations

As introduced in Chapter 1, GRMHD simulations typically track a single magnetized perfect fluid as a function of position and time described by the gas density ρ, internal energy density u, four-velocity u^μ, and magnetic field four-vector b^μ.

In this chapter, the fluid is expanded to *three* populations: thermal ions, thermal electrons, and an isotropic distribution of nonthermal electrons. All three populations are assumed to move with the same velocity u^μ. This assumption automatically preserves local charge neutrality and simplifies the evolution equations for the nonthermal spectrum (Section 4.1.3). Under this approximation, Equation 1.3 remains a valid description of the total stress-energy, although the equation of state relating p and u changes as explained below.

The electrons contribute negligibly to the mass density, $\rho = m_{\rm p} n_{\rm i}$.. Charge neutrality enforces the constraint:

$$n_{\rm e\,th} + n_{\rm e\,nth} = n_{\rm i} = \rho/m_{\rm p}, \tag{4.1}$$

where $n_{\mathrm{e\,th}}$, and $n_{\mathrm{e\,nth}}$ are the number densities of the thermal and nonthermal electrons, respectively.[1]

All three fluid populations can have substantial contributions to the net energy density and pressure of the fluid:

$$u = u_{\mathrm{i}} + u_{\mathrm{e\,th}} + u_{\mathrm{e\,nth}},$$
$$p = p_{\mathrm{i}} + p_{\mathrm{e\,th}} + p_{\mathrm{e\,nth}}. \tag{4.2}$$

The energy densities and pressures of the thermal species are determined by their respective temperatures $T_{\mathrm{i,e}}$ and corresponding adiabatic indices $\Gamma_{\mathrm{i,e}}\,(\Theta_{\mathrm{i,e}})$, which are functions of temperature through the dimensionless temperature $\Theta_{\mathrm{i,e}} = k_{\mathrm{B}} T_{\mathrm{i,e}} / m_{\mathrm{i,e}} c^2$ (Equation 1.23).

The nonthermal electrons are assumed to be isotropic in the fluid rest frame, with a distribution $n(\gamma)$ in Lorentz factor γ, sampled over a large range from a minimum γ_{min} to a maximum γ_{max}. The number density, energy density, and pressure of the nonthermal electrons are then given by integrals over the distribution $n(\gamma)$,

$$n_{\mathrm{e\,nth}} = \int_{\gamma_{\mathrm{min}}}^{\gamma_{\mathrm{max}}} n(\gamma)\,\mathrm{d}\gamma, \tag{4.3}$$
$$u_{\mathrm{e\,nth}} = m_{\mathrm{e}} c^2 \int_{\gamma_{\mathrm{min}}}^{\gamma_{\mathrm{max}}} n(\gamma)(\gamma - 1)\,\mathrm{d}\gamma, \tag{4.4}$$
$$p_{\mathrm{e\,nth}} = \frac{m_{\mathrm{e}} c^2}{3} \int_{\gamma_{\mathrm{min}}}^{\gamma_{\mathrm{max}}} n(\gamma)(\gamma - \gamma^{-1})\,\mathrm{d}\gamma, \tag{4.5}$$

where m_{e} is the electron mass.[2]

[1] As in Chapter 1, the procedure laid out in this Section assumes the plasma is pure ionized Hydrogen.

[2] The $m_{\mathrm{e}} c^2 (\gamma - \gamma^{-1})/3$ factor in the integrand for the pressure $p_{\mathrm{e\,nth}}$ is just $p^2/3E$, where $p^2 = m_{\mathrm{e}} c^2 (\gamma^2 - 1)$ is the square of the particle momentum, $E = \gamma m_{\mathrm{e}} c^2$ is the particle energy, and the $1/E$ factor comes from the relativistically invariant measure $d^3\boldsymbol{p}/E$.

The net adiabatic index of the combined three-species fluid with a nonthermal contribution is

$$\Gamma_{\rm gas} = 1 + \frac{p}{u} = 1 + \frac{p_{\rm i} + p_{\rm e\,th} + p_{\rm e\,nth}}{u_{\rm i} + u_{\rm e\,th} + u_{\rm e\,nth}}, \tag{4.6}$$

using Equation 1.23 for the thermal quantities and Equations 4.4 and 4.5 for the nonthermal energy density and pressure.

In addition to the three particle populations, KORAL evolves an additional fluid to represent radiation using the M1 closure scheme, described by the energy density $\bar{E}$ in the radiation frame, the photon number $\bar{n}_{\rm R}$, and the radiation frame four velocity $u^\mu_{\rm R} \neq u^\mu$ (Equations 1.7 – 1.8).[3]

4.1.2 Updated GRRMHD equations

The conservation equations that govern the evolution of the total fluid stress-energy $T^\mu{}_\nu$, electromagnetic field tensor $F^{*\mu\nu}$, and radiation field stress-energy $R^\mu{}_\nu$ are unchanged by the addition of nonthermal electrons (Equations 1.9–1.12). However, G_ν, the four-force density that couples the evolution of the radiation and gas, is modified by radiation from nonthermal electrons. The nonthermal electrons, like those in the thermal population, are assumed to radiate isotropically in the fluid rest frame. Thus, Equations 1.14– 1.15 now become

$$\hat{G}^0 = \tilde{\rho}\,(\kappa_{\rm P,a}\hat{E} - 4\pi\kappa_{\rm P,e}\hat{B}) + \hat{G}^0_{\rm IC\,th} + \hat{G}^0_{\rm nth}, \tag{4.7}$$

$$\hat{G}^i = (\tilde{\rho}\kappa_{\rm R} + \rho\kappa_{\rm es})\hat{F}^i. \tag{4.8}$$

As in Chapter 1, the κ factors are the grey, frequency-averaged opacities for the thermal radiative processes, $\hat{G}^0_{\rm IC\,th}$ is the thermal energy loss from inverse Compton scattering (Sądowski &

[3] As in Chapter 1, quantities in the radiation rest frame are denoted with bars, and quantities in the fluid frame are denoted with hats.

Narayan, 2015), $\hat{F}^i$ is the rest-frame radiation momentum flux ($\hat{F}^i = \hat{R}^{0\,i}$), and $\hat{B} = \sigma T_{\rm e}^4/\pi$ is the electron blackbody radiance. The new factors in Equations 4.7–4.8 are $\hat{G}^0_{\rm nth}$, the energy loss to radiation from the nonthermal population, and $\tilde{\rho}$, the reduced fluid density when accounting for the nonthermal population:

$$\tilde{\rho} = \rho \, \frac{n_{\rm e\,th}}{n_{\rm e\,th} + n_{\rm e\,nth}}. \tag{4.9}$$

The factors $\kappa_{\rm P,e}\hat{E}$, $4\pi\kappa_{\rm P,a}\hat{B}$, and $\kappa_{\rm R}\hat{F}^i$ account for emission and absorption from only the thermal electrons, so they are multiplied by $\tilde{\rho}$ instead of ρ. The full density ρ is used in the term for the electron scattering opacity (Equation 1.16), since nonthermal electrons also scatter the emission.

Neglecting absorption from the nonthermal electrons in the rest frame, the nonthermal population only contributes an emission factor $\hat{G}^0_{\rm nth}$. The contribution to the radiative power from the nonthermal electrons is the integral of the radiative cooling rate over the full distribution $n(\gamma)$,

$$\hat{G}^0_{\rm nth} = m_{\rm e}c^2 \int_{\gamma_{\rm min}}^{\gamma_{\rm max}} n(\gamma)\dot{\gamma}_{\rm rad}\,{\rm d}\gamma. \tag{4.10}$$

The quantity $\dot{\gamma}_{\rm rad}$ represents the cooling rate of a single electron with energy $\gamma m_{\rm e}c^2$ from radiative processes in the fluid rest frame; it is always negative. The total radiative cooling rate has contributions from synchrotron, bremsstrahlung, and inverse Compton scattering:

$$\dot{\gamma}_{\rm rad} = \dot{\gamma}_{\rm syn} + \dot{\gamma}_{\rm brem} + \dot{\gamma}_{\rm IC}, \tag{4.11}$$

where $\dot{\gamma}_{\rm syn}, \dot{\gamma}_{\rm brem}, \dot{\gamma}_{\rm IC}$ are given by Equations (4.19, 4.20, 4.21), respectively. Absorption by the nonthermal population is not included in G_ν.

The addition of nonthermal electrons also modifies the source term in the photon number evolu-

tion Equation 1.13. This frame invariant term, $\hat{\dot{n}}_{\rm R}$, becomes (c.f. Equation 1.17):

$$\hat{\dot{n}}_{\rm R} = \hat{\dot{n}}_{\rm syn} + \hat{\dot{n}}_{\rm brem\,th} + \hat{\dot{n}}_{\rm brem\,nth} - \tilde{\rho}\kappa_{n,{\rm a}}\hat{n}_{\rm R}. \tag{4.12}$$

The first term in Equation 4.12 from synchrotron emission of both thermal and nonthermal electrons is unchanged from the original expression (Equation 1.18) because the number of photons emitted in synchrotron is independent of the energy of the emitting particle. The second term is the production of photons from thermal bremsstrahlung emission, and the last term is the photon loss rate from absorption by the thermal electrons (see Sądowski & Narayan 2015; Sądowski et al. 2017). These terms are only modified by a factor $\tilde{\rho}/\rho$. The new, third term gives the corresponding rate of photon emission by bremsstrahlung from the nonthermal distribution. For an electron at γ, the bremsstrahlung photon production can be approximated by assuming that photons are only produced with energy $h\nu = \gamma m_{\rm e} c^2$, or

$$\hat{\dot{n}}_{\rm brem\,nth} = \int_{\gamma_{\rm min}}^{\gamma_{\rm max}} \frac{\dot{\gamma}_{\rm brem}}{\gamma}\, n(\gamma)\, {\rm d}\gamma. \tag{4.13}$$

4.1.3 The nonthermal distribution evolution equation

The evolution equation for the nonthermal distribution can be derived by taking angular moments of the relativistic Boltzmann equation and imposing the requirement that the distribution be isotropic in the fluid rest frame (Lindquist, 1966; Webb, 1985, 1989). The isotropy assumption truncates the hierarchy of moment equations and leaves a single equation:

$$[n(\gamma)u^{\alpha}]_{;\alpha} = -\frac{\partial}{\partial\gamma}\left[\dot{\gamma}_{\rm tot} n(\gamma)\right] + Q_{\rm I}(\gamma), \tag{4.14}$$

$$\dot{\gamma}_{\rm tot} = \dot{\gamma}_{\rm adiab} + \dot{\gamma}_{\rm C} + \dot{\gamma}_{\rm rad}. \tag{4.15}$$

Aside from the injection (source) term $Q_{\rm I}(\gamma)$, Equation 4.14 is essentially a conservation equation in five dimensions: four dimensions correspond to space and time (left-hand side of Equation 4.14), and the fifth dimension corresponds to the fluid frame particle Lorentz factor γ, through which particles move with velocity $\dot{\gamma}_{\rm tot}$. This velocity is broken into three parts: $\dot{\gamma}_{\rm adiab}$ from adiabatic heating and cooling due to gas compression and expansion, $\dot{\gamma}_{\rm C}$ from cooling due to the (weak) Coulomb coupling with the thermal electrons, and $\dot{\gamma}_{\rm rad}$ from the energy lost to radiation (Equation 4.11). Since nonthermal electrons are assumed in this chapter to only emit and never absorb photons, $\dot{\gamma}_{\rm rad}$ is always negative; furthermore, the Coulomb coupling term $\dot{\gamma}_{\rm C}$ is also negative, since the nonthermal population by assumption consists of particles more energetic than the thermal electrons that they couple to.

The adiabatic 'cooling' rate $\dot{\gamma}_{\rm adiab}$ can be positive or negative, depending on whether the gas is compressing or expanding. This term can be derived from the relativistic Boltzmann equation without interaction terms (Webb, 1989):

$$\dot{\gamma}_{\rm adiab} = -\frac{1}{3} u^{\alpha}{}_{;\alpha} (\gamma - \gamma^{-1}). \tag{4.16}$$

It is negative when the gas expands, ($u^{\alpha}{}_{;\alpha} > 0$), and it is positive when the gas is compressed ($u^{\alpha}{}_{;\alpha} < 0$).

The term $Q_{\rm I}(\gamma)$ in Equation 4.14 is the rate of injection of high energy electrons from the thermal to the nonthermal distribution at a given γ. In principle, $Q_{\rm I}(\gamma)$ is a function of local conditions and depends on microscopic plasma processes that accelerate electrons into the nonthermal distribution. For simplicity, this chapter assumes that the electrons are injected with a power-law distribution with index p,

$$Q_{\rm I}(\gamma) = C\gamma^{-p}, \tag{4.17}$$

where C is a normalization factor. In addition, this chapter assumes the total injection rate in nonthermal injection function $Q_{\rm I}$ is equal to a fraction $\delta_{\rm nth}$ of the total dissipation rate $q^{\rm v}$ at a given location and time. That is, given the total viscous heating rate $q^{\rm v}$, computed numerically from the simulation (see Equation 4.32), a fraction $\delta_{\rm e}$ of the energy goes to the electrons, of which a fraction $\delta_{\rm nth}$ goes into the nonthermal population. This assumption determines the normalization C in Equation 4.17 by relating

$$m_{\rm e}c^2 \int_{\gamma_{\rm inj\,min}}^{\gamma_{\rm inj\,max}} (\gamma-1) Q_{\rm I}(\gamma)\,{\rm d}\gamma = \delta_{\rm nth}\,\delta_{\rm e}\,q^{\rm v}$$

$$C = \frac{\delta_{\rm nth}\,\delta_{\rm e}\,q^{\rm v}}{m_{\rm e}c^2} \left[\int_{\gamma_{\rm inj\,min}}^{\gamma_{\rm inj\,max}} (\gamma-1)\gamma^{-p}\,{\rm d}\gamma\right]^{-1} \tag{4.18}$$

Apart from $\dot{\gamma}_{\rm adiab}$, the model includes additional cooling rates, $\dot{\gamma}_{\rm syn}$, $\dot{\gamma}_{\rm brem}$, $\dot{\gamma}_{\rm IC}, \dot{\gamma}_{\rm C}$, for synchrotron, bremsstrahlung, inverse Compton scattering, and Coulomb coupling. The expressions for these factors implemented in `KORAL` are taken from Manolakou et al. (2007) and Ginzburg & Syrovatskii (1964), valid in the relativistic limit ($\gamma > 1$),

$$\dot{\gamma}_{\rm syn} = -1.292\times10^{-11}\left(\frac{B}{1\,{\rm G}}\right)^2 \gamma^2\ {\rm s}^{-1}, \tag{4.19}$$

$$\dot{\gamma}_{\rm brem} = -1.37\times10^{-16}\left(\frac{n_{\rm i}}{1\,{\rm cm}^{-3}}\right)\gamma\,(\ln\gamma + 0.36)\ {\rm s}^{-1}, \tag{4.20}$$

$$\dot{\gamma}_{\rm IC} = -3.25\times10^{-8}\left(\frac{\hat{E}}{1\,{\rm erg\,cm}^{-3}}\right)\gamma^2\,F_{\rm KN}(\gamma)\ {\rm s}^{-1}, \tag{4.21}$$

$$\dot{\gamma}_{\rm C} = -1.491\times10^{-14}\left(\frac{n_{e\,{\rm th}}}{1\,{\rm cm}^{-3}}\right)\left[\ln\gamma + \ln\left(\frac{n_{e\,{\rm th}}}{1\,{\rm cm}^{-3}}\right) + 74.7\right]\ {\rm s}^{-1}. \tag{4.22}$$

The inverse Compton cooling rate $\dot{\gamma}_{\rm IC}$ includes a dimensionless Klein-Nishina factor $F_{\rm KN}$ which reduces the cooling rate at high γ. For a thermal distribution of photons at temperature $T_{\rm R}$, this

factor is (Manolakou et al., 2007; Moderski et al., 2005)

$$F_{\rm KN}(\gamma) = \left(1 + 11.2\gamma\frac{k_{\rm B}T_{\rm R}}{m_{\rm e}c^2}\right)^{-3/2}. \tag{4.23}$$

4.1.4 Thermal particle evolution, revisited

The evolution of the thermal ions and electrons is handled the same was as introduced in Chapter 1, Equations 1.28–1.29, but with additional terms added to describe the new interactions with nonthermal electrons. For both species, the thermal entropy per particle ($s_{\rm e}$, $s_{\rm i}$) evolves according to the first law of thermodynamics with source terms:

$$T_{\rm e}(n_{\rm e\,th}s_{\rm e}u^\mu)_{;\mu} = \delta_{\rm e}(1-\delta_{\rm nth})q^{\rm v} + q^{\rm C}_{\rm th} + \hat{G}^0_{\rm th} \tag{4.24}$$
$$+ q^C_{\rm nth} + \left(q^{\rm cool} - \mu\dot{n}^{\rm cool}\right),$$
$$T_{\rm i}(n_{\rm i}s_{\rm i}u^\mu)_{;\mu} = (1-\delta_{\rm e})q^{\rm v} - q^{\rm C}_{\rm th}. \tag{4.25}$$

As in the original, thermal-only Equations 1.28 and 1.29, the first term on the right-hand side in both Equations 4.24,4.25 represents the viscous heating of the thermal populations. As before, the total viscous heating rate $q^{\rm v}$ is identified numerically, modifying Equation 1.30 to account for the adiabatic heating and cooling of the nonthermal electrons (Equation 4.32). The fraction of the viscous heating that goes to the thermal ions is $(1-\delta_{\rm e})$, and the fraction that goes into thermal electrons is $(1-\delta_{\rm nth})\,\delta_{\rm e}$.

The second term in Equations 4.24,4.25 is the thermal Coulomb coupling $q^{\rm C}_{\rm th}$ between the thermal electron and ion populations (Stepney & Guilbert, 1983). The third term in the electron entropy equation is the net radiative power from emission and absorption by the thermal electrons, $\hat{G}^0_{\rm th}$ (Equation 4.7).

The nonthermal population also modifies the electron entropy evolution through a Coulomb coupling term $q^{\rm C}_{\rm nth}$, which is the total energy gained by the thermal electrons due to the Coulomb cooling of the high-energy particles:

$$q^{\rm C}_{\rm nth} = -m_{\rm e}c^2 \int_{\gamma_{\rm min}}^{\gamma_{\rm max}} n(\gamma)\dot{\gamma}_{\rm C}\,{\rm d}\gamma. \tag{4.26}$$

Finally, in order to conserve the total number of electrons, when nonthermal electrons cool below $\gamma_{\rm min}$, they are treated as thermalized and rejoin the thermal distribution. These cooling electrons join the thermal distribution at a rate $\dot{n}^{\rm cool}$, carrying energy density flux $q^{\rm cool}$. The energy and particle cooling rates from the nonthermal distribution to the thermal distribution are simply the flux of energy and particles at the boundary $\gamma_{\rm min}$:

$$\begin{aligned} \dot{n}^{\rm cool} &= -\left[\dot{\gamma}_{\rm tot} n(\gamma)\right]_{\gamma_{\rm min}}, \\ q^{\rm cool} &= -\left[m_{\rm e}\dot{\gamma}_{\rm tot}(\gamma-1) n(\gamma)\right]_{\gamma_{\rm min}}. \end{aligned} \tag{4.27}$$

Note that during adiabatic compression, there can be a (small) flux out of the nonthermal distribution at the top end, $\gamma_{\rm max}$. `KORAL` treats this flux identically to Equation 4.27, adding back the energy and particle number lost over this edge to the local thermal bath. This treatment is unphysical but necessary to conserve total energy among the three species in the simulation. Since the total amount of viscous heating is not increased by this procedure (see Equation 4.32), this choice will not increase the temperature of the thermal electron population above what it would be in a simulation without any nonthermal electrons. In any case, due to the steep power-law shape of the injection functions considered in this study (Equation 4.17), the outward flux at $\gamma_{\rm max}$ is always extremely small.

The expression $\mu\dot{n}^{\rm cool}$, where μ is the chemical potential, accounts for the increase in entropy from

the increase in particle number density. For the chemical potential μ, KORAL uses the following expression derived from the approximate form of the electron entropy per particle $s_{\rm e}$ (Equation 1.24):

$$\mu = m_{\rm e}c^2 \left[1 - \frac{3}{5}\ln\left(1+\frac{5}{2}\Theta_{\rm e}\right) + \Theta_{\rm e}\left(4 - \frac{3}{2}\ln\left(\Theta_{\rm e}^2 + \frac{2}{5}\Theta_{\rm e}\right) + \ln n_{\rm e\,th}\right)\right]. \tag{4.28}$$

4.2 Numerical method

The equations in Section 4.1 are implemented in the GRRMHD code KORAL (Sądowski et al., 2013a, 2014, 2017). The nonthermal electron distribution $n(\gamma)$ is sampled in N equally spaced logarithmic bins over a range $[\gamma_{\rm min}, \gamma_{\rm max}]$. These quantities $n(\gamma_j)$ are treated as N additional primitive quantities evolved in parallel with the other GRRMHD and thermodynamic primitives. The full vector of $15+N$ primitives P consists of the fluid density ρ, energy density u, fluid velocity u^i, magnetic field B^i, radiation energy density $\bar{E}$, radiation frame velocity $u^i_{\rm R}$, photon number $\bar{n}_{\rm R}$, thermal electron and ion entropy densities $s_{\rm e}n_{\rm e\,th}$ and $s_{\rm i}n_{\rm i}$, and the populations $n(\gamma_j)$ of the nonthermal electrons in the N bins:

$$P = [\rho, u, u^i, B^i, \bar{E}, u^i_{\rm R}, \bar{n}_{\rm R}, s_{\rm e}n_{\rm e\,th}, s_{\rm i}n_{\rm i}, n(\gamma_j)], \tag{4.29}$$

where the index j runs over the N bins sampled in γ-space. The corresponding conserved quantities are

$$U = [\rho u^0, T^0{}_0 + \rho u^0, T^0{}_i, B^i, R^0{}_0, R^0{}_i, n_{\rm R}u^0_{\rm R}, s_{\rm e}n_{\rm e\,th}u^0, s_{\rm i}n_{\rm i}u^0, n(\gamma_j)u^0]. \tag{4.30}$$

KORAL uses a Newton-Raphson solver to convert from the conserved quantities to primitives (Sądowski et al., 2013a, 2014). The fluid velocity u^μ is uniquely specified by inverting the MHD conserved quantities, so recovering $n(\gamma_j)$ requires simply dividing the conserved quantity $n(\gamma_j)u^0$ by the already-solved-for u^0.

Fixed floors and ceilings are applied on the evolved quantities as in Sądowski et al. (2013a, 2014, 2017). The floor on the nonthermal distribution is $n(\gamma_j) > 0$ at all γ_j. This floor is especially necessary when beginning from $n(\gamma_j) = 0$, as numerical effects can occasionally make $q^{\rm v}$ negative and bring the nonthermal number values below zero. Fixed ceilings also prevent the nonthermal number and energy densities from exceeding 50% of the total.

KORAL uses a second-order Runge-Kutta scheme to advance the fluid quantities in each time step. Within each Runge-Kutta step, there are three main sub-steps: explicit fluid evolution (Section 4.2.1), nonthermal adiabatic evolution and viscous heating (Section 4.2.2), and implicit radiation coupling (Section 4.2.3).

4.2.1 Explicit fluid evolution

In the explicit sub-step, the covariant conservation equations are evolved without source terms. The conservation equations evolved in this step consist of the GRRMHD equations (1.9–1.12), the photon number equation 1.13, the thermal entropy equation 4.24, and the nonthermal advection equation 4.14, all with their right hand sides set to zero. In particular, the nonthermal bins at each point in γ-space are treated independently and evolved as scalars with the fluid flow. The explicit evolution uses a Lax-Friedrichs method with van-Leer flux limiters to calculate fluxes of the conserved quantities at cell faces. Geometrical terms (i.e. the covariant derivative terms involving Christoffel symbols) are added as source terms at cell centers. The full explicit advective algorithm is described in Chapter 1, and Sądowski et al. (2013a, 2014).

4.2.2 Adiabatic nonthermal evolution and viscous heating

After evolving the bulk fluid quantities explicitly, the nonthermal distribution in each cell is evolved adiabatically through γ-space to provide the appropriate heating or cooling from gas compression or expansion. Then the dissipative heating is calculated and and applied to the thermal and nonthermal species using the viscous heating prescription (Equation 4.32). The steps are as follows:

1. The nonthermal distribution is evolved under adiabatic compression/expansion using the cooling rate $\dot{\gamma}_{\rm adiab}$ in Equation 4.16. From Equation 4.14, after explicit spatial evolution and before dealing with radiative and Coulomb coupling, the change in the nonthermal electron spectrum $n(\gamma_j)$ over a proper time interval $\Delta\tau$ at each bin j in γ-space is

$$\Delta n(\gamma_j) = \Delta\tau \left(\frac{u^\alpha{}_{;\alpha}}{3}\right) \left[\frac{\partial}{\partial\gamma}\left((\gamma - \gamma^{-1})n(\gamma)\right)\right]_j . \tag{4.31}$$

   The expansion parameter $u^\alpha{}_{;\alpha}$ is computed from the u^α obtained at the end of the explicit operator. For numerical stability, the derivative $\partial/\partial\gamma$ is approximated using explicit upwind finite differencing. The upwind direction depends on the sign of the expansion.

   Because the upwind evolution in Equation 4.31 conserves total particle number but not energy, the spectrum $\Delta n(\gamma)$ of particles added or subtracted to the distribution is scaled so that the total change in energy is equal to the amount predicted by Equation 4.33 (see Section 4.2.4).

2. If the expansion $u^\alpha{}_{;\alpha} > 0$, nonthermal electrons may escape out of the lowest bin of the distribution. The loss of energy and particles out of the lowest bin is calculated and added to the thermal distribution number and energy density. Similarly, if $u^\alpha{}_{;\alpha} < 0$, nonthermal electrons may escape out of the highest bin, and the corresponding flux of energy and number density is added to the thermal distribution. The thermal electron entropy per particle $s_{\rm e}$ is recomputed using the updated number and energy density.

3. Since each species has now gone through its full adiabatic evolution, the viscous dissipation rate $q^{\rm v}$ in each cell is computed by comparing the total fluid energy density after the explicit step with the sum of the current species energies (Sądowski et al., 2017),

$$q^{\rm v} = \frac{1}{\Delta\tau}\left(u - u_{i\,{\rm th\,adiab}} - u_{e\,{\rm th\,adiab}} - u_{e\,{\rm nth\,adiab}}\right). \tag{4.32}$$

   Here, u is the internal energy density of the total gas after the explicit step over a fluid frame proper time step $\Delta\tau$. $u_{i\,{\rm th\,adiab}}, u_{e\,{\rm th\,adiab}}$, and $u_{e\,{\rm nth\,adiab}}$ are the internal energy densities carried by thermal ions, electrons, and nonthermal electrons after adiabatic evolution. The

difference between u and the sum of the adiabatically evolved species energy densities gives the total energy gained from viscous dissipation during the time step.

4. The fraction of the viscous heating applied to the electrons $\delta_{\rm e}$, and the fraction of that applied to the nonthermal population $\delta_{\rm nth}$ are calculated depending on the subgrid prescription used.

5. Particles are added to the nonthermal population in a power-law distribution by adding the injection distribution $Q_{\rm I}(\gamma)\Delta\tau$ (Equation 4.17) to each bin.

6. The thermal energy densities $u_{\rm e\ th}$ and $u_{\rm i}$ are increased by their fraction of the remaining viscous heating. The corresponding changes in thermal entropies $s_{\rm e}$ and $s_{\rm i}$ are computed by Equations (1.23–1.24).

4.2.3 Implicit radiation and Coulomb coupling

The source terms representing the radiative and Coulomb coupling between the species are: the radiative coupling G_ν, the thermal Coulomb coupling $q^C_{\rm th}$, the photon source term $\dot{n}_{\rm R}$, the nonthermal cooling rates $\dot{\gamma}$, the nonthermal Coulomb coupling $q^C_{\rm nth}$, and the cooling from the nonthermal population to the thermal bath, $q^{\rm cool}$ and $\mu\dot{n}^{\rm cool}$. These coupling terms in Equations 1.9, 4.14, 4.24 and are applied through a semi-implicit operator using the methods described in Sądowski et al. (2013a, 2014).

The implicit solver uses a reduced set of primitives which includes the energy density and velocity of *either* gas or radiation, the photon number density, electron energy density, and the full nonthermal distribution. The other primitives, including the velocity not evolved, the gas density, and the ion entropy are continually updated during the iterations of the implicit solver to enforce the conservation of total energy among the species.

4.2.4 Energy and particle conservation

The equation that governs the evolution of the nonthermal distribution, Equation 4.14, is a conservation law for a particle current in five dimensions, three spatial, one time, and one corresponding to the individual particle energies. As this equation is evolved via finite volume algorithms, the total number of particles in the distribution is conserved. The total internal energy density in the distribution is given by integrating $n(\gamma)$ times the particle energy $(\gamma-1)m_{\mathrm{e}}c^2$ over γ (Equation 4.4). While particles are not lost from the distribution by evolving Equation 4.14 (excepting boundary effects), energy can be lost from the nonthermal distribution in radiative cooling, Coulomb coupling, or adiabatic expansion; the distribution can also gain energy in adiabatic compression. Because the finite volume form of Equation 4.14 does not conserve the particle energy current, and because KORAL uses a numerical approximation to the integral in Equation 4.4 to compute the internal energy in the nonthermal distribution, the evolution of the nonthermal distribution on its own does *not* conserve total energy.

KORAL accounts for this mismatch in two ways. In the implicit step, where nonthermal electrons lose energy to radiation and Coulomb coupling, it simply ensures that total energy is conserved by adjusting the nonthermal energy flux into radiation ($-\hat{G}^0_{\mathrm{nth}}$) to reflect the energy that is actually lost in cooling the nonthermal distribution. That is, instead of using Equation 4.10, $\hat{G}^0_{\mathrm{nth}}$ is computed by computing the difference in the total nonthermal energy density at a given sub-step in the implicit solver with the energy density computed before the implicit step, subtracting off the small part of the cooling that is due to Coulomb coupling (which goes into thermal electrons). When $\hat{G}^0_{\mathrm{nth}}$ is needed outside of the implicit solver, KORAL uses Equation 4.10. In this way, the total energy is conserved and the shape of the nonthermal distribution is not affected, although the total energy in the distribution may differ from the value computed from an analytic solution or found in a

simulation with finer sampling in γ.

In the intermediate step, where particles are heated or cooled by adiabatic compression or expansion, KORAL cannot account for the missing/extra energy by simply adding it to radiation or the thermal population. The adiabatic heating/cooling of the nonthermal distribution is part of the adiabatic evolution, and correctly computing the viscous heating via Equation 4.32 depends on properly evolving the independent species energies independently and adiabatically. In this case, the distribution is evolved explicitly, and then the computed $\Delta n(\gamma_j)$ at each sampled γ_j is scaled so that the total change in energy during the adiabatic step $\Delta u_{\mathrm{adiab,\,nth}}$ is equal to the amount given by the total instantaneous rate of adiabatic energy increase. That is, the distribution after adiabatic compression/expansion is scaled so that

$$\Delta u_{\mathrm{adiab,\,nth}} = m_{\mathrm{e}} c^2 \int_{\gamma_{\mathrm{min}}}^{\gamma_{\mathrm{max}}} \dot{\gamma}_{\mathrm{adiab}} n(\gamma) \mathrm{d}\,\gamma. \tag{4.33}$$

Scaling the distribution in this way can bias the shape of the distribution (Section 4.3.2). However, the energy gained and lost in this step is applied correctly (Section 4.3.1), and the computation of the viscous heating rate q^{v} is consequently not biased.

4.3 Tests

This section describes several test problems to demonstrate the accuracy of the nonthermal electron evolution as implemented in KORAL.

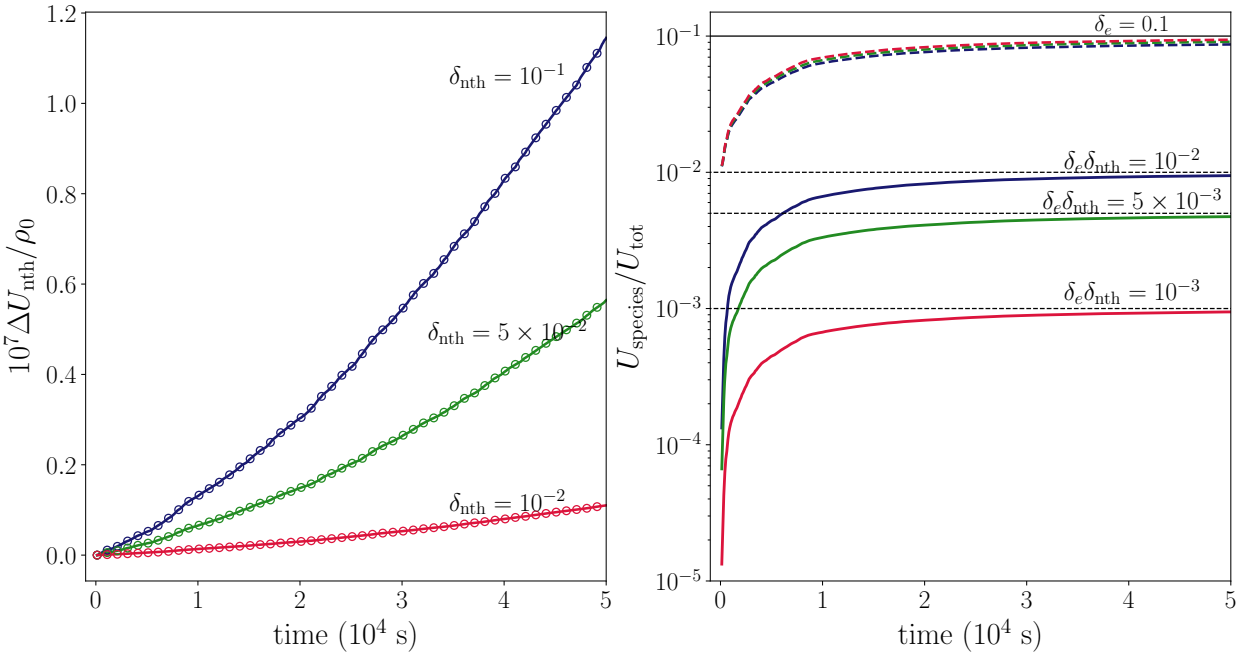


Figure 4.1: (Left) The increase of the total energy of nonthermal electrons $\Delta U_{\rm nth}$ integrated over the turbulent box of Section 4.3.1. The total electron heating fraction was set at $\delta_{\rm e} = 0.1$, and three runs were performed with nonthermal heating fraction $\delta_{\rm nth} = 0.01$ (red), $\delta_{\rm nth} = 0.05$ (green), and $\delta_{\rm nth} = 0.1$ (blue). The open circles indicate the increase of the internal energy of the nonthermal population, and the solid lines show the predicted increase, which is the fraction ($\delta_{\rm e}\delta_{\rm nth}$) of the increase in the total gas energy. (Right) The fraction $U_{\rm species}/U_{\rm gas}$ of the thermal energy of the thermal electrons (dashed lines) and nonthermal electrons (solid lines). As time proceeds and energy from viscous dissipation is divided among the different species, the energy fraction in each species asymptotes to the value given by the corresponding fixed viscous heating injection fractions: $\delta_{\rm e}(1 - \delta_{\rm nth})$ for thermal electrons, and $\delta_{\rm e}\delta_{\rm nth}$ for nonthermal electrons.

4.3.1 Driven turbulence

The first test validates the implementation of adiabatic evolution and viscous heating of the nonthermal population. This test is a modification of the turbulent box test from Sądowski et al. (2017), which was inspired by the MHD driven turbulence test of Ressler et al. (2015). A fraction $\delta_{\rm e} = 0.1$ of the dissipative heating $q^{\rm v}$ is deposited into the electrons, of which a fraction $\delta_{\rm nth}$ goes into the nonthermal population via Equation 4.17. The remaining fraction $\delta_{\rm e}(1-\delta_{\rm nth})$ goes into the thermal electrons by Equation 4.24.

This test begins with an initial uniform two dimensional system of size L with density ρ_0, zero velocity, speed of sound $c_{s0} = 8.6\times10^{-4}c$, a horizontal magnetic field with $\beta = p_{\rm gas}/p_{\rm mag} = 6$, no nonthermal electrons, and periodic boundary conditions. The system is driven with random, divergence-free Gaussian perturbations in the velocity with a power spectrum $P(|\delta v|^2) = k^6\exp(-8k/k_{\rm pk})$, where $k_{\rm pk} = 4\pi/L$. These perturbations add kinetic energy to the system which dissipates into internal energy of the gas, divided among the three species. Radiation and Coulomb coupling are turned off.

The open circles in the left panel of Figure 4.1 shows the resulting increase of the total energy in nonthermal electrons integrated over the simulation volume for three runs with $\delta_{\rm nth}$ =.01, .05, and 0.1, respectively (open circles). The circles are compared with the corresponding fraction $\delta_{\rm e}\delta_{\rm nth}$ of the increase in the total gas energy (solid lines). The close agreement shows that the combination of viscous heating and the change in energy from adiabatic compression and expansion (as a result of turbulence) is handled consistently among the three populations. In particular, the energy normalization performed on the nonthermal distribution during the adiabatic compression/expansion step (Section 4.2.4) is necessary to identify the correct amount of viscous heating and produce the good agreement shown in Figure 4.1.

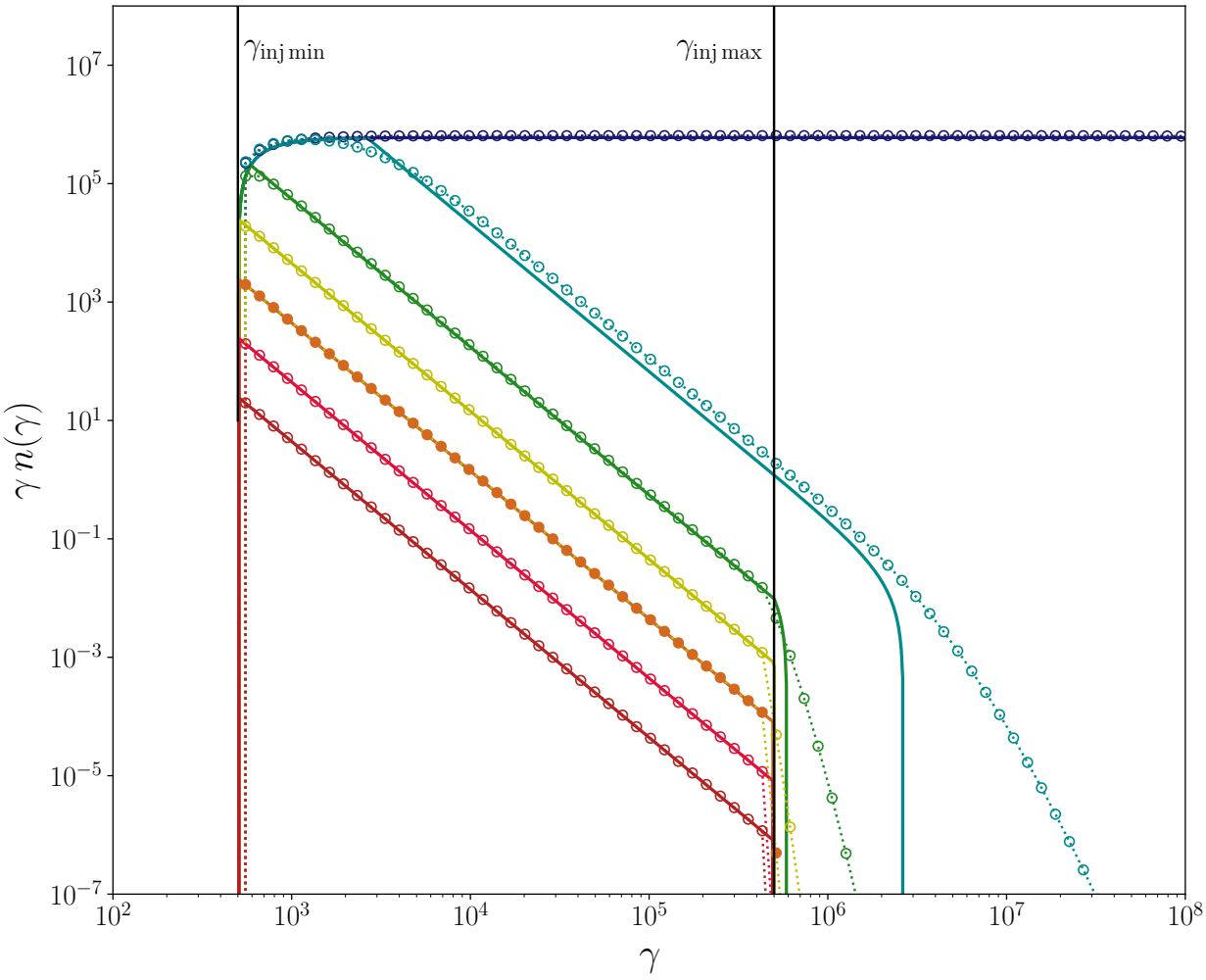


Figure 4.2: Results of a test of adiabatic compression with constant $u^\mu{}_{;\mu} = -5 \times 10^{-3}$ s^{-1} and particle injection with slope $p = 3.5$ between $\gamma_{\rm inj\,min} = 500$ and $\gamma_{\rm inj\,max} = 5 \times 10^5$. The injection distribution is normalized so that the total injection rate is 1000 particles cm^{-3} s^{-1}. The solid lines show the analytic solution to the problem at times $t = 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3$ and 10^4 seconds (progressing upward in $\gamma n(\gamma)$). The open circles show the KORAL solution at the corresponding times.

The right panel of Figure 4.1 shows the ratio of the energy densities of the two electron populations to the total gas energy density: $U_{\rm th}/U_{\rm gas}$, and $U_{\rm nth}/U_{\rm gas}$. As energy is dissipated and divided among the species, the ratios of the species energies to the total internal energy correctly asymptote to the injection fractions $\delta_{\rm e}(1-\delta_{\rm nth})$ and $\delta_{\rm e}\delta_{\rm nth}$ for thermal and nonthermal electrons, respectively.

4.3.2 Particle injection and adiabatic compression

This test validates the implementation of the adiabatic heating and cooling of electrons under gas compression and expansion (Equation 4.16) Consider a zero-velocity gas background with constant injection of nonthermal electrons with a power-law slope $p = 3.5$ between $\gamma_{\rm inj\,min} = 50$ and $\gamma_{\rm inj\,max} = 5 \times 10^5$. The test subjects this system to a constant artificial compression rate (not computed from the actual gas four-velocity) $u^{\mu}{}_{;\mu} = -5 \times 10^{-3}$ s^{-1}, similar to the compression rate found in the equatorial plane at a radius of $\sim 5\,r_{\rm g}$ in the accretion disk simulations described later in Section 4.4. The analytic solution to this problem (Manolakou et al. 2007, Appendix A) shows the development of a break from the injection power-law slope $-p$ to a slope of -1 at low γ, with the break propagating to higher γ with increasing time.

Figure 4.2 shows the results of the test at logarithmically spaced time intervals. The open circles, which denote the KORAL solution, mostly line up well with the analytic result. Deviations arise from two effects. First, the numerical scheme is diffusive and thus smooths out sudden breaks in the slope of $n(\gamma)$. This is seen as a tail above the maximum γ of the true distribution, and also at the break between slope -3.5 to -1, around $\gamma = 3000$ for $t = 1000$ s. Second, due to the smoothing out of breaks in solving Equation 4.14 numerically, energy tends to be lost from the distribution compared with the analytic value. Since KORAL enforces that the total energy is conserved by rescaling the injected particle distribution $Q_{\rm I}$ (Equation 4.33), this diffusion leads to a shift in the normalization (Section 4.2.4). This effect is obvious in the KORAL solution at $t = 10^3$ s. Once the spectrum has broken completely, the KORAL solution matches the analytic solution for all γ. In practice, a fluid parcel in a turbulent simulation will experience many phases of compression and expansion, which may wash out the energy correction effect illustrated in this test.

4.3.3 Synchrotron and inverse Compton cooling

The following two tests in a flat, zero-velocity gas background with constant injection of nonthermal electrons and radiative cooling check the implementation of radiative cooling in KORAL's implicit solver.

In the first test, particles are injected with a power-law slope $p = 3.5$ between $\gamma_{\mathrm{inj\,min}} = 1000$ and $\gamma_{\mathrm{inj\,max}} = 10^6$ and are then subjected to synchrotron cooling in a constant magnetic field of $B = 200$ G. Under constant injection and synchrotron cooling, and for $t < t_{\mathrm{syn}}$, the particle spectrum develops a cooling break from the injection power-law slope $-p$ to a slope $-(p+1)$ at a break Lorentz factor γ_{brk} given by

$$\gamma_{\mathrm{brk}} = (1/\gamma_{\mathrm{inj\,max}} - b_s t)^{-1}, \tag{4·34}$$
$$b_s = 1.292 \times 10^{-9} (B/\,1\,\mathrm{G})^2,$$

where the synchrotron cooling time t_{syn} is

$$t_{\mathrm{syn}} = (\gamma_{\mathrm{inj\,min}}^{-1} - \gamma_{\mathrm{inj\,max}}^{-1})/b_s. \tag{4·35}$$

At $t = t_{\mathrm{syn}}$, the cooling break reaches $\gamma_{\mathrm{inj\,min}}$, and at later times the spectrum cools to $\gamma < \gamma_{\mathrm{inj\,min}}$ with a power-law slope of -2.

The results from KORAL are compared with the analytic solution in Figure 4.3. The development of the synchrotron cooling break and its propagation to lower particle energies with time is clearly captured in the KORAL solution. The numerical solution from KORAL cannot capture the sharp cutoff at low particle energies, and it produces a tail extending to low γ (note that the vertical scale is over 14 orders of magnitude, so the discrepancy is not serious). However, the location of the peak

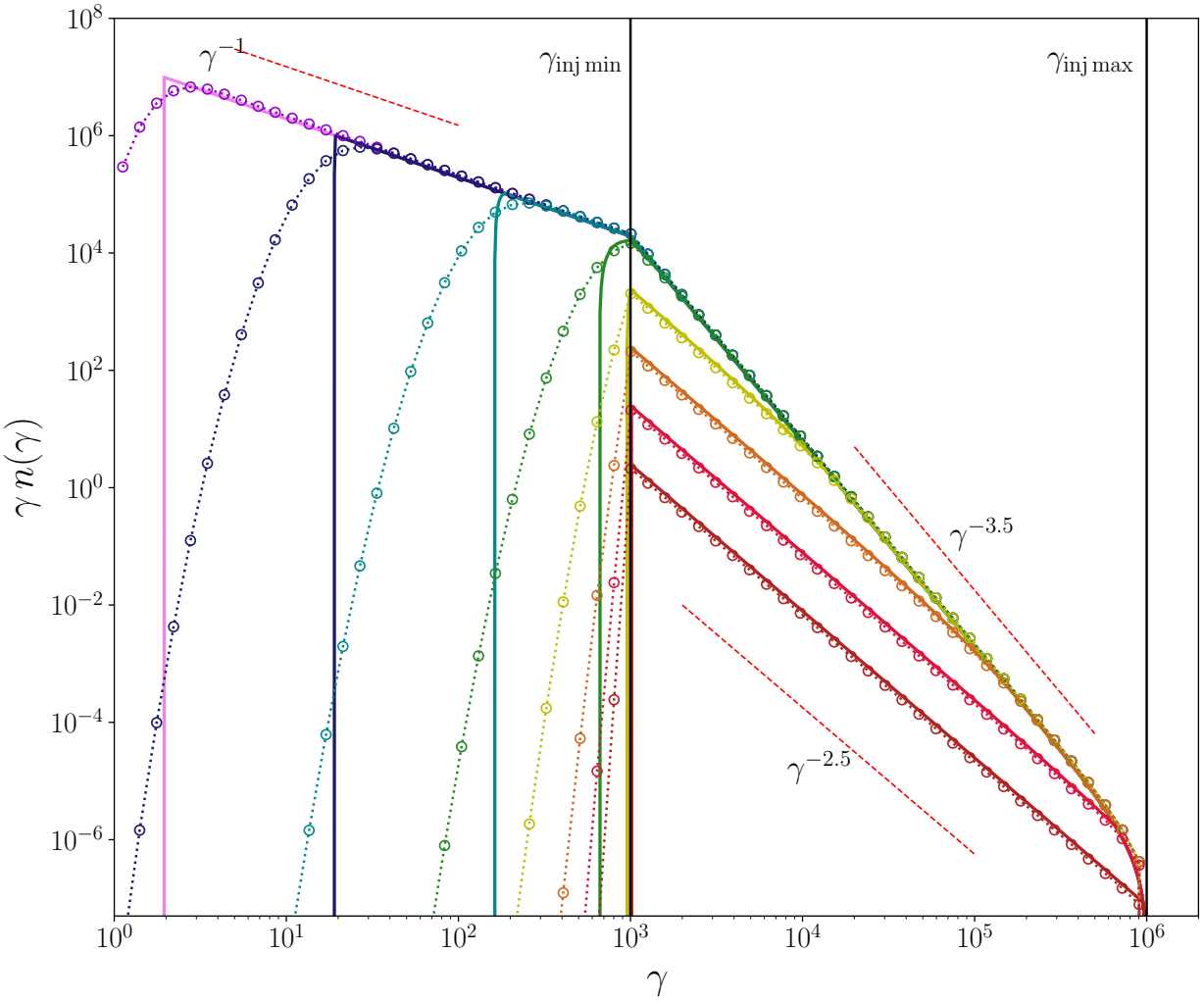


Figure 4.3: Results of a test with constant particle injection with slope $p = 3.5$ between $\gamma_{\rm inj\,min} = 1000$ and $\gamma_{\rm inj\,max} = 10^6$ coupled with synchrotron cooling in a uniform magnetic field with $B = 200$ G. The injection distribution is normalized so that the total injection rate is 1000 particles cm^{-3} s^{-1}. The numerical solution from KORAL's implicit solver (open circles) is compared with the analytic solution (solid lines) at times $t = 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3$ and 10^4 seconds. The spectrum develops a cooling break between the injection slope p for $\gamma < \gamma_{\rm brk}$ and $p + 1$ for $\gamma > \gamma_{\rm brk}$. The cooling break starts at large γ and propagates toward lower γ until the spectrum is broken over the entire injection range at $t = 10$ s. After this time, the spectrum cools to $\gamma < \gamma_{\rm inj\,min}$ with slope $p = 2$. The sharp discontinuity at the lower end of $n(\gamma)$ is smeared out in the numerical KORAL solution because of diffusion in the upwind finite differencing method. However, KORAL accurately captures the location of the peak of $\gamma n(\gamma)$ as it propagates to lower energies.

in the spectrum as a function of time is reproduced well.

The next test of the KORAL implicit solver for radiative nonthermal cooling replicates a problem from Manolakou et al. (2007) that demonstrates the effects of the Klein-Nishina cross section in the inverse Compton cooling term (Equations 4.21 and 4.23). Neglecting bremsstrahlung radiation and Coulomb coupling, the cooling rate is

$$\dot{\gamma} = b_{\rm syn}\gamma^2 \left[1 + \frac{u_{\rm rad}}{u_{\rm mag}} \left(1 + 4\gamma\epsilon_0\right)^{-3/2}\right], \tag{4.36}$$

where $b_{\rm syn} = -1.292{\times}10^{-11}\,(B/1\,{\rm G})^2$ and $\epsilon_0 = k_{\rm B}T_{\rm R}/m_{\rm e}c^2$. The test assumes a uniform background similar to a stellar environment dominated by hot young stars (Manolakou et al., 2007), with $B = 10\ \mu$G, $u_{\rm rad} = 7.95 \times 10^{-10}$ erg cm^{-3}, and $T_{\rm rad} = 30000$ K. Electrons are injected in a power law with $p = 2$ between $\gamma_{\rm inj\,min} = 100$ and $\gamma_{\rm inj\,max} = 10^9$, normalized so that the total injection rate is 10^{-3} particles cm^{-3} s^{-1}.

The results are displayed in Figure 4.4 at times $t = 10^5$, 5×10^5, and 10^6 yr. The KORAL solution (open circles) lines up well with the semi-analytic solution (Manolakou et al., 2007, solid lines;), demonstrating the code's ability to accurately capture details of the radiative cooling of nonthermal distributions beyond simple synchrotron cooling.

The solution in this test displays different behavior in three distinct regimes. From Equation 4.36, at the highest energies, $\gamma > \gamma_{\rm syn} = \left((u_{\rm rad}/u_{\rm mag})^{2/3} - 1\right)/4\epsilon_0$, the solution is dominated by synchrotron cooling. Hence the spectrum shows a characteristic synchrotron cooling break above $\gamma_{\rm syn}$, where the slope becomes $-(p+1) = -3$. Equation 4.36 also indicates that below $\gamma_{\rm KN} \approx 1/4\epsilon_0$, the Thomson limit applies. Between $\gamma_{\rm KN}$ and $\gamma_{\rm syn}$, the decrease in the cooling rate due to the Klein-Nishina cross section causes the spectrum to harden compared to what is predicted when only Thomson scattering is considered (dotted lines in Figure 4.4). As time progresses, electrons

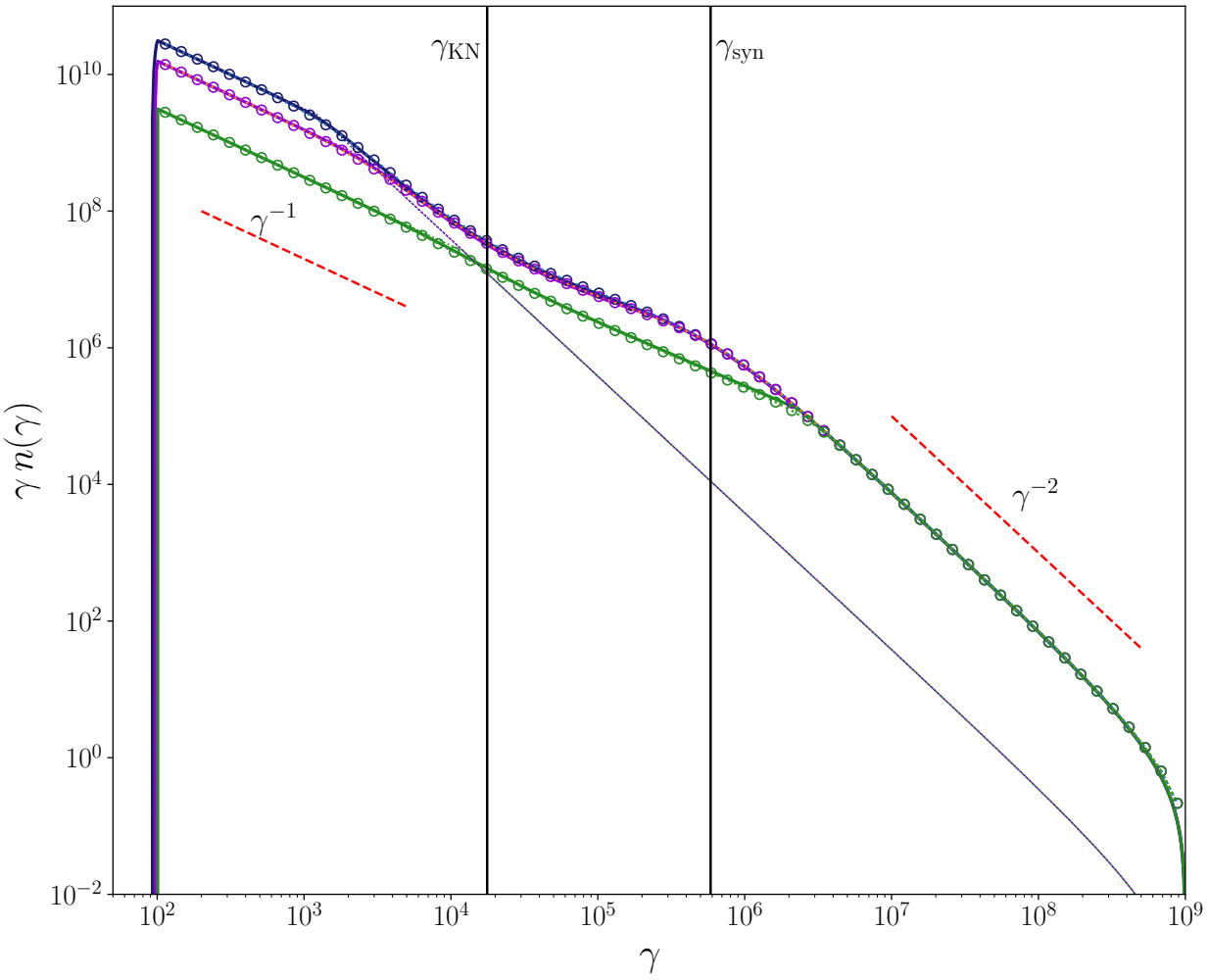


Figure 4.4: Nonthermal energy distribution evolution in an environment with $B = 10\,\mu$G, $u_{\rm rad} = 8.01 \times 10^{-10}$ erg cm^{-3}, and $T_{\rm rad} = 30000$ K. Particles are injected in a power law with $p = 2$ between $\gamma_{\rm inj,min} = 100$ and $\gamma_{\rm inj,max} = 10^9$. The numerical solution from KORAL's implicit solver (open circles) is compared with the semi-analytic solution (solid lines) at times $t = 10^5$ (green), 5×10^5 (purple) and 10^6 yr (blue). The analytic solution for the same problem neglecting the Klein-Nishina cross section of electrons (taking $F_{\rm KN} = 1$ in Equation 4.21) is also displayed (dotted lines).

initially injected at $\gamma_{\rm inj\,max}$ cool to lower energies $\gamma_{\rm cool}$. By the last time shown, $\gamma_{\rm cool} < \gamma_{\rm KN}$; electrons injected at the highest energies have cooled below the energies where the Klein-Nishina cross section dominates, and Thomson cooling begins to break the spectrum for $\gamma < \gamma_{\rm KN}$.

Because the cooling rates in this problem are so low, even over 10^6 years of evolution the spectrum does not have time to cool much below the injection range. Therefore, the implicit solver does not have to deal with abrupt discontinuities in the spectrum, and except for the slight smoothing out of the synchrotron break, the obvious diffusion seen in the tests in Figures 4.2 and 4.3 is not apparent here.

4.4 Test simulation of Sgr A*

As a test of the entire code with all elements included, this section presents 2D simulations of an accreting supermassive black hole with parameters appropriate for the accretion flow in Sgr A*. Two simulations were run using a pure thermal model with a two-temperature plasma (two fluid populations, thermal ions and thermal electrons, similar to Sądowski et al. 2017), and a nonthermal model with all three fluid populations (thermal ions, thermal electrons and nonthermal electrons, using the full method of Section 4.1.1).

For the injection properties of the nonthermal population these simulations use a very simple ad hoc prescription, which will need to be improved in the future for modeling real systems. A constant fraction of the local viscous electron heating rate goes into nonthermal electrons, with a fixed energy spectrum that is independent of location in the simulation box. The observed infrared and X-ray variability of Sgr A*(Dodds-Eden et al., 2011; Neilsen et al., 2013) suggests that the nonthermal acceleration mechanism is localized, either in magnetic reconnection regions (Sironi & Spitkovsky, 2014) or in shocks (Guo et al., 2014). Recently, Ball et al. (2016) showed that the

X-ray variability of Sgr A*could be qualitatively reproduced by adding a nonthermal distribution by hand in regions of high magnetization in a single-fluid GRMHD simulation. A future work will consider full 3D GRMHD+nonthermal electron simulations using more elaborate localized injection prescriptions informed by these studies (e.g., using the prescription for p and $\delta_{\rm nth}$ from the magnetic reconnection PIC simulations of Ball et al. 2018).

4.4.1 Units

These simulations assume a Schwarzschild spacetime (spin $a = 0$) with black hole mass $M = 4 \times 10^6\, M_\odot$, near the measured mass of Sgr A*(GRAVITY Collaboration et al., 2018a). The gravitational radius $r_{\rm g} = GM/c^2 = 6 \times 10^{11}\,{\rm cm} = 0.04\,{\rm AU}$, and the gravitational timescale $t_{\rm g} = r_{\rm g}/c = 20\,{\rm s}$. The Eddington luminosity $L_{\rm Edd} = 5 \times 10^{44}\ {\rm erg\ s^{-1}}$, and the Eddington accretion rate (Equation 0.3, with $\eta = 0.057$) is $\dot{M}_{\rm Edd} = 0.16\, M_\odot\ {\rm yr}^{-1}$.

4.4.2 Model setup

The simulations in this section used in Kerr-Schild coordinates with an axisymmetric 2D grid of resolution of 256×256 cells in radius and polar angle. The radial cells are distributed exponentially from inside the BH horizon at $1.85\, r_{\rm g}$ to 1000 $r_{\rm g}$, and the polar angle cells are sampled using the transformation described in Appendix B.

The initial fluid conditions are identical to the model `Rad8SMBH` in Sądowski et al. (2017). The simulation starts with a hydrostatic equilibrium torus with an inner edge at $10\, r_{\rm g}$ and is threaded by a weak magnetic field with dipolar field loops. The initial electron and ion temperatures are set equal to the initial gas temperature, and there are no nonthermal electrons. In the initial configuration, the torus is surrounded by a static atmosphere with negligible mass and radiation energy density, but with the radiation temperature everywhere set to 10^5 K.

The model ran for a total time of $2 \times 10^4\, t_g$ with nonthermal electron evolution turned off. The thermal electron and ion populations were heated using the viscous heating prescription of Howes (2010). They exchanged energy with each other via thermal Coulomb coupling, and the thermal electrons radiated via synchrotron, bremsstrahlung, and inverse Compton scattering. Throughout, the mean-field dynamo from Sądowski et al. (2015) prevents the decay of the axisymmetric magnetic field. This purely thermal simulation is referred to as the *control run*.

In the control run, gas begins accreting on the black hole around $t \approx 3000\, t_g$, and by $t = 10^4\, t_g$, the accretion is in steady state. At this time, the gas density and magnetic field strength were scaled to achieve the desired accretion rate of $\dot{M} \approx 4 \times 10^{-8} \dot{M}_{\rm Edd}$ appropriate for Sgr A*. The simulation was evolved with the rescaled density from $t = 10^4\, t_g$ up to $2 \times 10^4\, t_g$. The data from the time period $1.5 \times 10^4 - 2 \times 10^4\, t_g$ were then used to study the properties of the accretion flow.

A system with nonthermal electrons included (*nonthermal run*) was initialized with the rescaled output from the control run at time $10^4\, t_g$ and was evolved from $t = 10^4\, t_g$ up to $2 \times 10^4\, t_g$ with all the nonthermal interactions turned on. The nonthermal electron energy distribution is tracked over $N = 32$ bins ranging from $\gamma_{\rm min} = 200$ to $\gamma_{\rm max} = 2 \times 10^6$, a resolution of 8 bins per decade. $\gamma_{\rm min}$ is set above the characteristic electron energy $\Theta_e = k_B T_e / m_e c^2$ for a temperature at the high end of the range observed in the control model, around $T_e \sim 10^{12}$ K.

For the nonthermal injection, the power-law index is fixed at $p = 3.5$, consistent with past studies (Özel et al., 2000; Yuan et al., 2003) and with observational constraints (Porquet et al., 2008; Barrière et al., 2014). Electrons are injected between $\gamma_{\rm inj\,min} = 500$ and $\gamma_{\rm inj\,max} = \gamma_{\rm max} = 2 \times 10^6$. The nonthermal heating fraction was fixed at 1.5%, $\delta_{\rm nth} = 0.015$ (Özel et al., 2000; Yuan et al., 2003; Ball et al., 2016; Mao et al., 2016). The total electron heating fraction δ_e, of which 98.5% goes to the thermal species, is determined using the prescription of Howes (2010), which is a (strong) function of the magnetization parameter $\beta_i = p_{\rm gas}/p_{\rm mag}$ (Equation 1.33).

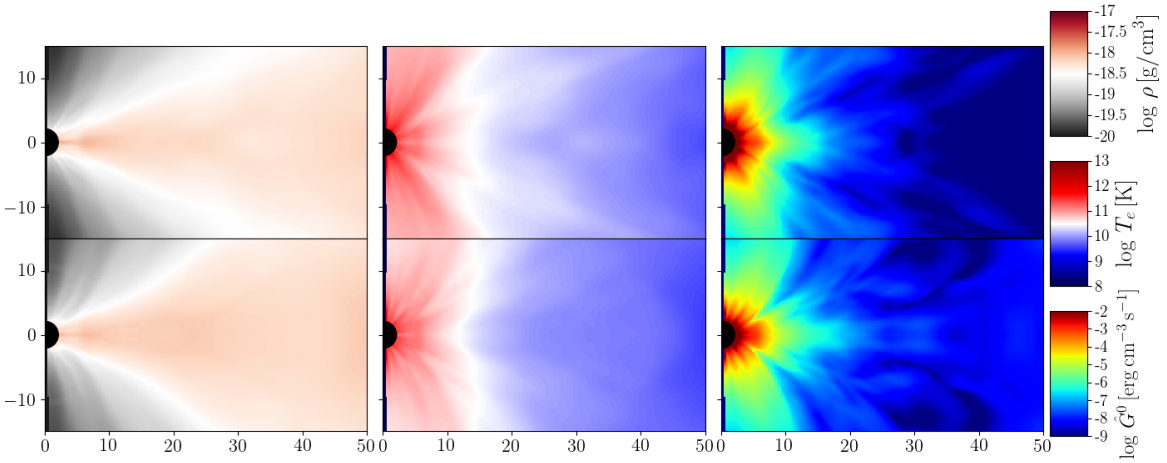


Figure 4.5: Comparison of time-averaged quantities in the control thermal run (top row) and the nonthermal run (bottom row). Averages are taken over the period $t = 1.5 \times 10^4 - 2 \times 10^4\, t_g$, and the distributions are symmetrized about the equatorial plane. In each column, the same color scale is used in the upper and lower panels. From left to right, the quantities shown are the gas density ρ, the electron temperature T_e, and the fluid frame radiation power $-\hat{G}^0$. The presence of nonthermal electrons does not significantly affect either ρ or T_e. However, the nonthermal model has significantly more radiative power, especially at larger radii.

4.4.3 Comparison of thermal and nonthermal models

Figure 4.5 compares time-averaged spatial distributions of several quantities in the control (thermal) run with those in the nonthermal run. For each model, the quantities are averaged over the time range $t = 1.5 \times 10^4 - 2 \times 10^4\, t_g$, and also symmetrized around the equatorial plane for additional smoothing of the results.

Figure 4.5 indicates that the overall structure and distribution of the gas density and electron temperature are similar in the two models. This is expected, since the fraction of electron energy that goes into the nonthermal electrons is only 1.5%. Furthermore, the accretion flow in these simulations is optically thin and radiatively inefficient, so the emission from the nonthermal electrons does not significantly alter the gas dynamics. Indeed, the gas dynamics and electron and ion thermodynamics in both the control run and the nonthermal run are quite similar to the thermal model `Rad8SMBH` presented in Sądowski et al. (2017).

The last column of Figure 4.5, however, shows that the rest frame power of the emitted radiation

is not the same in the control and nonthermal runs — it is enhanced in the latter run, most significantly at large radii. The spatial distribution of the nonthermal emission is purely the result of the particular injection prescription used. Nonthermal electrons start with the same energy fraction $\delta_{\rm nth} = 0.015$, in the same power-law distribution, everywhere in the simulation. In addition, the magnetic field strength is fairly constant ($B \sim 10\,{\rm G}$) over much of the region of interest. Therefore, the amount of nonthermal synchrotron emission is directly proportional to the viscous heating rate of the gas. On the other hand, the thermal electron temperature varies substantially with radius, falling to below $\sim 10^{10}$ K by a radius of $30\,r_{\rm g}$. Since thermal synchrotron power varies as $T_{\rm e}^2$, the thermal emission falls rapidly with increasing radius. Thus, the thermal electrons are more advection-dominated at large radii compared to the nonthermal electrons.

4.4.4 Nonthermal simulation properties

Figures 4.6 and 4.7 show time-averages and snapshots of several quantities in the nonthermal run. The snapshot in these comparisons (right side of each panel) corresponds to $t = 1.8 \times 10^5\,t_{\rm g}$ and the time-averaging (left side of each panel) is done from $t = 1.5 \times 10^4 - 2 \times 10^4\,t_{\rm g}$.

The top panel in Figure 4.6 shows the gas density ρ. As expected, and as seen also in the control run (Figure 4.5), the disk is geometrically thick and turbulent, the latter evident in the snapshot density distribution (even more so in the temperature distribution discussed next). The blue contour corresponds to the location where the accretion time-scale $t_{\rm acc}$ in the time-averaged model is equal to the time-averaging duration $5000\,t_{\rm g}$. The accretion time scale is defined as

$$t_{\rm acc} \equiv \frac{r}{\sqrt{v_r^2 + r^2 v_\theta^2}}. \tag{4.37}$$

Since the total duration of the nonthermal run is $10^4\,t_{\rm g}$, the above limit is a conservative estimate

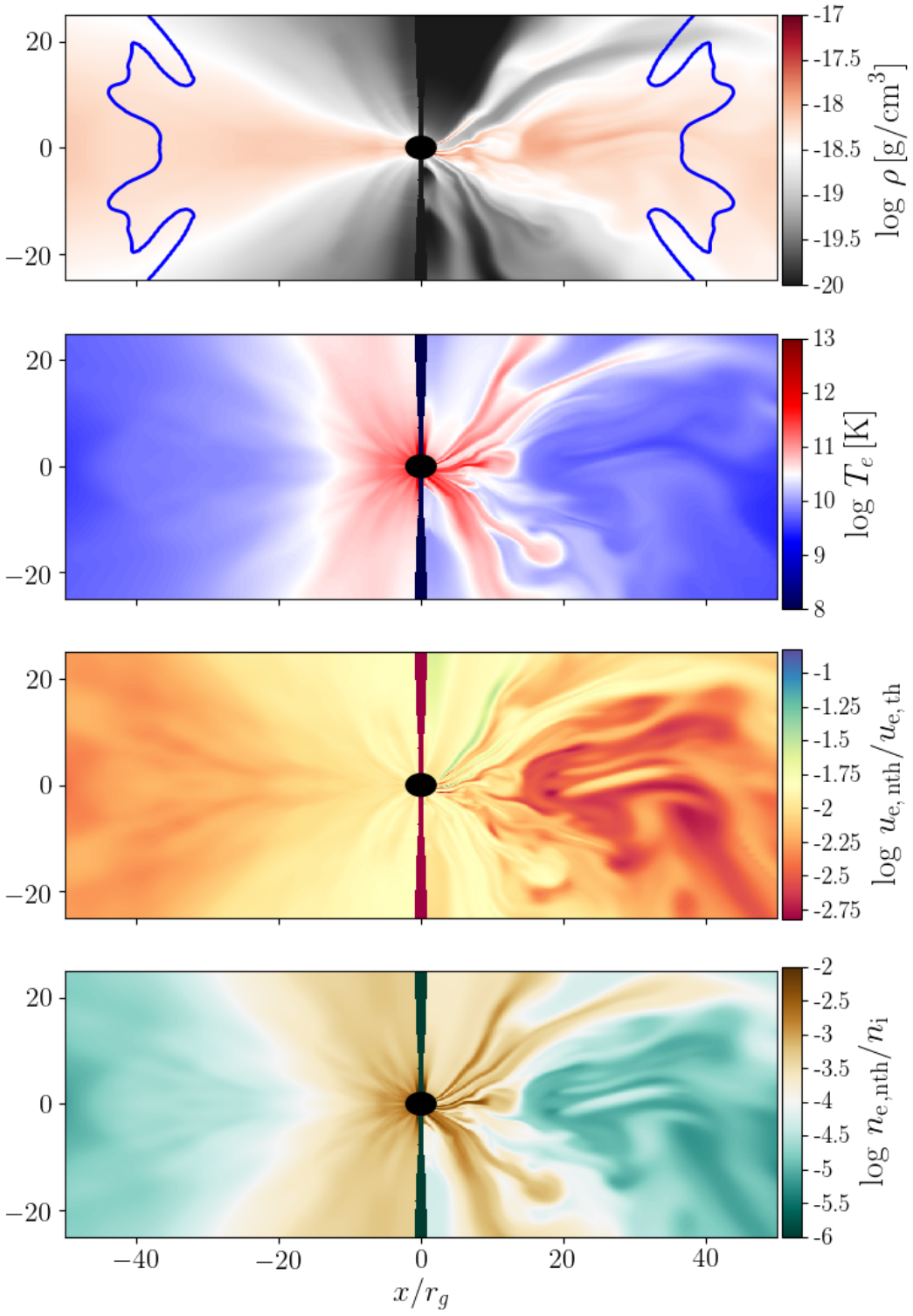


Figure 4.6: Snapshot (right) and time-averaged (left) distributions of (from top to bottom) gas density ρ, thermal electron temperature $T_{\rm e}$, ratio of nonthermal to thermal electron energy densities $u_{\rm e\,nth}/u_{\rm e\,th}$, and fraction of electrons in the nonthermal distribution $n_{\rm e\ th}/n_{\rm i}$. The blue contour in the first panel encloses the region of the simulation that is in inflow equilibrium, as determined by Equation 4.37.

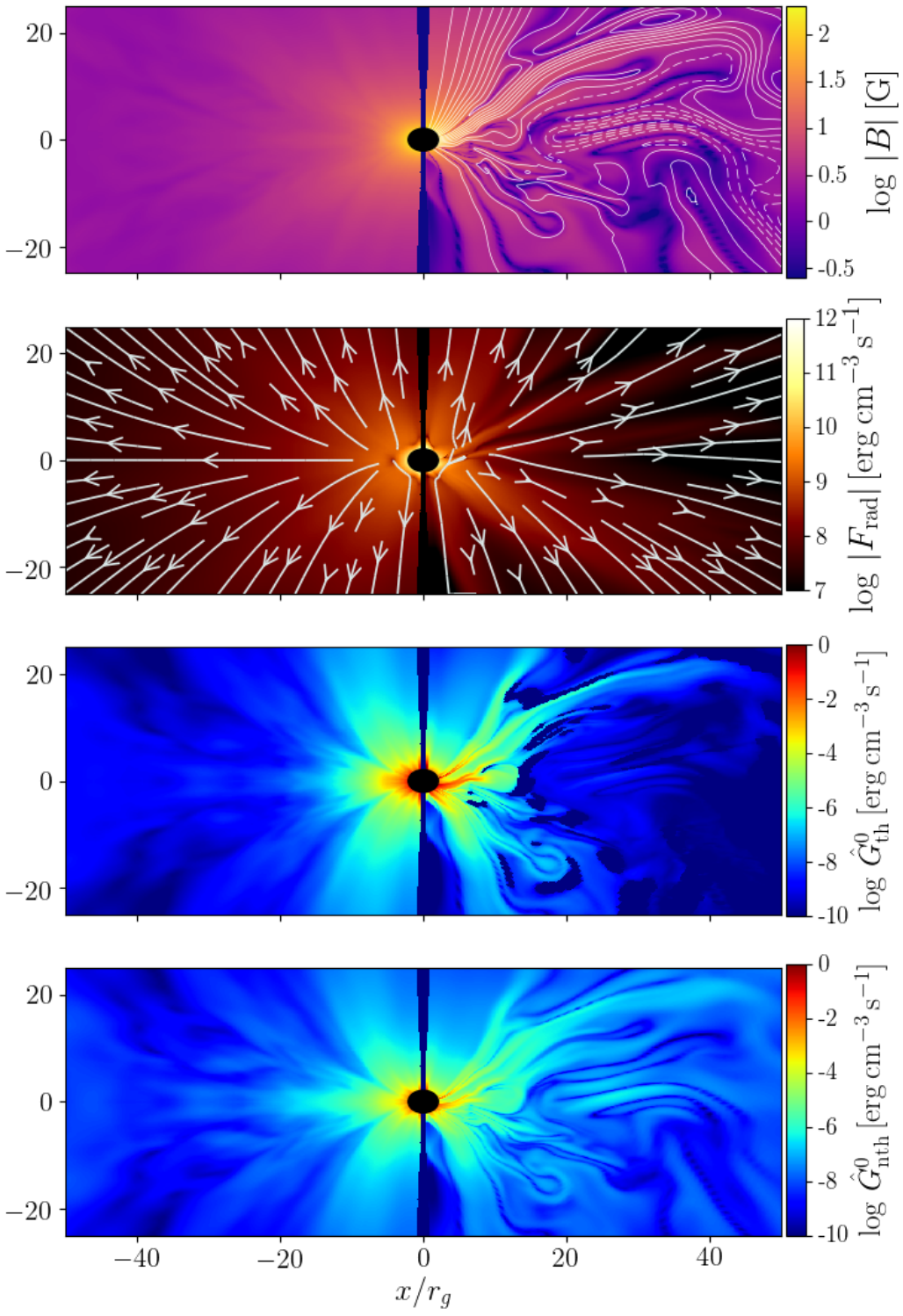


Figure 4.7: Snapshot (right) and time-averaged (left) distributions of (from top to bottom) magnetic field strength $|B|$, magnitude of the radiation flux $|F|$, fluid frame radiation power from thermal electrons $-\hat{G}^0_{\rm th}$, and fluid frame radiation power from nonthermal electrons $-\hat{G}^0_{\rm nth}$. Contours in the first panel show poloidal magnetic field lines. Streamlines in the second panel show the direction of the radiation flux.

of the region of inflow equilibrium (it corresponds to the 'strict' criterion, as defined in Narayan et al. 2012).

The second panel in Figure 4.6 shows the electron temperature, which ranges from $\sim 10^{10}$ K in the disk at $r \approx 30\, r_{\rm g}$ to 10^{12} K in the funnel region. In very low accretion rate systems such as Sgr A*, both radiative cooling and Coulomb coupling are weak and neither is capable of controlling the electron temperature (Yuan & Narayan, 2014). The temperature is thus primarily determined by the viscous heating and is highly dependent on the heating fraction $\delta_{\rm e}$. As in the 3D simulations in Chapter 2, The Howes (2010) prescription has high $\delta_{\rm e} \approx 1$ in regions of high magnetization, which explains the high temperature in the polar region (where $\beta_{\rm i} < 1$) compared to the equatorial plane (where typically $\beta_{\rm i} > 5$).

The third panel in Figure 4.6 shows the ratio of the energies in nonthermal and thermal electrons. Since the radiative and Coulomb coupling between the two species is weak, the energy ratio should be set primarily by the injection ratio $\delta_{\rm nth}$, which was fixed at 1.5% throughout. In much of the equatorial plane out to $r \approx 30\, r_{\rm g}$, the energy ratio is indeed approximately equal to $\delta_{\rm nth}$. Regions where the ratio is lower than $\delta_{\rm nth}$ correspond to places where the electron temperature is lowest. In these regions, the overall electron heating fraction $\delta_{\rm e}$ is small and there has not been enough injection of nonthermal particles to bring the energy up to the injection value. Conversely, in the snapshot distribution, some regions have a nonthermal-to-thermal energy ratio exceeding $\delta_{\rm nth}$. In these regions, the thermal electrons are heated to high temperatures $\sim 10^{12}$ K. At these temperatures, the thermal electrons that produce most of the synchrotron emission have Lorentz factors $\gamma > 500$, greater than the minimum $\gamma_{\rm inj\,min}$ of the injected nonthermal electrons. These high-γ thermal electrons lose energy rapidly to radiation more rapidly than electrons at the peak of the nonthermal distribution.

Finally, the fourth panel in Figure 4.6 shows the overall fraction of the electron population that

is in the nonthermal distribution. In the snapshot image, regions with a high ratio of nonthermal electrons to the total population are coincident with regions of high thermal electron temperature (second panel). This is because both distributions are primarily driven by the fraction $\delta_{\rm e}$ of electron viscous heating (since $\delta_{\rm nth}$ is fixed).

Figure 4.7 displays quantities related to the cooling and radiation from nonthermal particles. The top panel shows the magnetic field strength, which is on average $\sim 10\,{\rm G}$ throughout much of the region in inflow equilibrium ($r \lesssim 40\,r_{\rm g}$). However, the snapshot on the right shows considerable evidence for turbulence and deviations from the mean. Regions with a stronger magnetic field in the snapshot image correlate with regions of higher thermal electron temperature (second panel of Figure 4.6); this is expected since the electron energy injection fraction $\delta_{\rm e}$ increases with magnetization. In addition, since the nonthermal injection rate is proportional to $\delta_{\rm e}$, the same regions also stand out in the snapshot distribution in the fourth panel of Figure 4.6.

The second panel in Figure 4.7 shows the magnitude of the radiation flux $\hat{F}^i$, represented by the color scale, with streamlines indicating the direction of the flux vector. Since the accretion flow is highly optically thin, radiation is emitted more-or-less isotropically and freely streams out of the system.

The third and fourth panels in Figure 4.7 show the fluid frame power in radiation from thermal and nonthermal electrons, respectively. The thermal emission dominates in the inner regions up to $r \sim 10\,r_{\rm g}$, and then declines rapidly at larger radii where the electrons are cooler. However, as previously discussed, highly energetic nonthermal electrons are present even at large radii, because of the simple injection prescription. Therefore, there is significant nonthermal synchrotron emission out to $r \sim 50\,r_{\rm g}$. The two snapshot panels show that the instantaneous radiation power in both thermal and nonthermal emission traces the regions of strongest magnetic field in the top panel.

4.4.5 Synchrotron break

The dominant physical processes shaping the evolution of the nonthermal electron energy distribution $n(\gamma)$ in the nonthermal simulation are electron injection and synchrotron cooling. As the nonthermal particles cool via synchrotron emission, the spectrum will break from the injection power-law slope $-p = -3.5$ to $-(p+1) = -4.5$. The $\gamma_{\rm brk}$ at which the break occurs moves to lower values with increasing time. From Equation 4.35, under constant injection and given a characteristic magnetic field strength of $B \sim 10\,{\rm G}$, $\gamma_{\rm brk}$ will move all the way down to $\gamma_{\rm inj\,min}$ in 1.5×10^4 s, or $780\,t_{\rm g}$. However, in the actual simulation, non-constant particle injection rates, adiabatic compression, and advection modify the development of the synchrotron break and can locally shift the break Lorentz factor to higher γ, with advection having the strongest effect.

The top panel of Figure 4.8 plots the ratio of the synchrotron cooling time $t_{\rm syn}$ (Equation 4.35) to the accretion time-scale $t_{\rm acc}$ (Equation 4.37). This ratio is > 1 almost everywhere in the region considered, which indicates that, before the spectrum can break fully, the gas is advected away or falls into the black hole.

The second panel of Figure 4.8 shows the Lorentz factor $\gamma_{\rm brk}$ of the synchrotron cooling break in the nonthermal distribution. By the late times considered, the cooling break has propagated to low Lorentz factors, but since the accretion time-scale is shorter than the cooling time-scale, the break still lies above $\gamma_{\rm inj\,min}$. In much of the disk, the break Lorentz factor $\gamma_{\rm brk} \sim 3000$. In the funnel region, gas moves with high velocities either into the BH or out along the axis; the corresponding small inflow/outflow (advection) time-scale means that electrons do not have enough time to cool before being swept away. Thus, the break Lorentz factors in the funnel are typically higher than in the rest of the simulation, $\gamma_{\rm brk} \sim 10^4$.

In the time-averaged distribution, the ratio of synchrotron to advection times can provide a quick

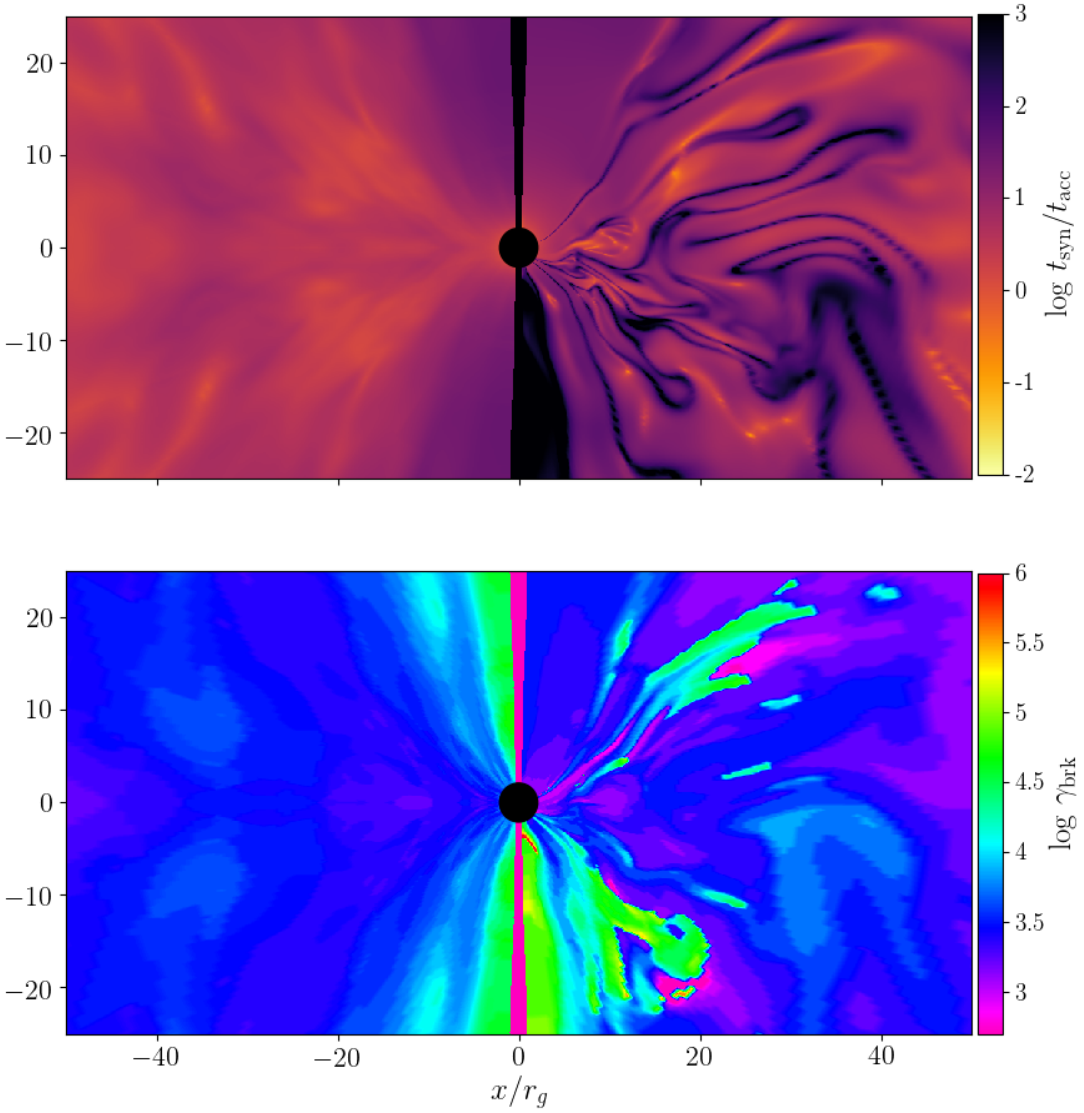


Figure 4.8: (Top) Ratio of synchrotron cooling time-scale to accretion time-scale, $t_{\rm syn}/t_{\rm acc}$, for a snapshot at $t = 1.8 \times 10^4\, t_{\rm g}$ (right) and corresponding ratio computed from time-averaged primitives (left). (Bottom) Location of the synchrotron cooling break Lorentz factor $\gamma_{\rm brk}$. The cooling break is at higher γ in regions where $t_{\rm syn}/t_{\rm acc}$ is large. Electrons in such regions are advected away before they can be cooled by the magnetic field.

estimate of the break Lorentz factor. In the funnel regions, where $t_{\rm syn}/t_{\rm acc} \approx 100$, the position of the break is estimated by substituting $t_{\rm syn}/100$ in Equation 4.34; the result is $\gamma_{\rm brk} \approx 5 \times 10^4$. A comparison with the second panel shows that this quick estimate is reasonably good.

The snapshot distribution of break Lorentz factor shows more structure than the average. Much of this structure is due to the turbulent magnetic field, which creates regions of short and long synchrotron cooling times. However, the regions with high $\gamma_{\rm brk}$ do not always have a one-to-one correspondence with regions of large $t_{\rm syn}/t_{\rm acc}$ (see e.g., around $x = 20\,r_{\rm g}$, $z = 10\,r_{\rm g}$). In addition to synchrotron cooling and advection, other processes – particularly adiabatic compression – can shape the spectrum. Compression acts to push the entire distribution to higher γ, so it naturally pushes the break Lorentz factor to a higher γ than predicted by Equation 4.34.

4.4.6 Spectra and images

Spectral energy distributions (SEDs) and images for both the thermal-only control model and the full nonthermal model were computed using `grtrans`, modified to compute the nonthermal synchrotron emissivity j_ν and absorption coefficient α_ν directly from the local magnetic field and the appropriate integrals over the nonthermal electron energy distribution $n(\gamma)$ (Rybicki & Lightman 1979, equations 6.33 and 6.50).[4] The integrals for j_ν and α_ν are

$$j_\nu = \frac{\sqrt{3}}{4\pi^2}\frac{e^3 B \sin\alpha}{mc^2}\int n(\gamma) F\left(\frac{\nu}{\nu_c}\right) {\rm d}\gamma, \tag{4.38}$$

$$\alpha_\nu = \frac{4\pi}{3\sqrt{3}}\frac{e}{B\sin\alpha}\int \frac{n(\gamma)}{\gamma^5} K_{5/3}\left(\frac{\nu}{\nu_c}\right) {\rm d}\gamma. \tag{4.39}$$

[4]To derive Equation 4.39 from Rybicki & Lightman (1979) equation 6.50, perform an integration by parts and discard the boundary term. For recent work on integrating polarimetric synchrotron emissivities from various electron distribution functions see Leung et al. 2011 and Pandya et al. 2016.

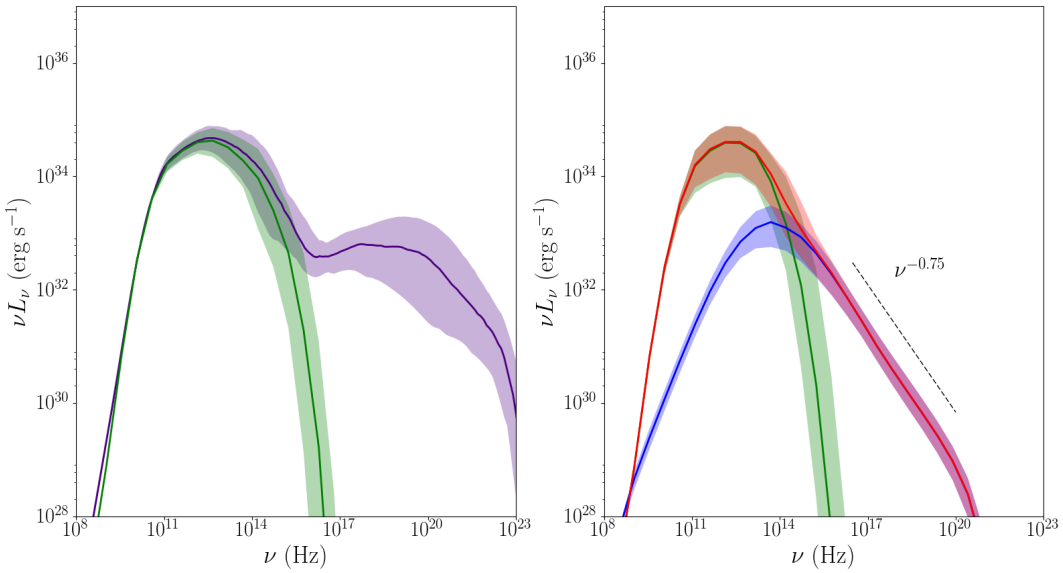


Figure 4.9: (Left Panel) Median spectral energy distribution (solid lines) of the thermal control run computed from the snapshot data from 15,000-20,000 $t_{\rm g}$, as observed at an angle of 60° with respect to the disk polar axis. The shaded regions represent the 68% confidence interval (nominal 1σ range) for the time-variability of the spectra in this interval. The green spectrum is obtained using `grtrans` (Dexter, 2016), which includes only synchrotron radiation. The indigo spectrum was computed with `HEROIC` (Narayan et al., 2016), including bremsstrahlung emission and inverse Compton scattering. (Right Panel) Synchrotron-only spectra of snapshots from the nonthermal simulation in the range 15,000-20,000 $t_{\rm g}$ computed with `grtrans`. The green and blue lines show the spectra of the thermal and nonthermal electrons, respectively, and the red line shows the total spectrum of both populations combined. The dashed line shows the expected power-law slope produced by the broken spectrum of nonthermal electrons, $L_\nu \propto \nu^{-p/2} \propto \nu^{-1.75}$.

In the above expression, α is the pitch angle between the line of sight and the magnetic field in the fluid frame, $F(x) = x \int_x^\infty K_{5/3}(y) dy$ is the synchrotron function, and ν_c is the characteristic synchrotron frequency,

$$\nu_c = \frac{3\, eB\, \gamma^2 \sin\alpha}{4\pi m_e c}. \tag{4.40}$$

To speed up the computations, the modified `grtrans` code uses fitting functions for the synchrotron function $F(x)$ and Bessel function $K_{5/3}(x)$ from Fouka & Ouichaoui (2013).

The green curve in the left panel in Figure 4.9 shows the median `grtrans` synchrotron SED from the thermal control run, computed from the snapshot data from 15,000-20,000 t_g, as observed at an angle of 60° with respect to the disk polar axis. The shaded region represents the 68% confidence interval (nominal 1σ range) for the time-variability of the spectrum in this time interval. The spectrum peaks at $\nu \sim 10^{12}$ Hz, with a steep fall-off at lower frequencies because of self-absorption and a fall-off at higher frequencies because of the rapid decline in the number of thermal electrons at large Lorentz factors.

The indigo curve in the same panel was computed using `HEROIC` to self-consistently solve for the spectrum and angular distribution of radiation at each position using the radiative transfer equation. The `HEROIC` radiative transfer includes all radiation processes — synchrotron, bremsstrahlung, and inverse Compton scattering. In the synchrotron component, the `HEROIC` spectrum agrees very well with the `grtrans` spectrum except at frequencies below 10^{10} Hz. This small discrepancy arises because the `HEROIC` computations were done using simulation data out to a radius of $300\, r_g$, whereas the `grtrans` calculations were limited to $50\, r_g$.

The right panel in Figure 4.9 shows spectra of the nonthermal run computed with `grtrans` using the same parameters as the left panel. Similarly to the left panel, the solid lines are the median SEDs from the interval 15,000-20,000 t_g, and the shaded regions are the 68% confidence range of the time

variability. HEROIC does not presently include nonthermal electrons in its calculations. Comparing the thermal-only (green curve) and the nonthermal-only (blue curve) grtrans spectra, thermal emission dominates by far in the submillimeter band, nonthermal emission is modestly stronger at infrared wavelengths. Nonthermal particles make the only contribution to the synchrotron emission at X-ray wavelengths. The power-law synchrotron emission is optically thin, and shows a characteristic slope $L_\nu \propto \nu^{1/3}$ at low frequencies. The power-law tail in the spectrum at high frequencies has a spectral slope $L_\nu \propto \nu^{-p/2} \propto \nu^{-1.75}$, as expected for a cooled population of electrons with a distribution mostly broken to a power-law slope $-(p+1) = -4.5$.

The red curve in Figure 4.9 shows the combined synchrotron emission from both thermal and nonthermal electrons. By and large, the combined spectrum is a direct sum of the two independent contributions, except at the lowest frequencies, where absorption by thermal electrons suppresses the nonthermal emission (Özel et al., 2000; Yuan et al., 2003). This effect is seen also in other recent studies in which synchrotron spectra from thermal and nonthermal electrons are computed by post-processing single temperature GRMHD simulations (Ball et al., 2016; Mao et al., 2016). Note that the spectra shown here include only thermal and nonthermal synchrotron emission. For more realistic nonthermal spectra, it will be necessary to incorporate synchrotron, bremsstrahlung, and inverse Compton scattering from nonthermal electrons into a global radiative transfer solver like HEROIC or a Monte Carlo transfer code such as grmonty (Dolence et al., 2009).

Sgr A* is known to be more variable in the infrared compared to millimeter/submillimeter, and it is even more variable in X-rays (Eckart et al., 2006; Yusef-Zadeh et al., 2006; Dodds-Eden et al., 2009; Neilsen et al., 2013). From the variability in the spectra shown in the right panel of Figure 4.9, it is clear that the uniform injection prescription used in this test generates little variability in the nonthermal synchrotron emission at high frequencies. However, the present simulations are not suitable for exploring the variability in detail, both because they are in 2D and because they use a

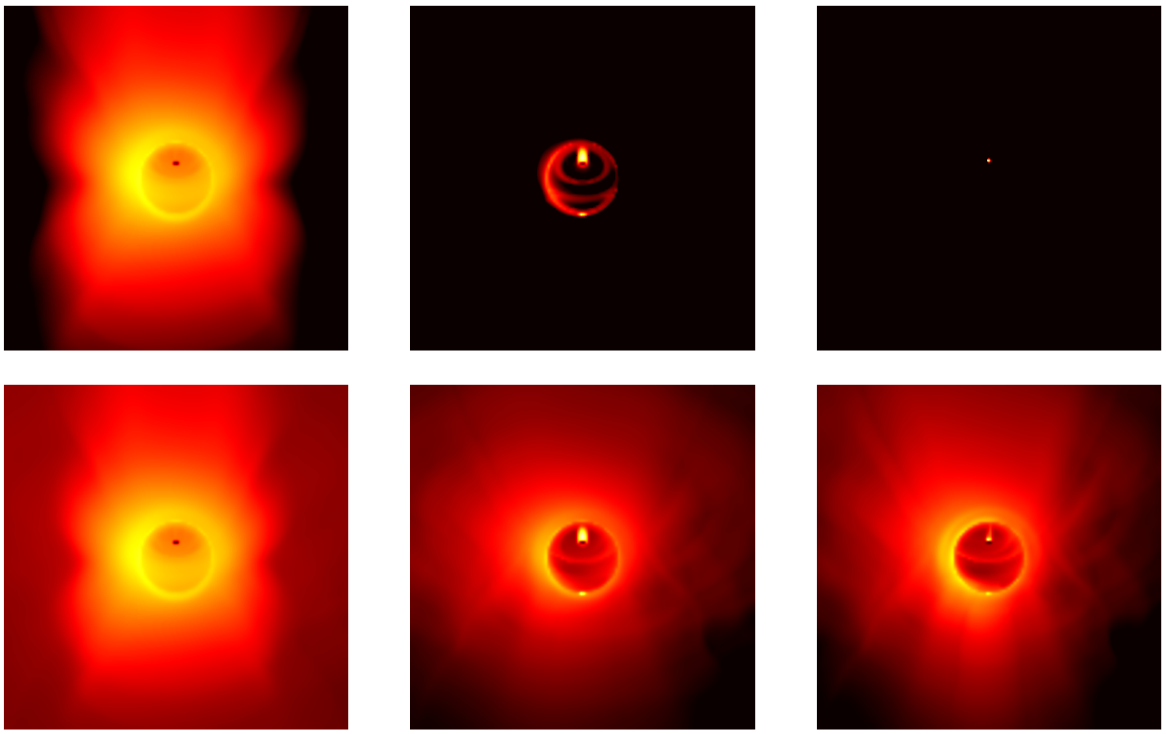


Figure 4.10: Images ($50\,r_{\rm g}$ wide, using logarithmic color maps) of synchrotron emission only, computed using `grtrans`, for the time-averaged control run (top row) and the nonthermal run (bottom row). The images correspond to 230 GHz millimeter wavelength emission (left), 136 THz near-infrared emission (middle), and 2 keV X-ray emission (right).

toy prescription for nonthermal energy injection. The thermal spectrum in Figure 4.9 shows that variability in the thermal X-ray inverse Compton spectrum exceeds that in the direct synchrotron emission at lower frequencies. Furthermore, a direct comparison of the thermal and nonthermal frequency-integrated inverse Compton power shows that while the thermal IC power dominates in the disk in the densest regions at small radii, the high energy of the nonthermal electrons (and the fact that the IC power grows as γ^2) leads to the nonthermal IC power exceeding the thermal IC power in the funnel region and in the disk at radii $\gtrsim 40\,r_{\rm g}$. Thus, nonthermal electrons should make a significant contribution to the high frequency spectrum and variability from IC emission. To accurately explore variability and flares from nonthermal electrons, these simulations must be expanded to 3D with local injection prescriptions, and nonthermal bremsstrahlung and inverse Compton emission should be included in the radiative transfer.

Figure 4.10 shows `grtrans`-generated ray-traced images of the synchrotron emission from the time-averaged simulations at 3 frequencies: 230 GHz, which is near the thermal synchrotron peak and corresponds to the observing frequency of the Event Horizon Telescope (Doeleman et al., 2008), 136 THz in the near-infrared, and 4.8×10^{17} Hz (2 keV) in X-rays. The images are 50 projected gravitational radii across and displayed in a log scale. The bright regions of the image at 230 GHz are practically the same for the thermal and nonthermal runs, confirming that much of the emission is from thermal electrons. There is, however, additional extended flux at large radii in the bottom panel from emission by nonthermal electrons. The lensed photon ring in the infrared image is brighter when nonthermal electrons are included, and the emission extends to noticeably larger radii. The X-ray synchrotron image is almost entirely from nonthermal emission. As for the spectra in Figure 4.9, these results depend sensitively on the simple nonthermal energy injection prescription used.

4.5 Summary and conclusions

This chapter introduces a new algorithm to self-consistently evolve a population of nonthermal electrons with an arbitrary distribution function in a black hole spacetime, in parallel with magnetized thermal gas and radiation. In each time step, a fraction of the viscously generated heat is used to heat some of the thermal electrons and to transfer them to the nonthermal population. The nonthermal electrons move with the fluid, and their energy distribution is modified by gas compression and expansion, Coulomb coupling, and radiative cooling. The back-reaction of the nonthermal electrons on the thermal population is automatically included.

The algorithm performs well on a variety of test problems, including the first test with a 2D black hole accretion flow. This simulation has a low mass accretion rate, roughly equal to the rate

estimated in Sgr A*. As a result, the nonthermal distribution does not significantly affect the gas dynamics or thermodynamics of the thermal electrons or ions. However, the radiation power is enhanced, since the nonthermal electrons radiate more efficiently than their thermal counterparts. Furthermore, the energy distribution of the nonthermal electrons varies with location in the accretion flow. The distribution exhibits a synchrotron cooling break, and the break Lorentz factor $\gamma_{\rm brk}$ varies with position, set by local conditions such as the magnetic field strength (which determines synchrotron power) and the gas velocity (which sets the effective advection time). Furthermore, $\gamma_{\rm brk}$ is also modified by other factors such as strong adiabatic compression.

The accretion simulation in Section 4.4 considers only one particularly simple prescription for injection into the nonthermal population, and the resulting simulation results are strongly influenced by this choice. A constant injection range of γ, independent of radius, ensures that nonthermal synchrotron emission dominates over thermal emission at large radii, where the temperature of the thermal electrons falls off rapidly. This behavior is reflected in Figure 4.10, which shows that at high frequencies, nonthermal electrons from farther out in the disk dominate the raytraced synchrotron image of the accreting gas. Furthermore, the choice of a minimum injection Lorentz factor $\gamma_{\rm inj\,min} = 500$ means that most of the nonthermal emission is concentrated at infrared or higher frequencies, while the image at 230 GHz is basically unchanged compared to a purely thermal model. In principle, $\gamma_{\rm inj\,min}$ should be chosen such that the nonthermal population connects smoothly to the thermal distribution, without a gap between the two.

Another consequence of the choice of injection parameters is that the high frequency nonthermal emission in the simulation shows relatively little time variability. This stability arises because nonthermal electrons are distributed smoothly and uniformly throughout the simulation with $\delta_{\rm nth} = 0.015$ everywhere. In contrast, the rapid variability that is observed in Sgr A*is likely driven by strong localized injection, perhaps from shocks or magnetic reconnection. This suggests a much

more sporadic and localized injection of nonthermal energy, with small regions where the fraction of energy going into the nonthermal electrons, $\delta_{\rm nth}$, is much larger than the 1.5% used in this work, and large regions elsewhere with $\delta_{\rm nth}$ near-zero (see Ball et al. 2016). Furthermore, particle-in-cell simulations show that electrons accelerated in reconnection events attain progressively harder energy spectra as the magnetization of the plasma increases (Sironi & Spitkovsky, 2014; Ball et al., 2018). This effect will again have a strong impact on variability.

The next, natural application of the method presented in this chapter is thus to perform full 3D simulations with nonthermal particle evolution sourced by localized injection in order to track the origins and spatial and spectral evolution of flares from Sgr A*. Ball et al. (2018) recently published simple prescriptions for the power law index p and acceleration fraction $\delta_{\rm nth}$ as a function of local plasma parameters in simulations of magnetic reconnection. Their results predict that the power law index p is a steep function of the magnetization $\sigma_{\rm i}$, with values $p \approx 2$ in reconnecting plasmas at $\sigma_{\rm i} \approx 1$.[5] This behavior implies that, under this prescription, flares should be sourced in high $\sigma_{\rm i}$ regions close to the black hole in Sgr A*, as nonthermal distributions from less magnetized regions fall off quickly with energy at larger values of p. The recent detection of horizon-scale circular motion in near-infrared flares by the GRAVITY interferometer (GRAVITY Collaboration et al., 2018a) suggests that flares may arise in "hot spots" of high-magnetization plasma near the black hole. New results from GRAVITY, the EHT, and other instruments in characterizing the amplitudes, lifetimes, polarization, and orbits of Sgr A* flares in the submm, NIR, and X-ray will provide a rich dataset for testing the particle acceleration mechanisms explored in full 3D simulations performed using the method introduced in this chapter.

[5]This result is consistent with the measured near-infrared spectral index from Ponti et al. 2017 if the power law is broken by cooling and flares originate from synchrotron emission.

Part II

Imaging

Page intentionally left blank

Text in this chapter was previously published in *ApJ* 829 (2016), 1, 11 (A. Chael, M. Johnson, R. Narayan, S. Doeleman, J. Wardle, and K. Bouman) and in *ApJ* 857 (2018), 1, 23 (A. Chael, M. Johnson, K. Bouman, L. Blackburn, K. Akiyama, and R. Narayan).

5

Interferometric imaging with Regularized Maximum Likelihood

As described in the Introduction, the Event Horizon Telescope (EHT) is a global Very Long Baseline Interferometry (VLBI) array with eight participating telescopes at six distinct geographical sites (Paper II). With an operating wavelength of $\lambda = 1.3$ mm and with its longest baselines spanning nearly the Earth's diameter $b_{\rm max} \sim 10^4$ km, the EHT's nominal resolution, or its observing wavelength divided by the longest baseline, is $b_{\rm max}/\lambda \approx 25\,\mu$as. Together, the EHT's long baselines and short operating wavelength provide the extremely fine resolution necessary to resolve and make images of the lensed photon rings around the supermassive black holes in M87 ($d_{\rm shadow} \approx 40\,\mu$as) and Sgr A* ($d_{\rm shadow} \approx 48\,\mu$as).

As an interferometer, the EHT does not measure the sky intensity distribution directly; rather, it measures an incomplete set of complex "visibilities" which sample the Fourier components of the underlying image. Reconstructing images from the measured interferometric data must be done computationally in a process called "synthesis imaging." However, because of the incomplete Fourier sampling of the sparse EHT array, an infinite number of images can fit the measured data. Synthesis imaging is an ill-posed problem.

Standard inverse model approaches to interferometric imaging, such as the CLEAN algorithm (Högbom, 1974; Clark, 1980), begin with an inverse Fourier transform of the sampled visibilities (the "dirty image") and then proceed to deconvolve artifacts introduced by the sparse Fourier sampling (the Fourier transform of the baseline coverage pattern, or "dirty beam"). In particular, CLEAN performs this deconvolution by decomposing the image into point sources. To use traditional deconvolution imaging algorithms like CLEAN, the interferometric visibilities must be calibrated for amplitude and phase errors. At high frequencies like the 230 GHz operating frequency of the EHT, however, the atmospheric coherence time can be as short as seconds, and rapid phase variations effectively eliminate the absolute interferometric phase. Absolute amplitude calibration also becomes more difficult at high frequencies, where pointing errors can introduce large, time-varying errors in station gain terms.

Most VLBI calibration errors can be decoupled into station-based gain terms (Thompson et al., 2017, hereafter TMS). An interferometric array consisting of N_s stations samples $N_s(N_s-1)/2$ visibilities at each time, but has only N_s unknown complex gains. Hence, the calibration is over-constrained and combinations of the measured visibilities–closure quantities–can be formed that are unaffected by calibration errors. For example, a closure phase is formed by adding together three visibility phases around a triangle, canceling out the station-based phase errors on each individual visibility (Jennison, 1958; Rogers et al., 1974). Likewise, the closure amplitude is a combination of four visibility amplitudes that cancels out amplitude gain errors in a specified ratio (Twiss et al., 1960). Despite the challenges in *absolute* calibration of a VLBI array, closure quantities provide robust measurements of certain *relative* quantities, which carry information about source structure that is only limited by the level of thermal noise.

When the data's calibration is uncertain, the usual approach in VLBI imaging is to iterate between imaging with CLEAN and deriving new calibration solutions using information from the last-

solved-for image — a so-called "self-calibration" or "hybrid mapping" loop (e.g., Wilkinson et al., 1977; Readhead & Wilkinson, 1978; Readhead et al., 1980; Schwab, 1980; Cornwell & Wilkinson, 1981; Pearson & Readhead, 1984; Cornwell & Fomalont, 1999; Thompson et al., 2017). The results and time to convergence of this approach depend on many assumptions made in the course of the hybrid process; these choices include the initial source model used for self-calibration, which regions to clean in a given iteration, the method used for deriving complex gains from a given image, and how frequently to re-calibrate the data. The procedure used for each data set is not standard and is typically driven by expert judgment; the sensitivity of the final image to these assumptions cannot be directly inferred from the result.

In contrast with CLEAN's approach of deconvolving the dirty image into point sources, imaging algorithms in the family of Regularized Maximum Likelihood (RML) methods directly solve for the source image pixels by finding the best-fit image to data, constrained by additional convex "regularization" terms. The most familiar of these techniques is the Maximum Entropy Method (MEM, see e.g Narayan & Nityananda 1986), but other regularizing functions such as image sparsity, or smoothness can also be used. In contrast with CLEAN, RML methods only rely on comparing the data computed from a trial image to the specified measurements. In other words, RML methods never need to perform an inverse Fourier transform from calibrated data. As a result, they can be used directly with robust data products derived from complex visibilities. The field of optical interferometry has pioneered the use of imaging directly from the measured visibility amplitudes and closure phases, bypassing the corrupted visibility phase (Buscher, 1994; Baron et al., 2010; Thiébaut, 2013; Thiébaut & Young, 2017). Recently, several other methods have built on these techniques in preparing imaging algorithms for EHT data, fitting some combination of closure phases and visibility amplitudes directly while using different regularizing functions (e.g., Bouman et al., 2016; Akiyama et al., 2017b). Extending these approaches further, RML methods provide a

natural way to use both closure phases and amplitudes together as the fundamental data product, bypassing the self-calibration loop entirely.

This chapter presents a framework for RML imaging in interferometry, focusing on a method (first described in Chael et al., 2018b) which is the first to reconstruct images directly using only closure amplitudes and closure phases. These reconstructions require no assumptions about absolute phase or amplitude calibration beyond stability during the integration time used to obtain the visibilities. The extreme case of complete *closure-only imaging* is particularly attractive for imaging with the EHT, but in general it is the overall flexibility of RML methods that makes them excellent choices for interferometric imaging. Chapter 6 builds off the general framework established here to discuss the RML imaging software implemented in the `eht-imaging` Python library, which is a workhorse of the EHT's data processing and imaging pipelines.

5.1 Visibilities and closure quantities

5.1.1 Interferometric visibilities

Two telescopes i and j in an interferometer separated by a baseline vector $\vec{b}_{ij}$ measure a complex visibility V_{ij} by cross-correlating the electric field recorded at each station.[1] The van Cittert-Zernike theorem (van Cittert, 1934; Zernike, 1938) identifies the *ideal* visibility $\mathcal{V}_{ij}$ measured by these two stations with a Fourier component of the source brightness distribution $I(x, y)$ on the sky:

$$V_{ij} = \tilde{I}(u, v) = \int\int I(x, y) \mathrm{e}^{-2\pi \mathrm{i}(ux+vy)} \, \mathrm{d}x \, \mathrm{d}y. \tag{5.1}$$

[1] Interferometric visibilities are in general polarimetric quantities; at a given time and frequency, two stations can measure visibilities corresponding to the four Stokes parameters I, Q, U, V by cross-correlating different combinations of the two polarizations recorded at each station. This chapter focuses on total intensity imaging; see Roberts et al. (1994) for a discussion of polarimetric quantities in VLBI and Chael et al. (2016) for the RML polarimetric imaging method implemented in the `eht-imaging` library.

Here, (x, y) are real space angular coordinates and (u, v) are the coordinates of the baseline vector $\vec{b}_{ij}$, projected in the plane perpendicular to the source line of sight $\vec{s}$ and measured in wavelengths λ.[2]

Since the sky intensity distribution $I(x, y)$ is real valued, the visibility is conjugate-symmetric in the Fourier plane, $\mathcal{V}(-u, -v) = \mathcal{V}^*(u, v)$. When N_s stations can observe the source, the number of independent instantaneous visibilities N_V is given by the binomial coefficient

$$N_V = \frac{N_s(N_s - 1)}{2}. \tag{5.2}$$

To fill in samples of the Fourier plane from the small number N_V available at a single instant in time, interferometric observations use "earth rotation aperture synthesis." As the Earth rotates, the projected baseline coordinates (u, v) trace out elliptical curves in the Fourier domain, providing measurements of new visibilities from the same pair of telescopes.

The ideal identification of measured visibilities with Fourier components of the image is complicated by several factors. First, thermal noise from the telescope receiver chains, Earth's atmosphere, and astronomical background corrupts the measured visibility. This thermal noise, ϵ_{ij}, is assumed to be drawn from a complex Gaussian distribution with a time- and baseline-dependent standard deviation σ_{ij}. The noise level depends on the telescope sensitivities, bandwidth, and integration time. Second, each station transforms the measured incoming polarized waveform according to its own (time-dependent) 2×2 Jones matrix $\mathbf{J}_i$ that adjusts the level of the measured signal amplitude and mixes the measured polarizations (e.g.,Hamaker et al. 1996; TMS). For the total intensity imaging considered in this chapter, each station is treated as contributing a single (time-dependent)

[2]Typically, u measures spatial frequencies projected along the east-west axis in the plane of the sky (positive in the East), and v measures frequencies projected along the north-south axis (positive in the North). The real space angular coordinates (x, y) are measured with the same conventions. The w component of the baseline vector projected along the line of sight is unimportant unless the interferometer field of view is very large (TMS).

complex gain $G_i\, e^{i\phi_i}$ to the visibility. Appendix C discusses the full method for simulating realistic data from polarimetric images in `eht-imaging`, including polarimetric leakage.

The station-based phase error ϕ_i results from uncorrected propagation delays and clock errors. In particular, atmospheric turbulence contributes a rapidly varying stochastic term to each ϕ_i, which for EHT observations at 1.3 mm has a coherence time on the order of seconds. The amplitude gain term G_i arises from uncertainty in the conversion of the correlation coefficients measured on each baseline to units of flux density. In general, G_i is more slowly varying than ϕ_i.

Including all of these corrupting factors, the full complex visibility is

$$V_{ij}(u,v) = G_i G_j e^{i(\phi_i - \phi_j)} \left(\mathcal{V}_{ij}(u,v) + \epsilon_{ij}\right), \tag{5.3}$$

where the measured visibility, gain amplitudes, phases, and thermal noise all vary in time.

Note that Equation 5.3 represents all systematic errors (e.g., those other than thermal noise) as station-based effects. In practice, effects such as polarization leakage and bandpass errors will also contribute small baseline-based effects that can bias closure quantities. However, these errors are generally much more slowly varying than ϕ_i or G_i and can often be removed with a priori calibration of the complex visibilities.

5.1.2 Closure phases and closure amplitudes

Two types of "closure quantities" can be formed from interferometric visibilities that are insensitive to station-based complex gain terms. While these quantities are robust to the presence of arbitrarily large complex gains on the visibilities, they contain less information about the source than the full set of complex visibilities. Furthermore, because closure quantities mix different Fourier components, they can be difficult to interpret physically.

First, multiplying three complex visibilities around a triangle of baselines eliminates the complex gain phase terms (Jennison 1958; Rogers et al. 1974; TMS). For three stations on a triangle i–j–k, the visibility bispectrum is

$$V_{\mathrm{B},\,ijk} \equiv |V_{\mathrm{B},\,ijk}| e^{i\psi_{ijk}} = V_{ij}\, V_{jk}\, V_{ki}. \tag{5.4}$$

While the bispectral amplitude $|V_{\mathrm{B}}|$ is affected by the amplitude gain terms G_i in Equation 5.3, the phase of the bispectrum is preserved under any choice of station-based phase error. This closure phase ψ_i is thus a robust interferometric observable; apart from thermal noise, the measured closure phase is the same as the closure phase of the observed image.

The total number of closure phases at a moment in time is equal to the number of triangles that can be formed from stations in the array, $\binom{N_s}{3}$. However, not all of these closure phases are independent, as some can be formed by adding or subtracting other closure phases in the set. The total set of independent closure phases can be obtained by selecting an antenna as a reference and choosing only the triangles that include that antenna (Twiss et al. 1960; TMS). The total number of such independent closure phases is

$$N_\psi = \binom{N_s - 1}{2} = \frac{(N_s - 1)(N_s - 2)}{2}. \tag{5.5}$$

N_ψ is less than the number of measured visibilities at a given time (Equation 5.2) by the fraction $1 - 2/N_s$.

Second, on any set of four stations, *closure amplitudes* are formed by taking ratios of visibility amplitudes so as to cancel all the amplitude gain terms G_i in Equation 5.3. Up to inverses, the baselines among any set of four simultaneously observing stations $\{ijk\ell\}$ can form three quadrangles

with three corresponding closure amplitudes A_{C}:

$$A_{\mathrm{C},\,ijk\ell} = \frac{A_{ij}A_{k\ell}}{A_{ik}A_{j\ell}} \quad , \quad A_{\mathrm{C},\,ikj\ell} = \frac{A_{ik}A_{j\ell}}{A_{ij}A_{k\ell}},$$
$$A_{\mathrm{C},\,i\ell jk} = \frac{A_{i\ell}A_{jk}}{A_{ij}A_{\ell k}}, \tag{5.6}$$

where the A terms are the (debiased) visibility amplitudes (Equation 5.9). Since the product of the three closure amplitudes in Equation 5.6 is unity, only two closure amplitudes in the set are independent. The total number of instantaneous closure quadrangles is $3\binom{N_s}{4}$, but the number of independent closure amplitudes is

$$N_{\mathrm{C}} = \frac{N_s(N_s-3)}{2}. \tag{5.7}$$

N_{C} is equal to the total number of visibilities minus the number of unknown station gains (TMS). At any given time, the number of closure amplitudes is less than the number of visibilities by a fraction $1 - 2/(N_s - 1)$. Like the visibility amplitude and bispectrum, closure amplitudes are biased by thermal noise, and their distribution becomes highly non-Gaussian even at moderate SNR (Blackburn et al., 2019). For this reason, the logarithm of the closure amplitudes $\ln A_{\mathrm{C}}$ is often used in the RML imaging methods presented later in this chapter (see Section 5.2.2).

The robustness of closure phases and amplitudes to calibration errors comes at a cost of the loss of some information about the source. For instance, closure phases are insensitive to the absolute position of the image centroid, and closure amplitudes are insensitive to the total flux density. These can be constrained separately, either through arbitrary choices (e.g., centering the reconstructed image) or through additional data constraints (e.g., specifying the total image flux density through a separate measurement).

5.1.3 Redundant and trivial closure quantities

Some VLBI arrays include multiple stations that are geographically co-located. For instance, the EHT includes two stations on Maunakea in Hawai'i (the SMA and the JCMT) as well as two stations in the Atacama desert in Chile (the ALMA array and the APEX telescope). Practically, any two stations that form a baseline that does not appreciably resolve any source structure can be considered co-located.

These "redundant" stations can still be used to form closure quantities. In the case of closure phase, the added triangles provide no new source information. Specifically, any triangle $\{\vec{b}_{12}, \vec{b}_{23}, \vec{b}_{31}\}$ that includes two co-located stations $\{1, 2\}$ will include one leg that measures the zero-baseline visibility; $V_{12} = \tilde{I}(0, 0)$. The zero-baseline visibility is the integrated flux density of the source; it has zero phase (see Equation 5.1). The remaining two long legs from the pair of co-located stations to the third station will have $\vec{b}_{23} = -\vec{b}_{31}$, and consequently $V_{23} = V_{31}^*$. Thus, the bispectrum will be a positive real number, and the closure phase must be zero regardless of the source structure. These *trivial* triangles are not useful for imaging, but they provide valuable tests of the closure phase statistics and systematic bias (e.g., Fish et al., 2016).

Redundant stations also give rise to trivial closure amplitudes which have a value of unity regardless of the source. For instance, if in a set of four stations $\{1, 2, 3, 4\}$ the stations $\{1, 2\}$ are co-located, the numerator and denominator in the closure amplitude $A_{13}A_{24}/A_{14}A_{23}$ will always be equal, regardless of the underlying source structure. However, redundant stations also yield non-trivial closure amplitudes that provide additional information on the source structure. As an example, one can measure the normalized visibility amplitude, $\left|V(\vec{u})/V(\vec{0})\right|$, as a closure quantity on any baseline joining two sets of co-located stations (Johnson et al., 2015). In the limiting case where every station in an array has a redundant companion, the complete source visibility am-

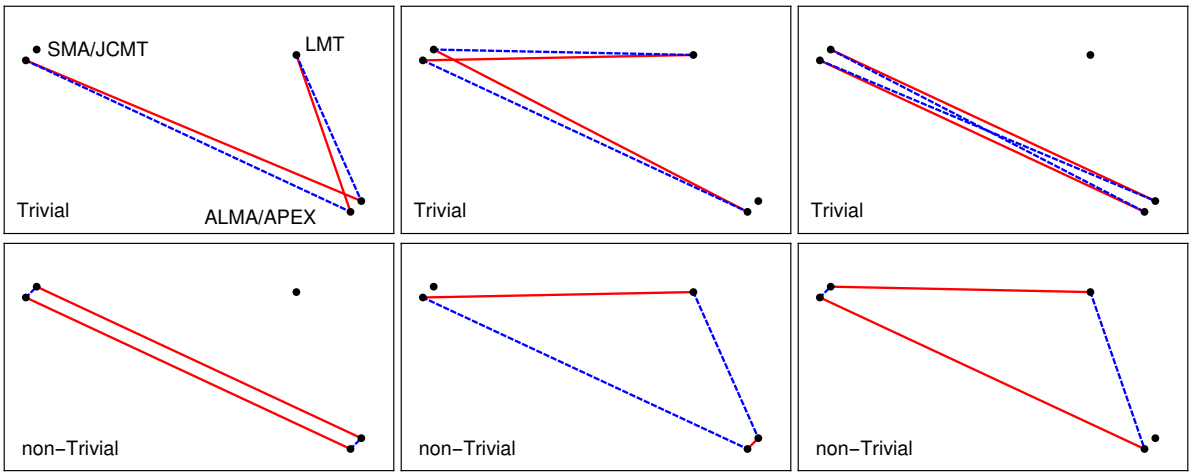


Figure 5.1: Example closure amplitudes for a portion of the EHT. Solid red lines connecting stations denote visibilities in the numerator of the closure amplitude; dashed blue lines denote visibilities in the denominator. An array containing redundant stations (such as SMA/JCMT and ALMA/APEX in the EHT) will produce trivial closure amplitudes, which are equal to unity (plus thermal noise), as well as non-trivial closure amplitudes, which yield new information about the source. Without redundant stations, there would be no closure amplitudes from this portion of the array.

plitude information could be recovered through closure amplitudes, except for a single degree of freedom for the total flux density.

Figure 5.1 shows examples of the trivial and non-trivial closure amplitudes for an array with partial redundancy. As these examples illustrate, redundant stations can significantly inform and improve calibration and imaging. Figure 5.2 shows the number of closure amplitudes and phases for the EHT with and without redundant stations as a function of observing hour. The two redundant stations of the 2017 EHT array more than double the amount of information contained in the set of closure amplitudes over the same array considered without these stations.

5.1.4 Thermal noise on closure quantities

The thermal noise ϵ_{ij} on the baseline i–j in Equation 5.3 is a circularly-symmetric complex Gaussian random variable with zero mean that is independently sampled for each visibility measurement. The standard deviation σ_{ij} of the thermal noise on this baseline is determined by the radiometer

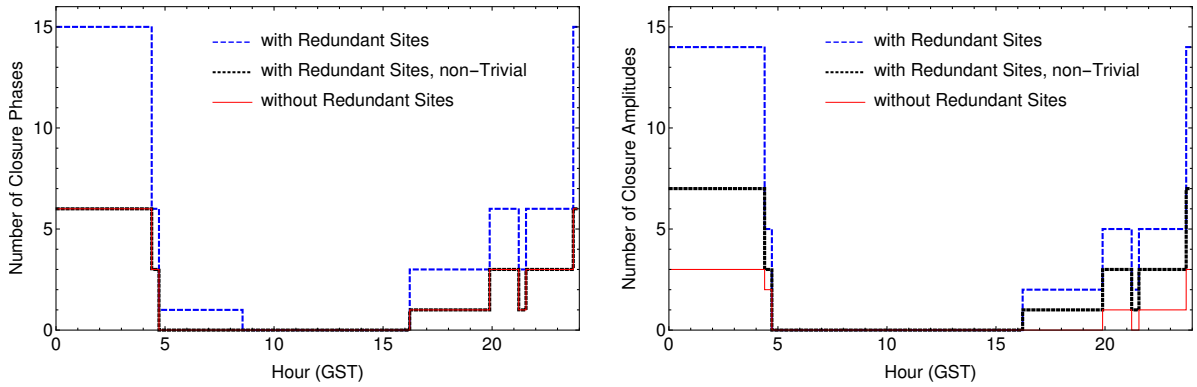


Figure 5.2: (Left) Number of independent closure phases for the 2017 EHT over 24 hours GMST while observing Sgr A*. The blue line shows the total number of independent closure phases in the array containing redundant stations, the black line shows the number of independent closure phases that measure source structure, and the red line shows the number of independent closure phases in the array when the redundant stations are excluded. Redundant stations do not add any closure phase information to the array. (Right) Total independent (blue) and non-trivial (black) closure amplitudes over 24 hours for the EHT including redundant stations. Unlike for closure phases, adding only two redundant stations significantly increases the amount of information contained in the set of independent closure amplitudes compared to the same array without these stations (red) because not all closure amplitudes containing a baseline between two co-located stations are trivial (see Figure 5.1).

equation (TMS):

$$\sigma_{ij} = \frac{1}{\eta}\sqrt{\frac{\mathrm{SEFD}_i \times \mathrm{SEFD}_j}{2\Delta\nu\,\Delta t}}. \tag{5.8}$$

In Equation 5.8, SEFD_i and SEFD_j are the "system equivalent flux densities" of the two telescopes.[3] The observing bandwidth of the visibility measurement is $\Delta\nu$, and Δt is the integration time. The factor of $1/\eta$ in Equation 5.8 is due to quantization losses in the signal digitization; for the 2-bit quantization used by the EHT, $\eta = 0.88$ (TMS).

When the signal-to-noise ratio (SNR) is high, the visibility amplitudes will also be Gaussian distributed with standard deviation σ_{ij} given by Equation 5.8. At lower SNR > 1, the distribution of the amplitudes becomes non-Gaussian, and the estimate of the visibility amplitude taken directly from the norm of the complex visibility is biased upward by the noise. To first order, the amplitudes can be debiased with the equation (TMS)

$$A_{ij} = \sqrt{\left(|V_{ij}|^2 - \sigma_{ij}^2\right)\Theta\left(|V_{ij}|^2 - \sigma_{ij}^2\right)}. \tag{5.9}$$

[3]For a telescope with system temperature T_{sys} and effective area A_{eff}, the SEFD is $2k_{\mathrm{B}}T_{\mathrm{sys}}/A_{\mathrm{eff}}$.

The Heaviside Θ-function in Equation 5.9 ensures that the debiased amplitudes are always real.

In the high signal-to-noise limit, the baseline-based thermal noise on the closure amplitudes and phases introduced in Section 5.1 will also be Gaussian distributed. To first order, the standard deviation $\sigma_{\rm B}$ of the complex noise on the bispectrum $V_{\rm B}$ due to the thermal noise on the 3 component visibilities is (Rogers et al., 1995)

$$\sigma_{{\rm B},\,ijk} = |V_{{\rm B},\,ijk}| \sqrt{\frac{\sigma_{ij}^2}{|V_{ij}|^2} + \frac{\sigma_{jk}^2}{|V_{jk}|^2} + \frac{\sigma_{ki}^2}{|V_{ki}|^2}}. \tag{5.10}$$

In the high SNR regime, the variance of the phase of a complex quantity X drawn from a circular Gaussian distribution is $\sigma^2_{{\rm Arg}[x]} = \sigma^2_X/|X|^2$. Thus, the closure phase uncertainty σ_ψ depends only on the SNR values on the three baselines:

$$\sigma_{\psi\,,ijk} = \frac{\sigma_{{\rm B}\,,ijk}}{|V_{{\rm B}\,,ijk}|} = \sqrt{\frac{\sigma_{ij}^2}{|V_{ij}|^2} + \frac{\sigma_{jk}^2}{|V_{jk}|^2} + \frac{\sigma_{ki}^2}{|V_{ki}|^2}}. \tag{5.11}$$

The standard deviation $\sigma_{\rm C}$ of the thermal noise of a closure amplitude $A_{\rm C}$ is, to leading order in the inverse SNRs,

$$\sigma_{{\rm C}\,,ijk\ell} = A_{{\rm C},\,ijk\ell} \sqrt{\frac{\sigma_{ij}^2}{|V_{ij}|^2} + \frac{\sigma_{k\ell}^2}{|V_{k\ell}|^2} + \frac{\sigma_{ik}^2}{|V_{ik}|^2} + \frac{\sigma_{j\ell}^2}{|V_{j\ell}|^2}}. \tag{5.12}$$

To first order the variance on the logarithm of a quantity x is $\sigma^2_{\log(x)} = \sigma_x^2/x^2$, so like the closure phase, the noise on the log closure amplitude is only determined by the component SNRs:

$$\sigma_{\log{\rm C}\,,ijk\ell} = \sqrt{\frac{\sigma_{ij}^2}{|V_{ij}|^2} + \frac{\sigma_{k\ell}^2}{|V_{k\ell}|^2} + \frac{\sigma_{ik}^2}{|V_{ik}|^2} + \frac{\sigma_{j\ell}^2}{|V_{j\ell}|^2}}. \tag{5.13}$$

At moderately low SNR, the Gaussianity of the thermal noise on phase and amplitude breaks down, as does the appropriateness of using the measured SNR as an estimate of the true SNR when estimating σ_ψ and $\sigma_{\rm C}$. Because the measured phase is unbiased by thermal noise and wraps at 2π,

the true σ_ψ is *smaller* than the estimate in Equation 5.11 in the low-SNR limit. Low-SNR closure phases are not prone to extreme outliers, so the Gaussian approximation is reasonable to use over a broad range of signal-to-noise.

In contrast, the distribution for the reciprocal visibility amplitude, which appears in the denominator of Equation 5.12 for σ_C, takes on an extreme tail at low SNR that extends to positive infinity. This tail causes a large positive bias in the measured closure amplitudes, and it gives the closure amplitude a severely non-Gaussian distribution. Fitting to log closure amplitudes instead of the closure amplitudes themselves has the dual benefits of mitigating the tail of the reciprocal amplitude distribution and symmetrizing the numerator and denominator. Furthermore, debiasing the component visibility amplitudes with Equation 5.9 corrects the estimate of the log closure amplitude to first order. Detailed analysis of the statistics of closure quantities will be explored in a forthcoming work (Blackburn et al., 2019).

5.2 RML imaging

5.2.1 Imaging framework

The standard methods of interferometric imaging are based on the CLEAN algorithm (Högbom, 1974; Clark, 1980). CLEAN operates on the so-called "dirty image" obtained by directly taking the Fourier transform of the sparsely sampled visibilities. To produce an image of the source, CLEAN attempts to deconvolve the "dirty beam", or point-spread function that results from incomplete sampling of the Fourier domain. To perform the initial inverse Fourier transform, CLEAN requires well-calibrated complex visibilities. When a priori calibration is ineffective – which is often the case at EHT frequencies where atmospheric phase terms vary rapidly – the visibilities must be

"self-calibrated".[4]

The CLEAN self-calibration procedure starts from an initial model image and solves for the set of time-dependent complex gains in Equation 5.3 either by fixing a sufficient set of amplitudes or phases directly from the image and solving for the rest analytically (Wilkinson et al., 1977; Readhead et al., 1980), or by finding a set that minimizes the sum of squares of the differences between the measured and model visibilities (Schwab, 1980; Cornwell & Wilkinson, 1981).[5] Self-calibration is often performed in practice by first solving only for the phases of the complex gains, correcting the amplitudes at a later stage (Cornwell & Fomalont, 1999). At each round of self-calibration, the estimated inverse gain terms are applied to the measured visibilities, and the imager is run again to obtain a new source model. These two steps are repeated many times until convergence. There are several assumptions in this procedure which may affect the final image. Most critical are the choice of the initial source model (often taken as a point source) and the choice of where to clean the image in each iteration (the so-called "clean boxes"). These choices enforce assumptions about the source brightness distribution early on in the self-calibration process which then propagate to later rounds via the self-calibrated complex visibilities.

In contrast, the various methods of interferometric imaging explored in this paper all fall under the category of regularized maximum likelihood algorithms. RML methods search for an image that maximizes the sum of a "data term" that enforces image-data consistency and a "regularizer" function that prefers images with certain features when the data are not sufficient to constrain the structure. RML methods can often be interpreted in a Bayesian framework, where the data term is identified with a log-likelihood and the regularizer term with a log-prior; however, many regularizing functions do not have straightforward probabilistic interpretations. RML methods

[4]Although self-calibration is used most frequently with CLEAN, it can be used in conjunction with any imaging method that requires calibrated complex visibilities.

[5]This is the strategy adopted in `eht-imaging`'s self-calibration routines (Section 6.6.1).

require only a forward Fourier transform from trial images to the visibility domain. Consequently, they can fit directly to data terms like closure quantities that are derived from the visibilities, even if the visibilities themselves are corrupted by gain and phase errors.

In astronomy, the most familiar of these RML methods is the Maximum Entropy Method (e.g., Frieden, 1972; Gull & Daniell, 1978; Cornwell & Evans, 1985; Narayan & Nityananda, 1986). While traditional MEM algorithms use calibrated complex visibilities as their fundamental data product, other algorithms have gone beyond complex visibilities as the fundamental data product to produce images directly from the image bispectrum. Development of imaging algorithms that use different fundamental data products from complex visibilities has been particularly fruitful in optical interferometry, where the absolute visibility phase is almost never accessible (e.g., Buscher, 1994; Baron et al., 2010; Thiébaut, 2013; Thiébaut & Young, 2017), and has also been recently explored in the context of the EHT (Lu et al., 2014; Bouman et al., 2016; Akiyama et al., 2017b). RML methods have been developed beyond MEM regularization using regularizers such as the ℓ_1 norm (Honma et al., 2014), image smoothness (Kuramochi et al., 2018), or a data-driven Gaussian patch prior (Bouman et al., 2016). Regularized maximum likelihood methods have also been extended to polarization (Ponsonby, 1973; Nityananda & Narayan, 1983; Holdaway & Wardle, 1990; Chael et al., 2016; Coughlan & Gabuzda, 2016; Akiyama et al., 2017a), to the mitigation of interstellar scattering (Johnson, 2016), and to dynamical imaging to reconstruct movies of time-variable sources (Johnson et al., 2017; Bouman et al., 2018).

In the most general case, where multiple data terms and multiple regularizers may inform a reconstruction, RML finds the image $\mathbf{I}$ that minimizes an objective function $J(\mathbf{I})$:

$$J(\mathbf{I}) = \sum_{\text{data terms}} \alpha_D \chi^2_D(\mathbf{I}, \mathbf{d}) - \sum_{\text{regularizers}} \beta_R S_R(\mathbf{I}). \tag{5.14}$$

In the above expression, the χ^2_D are the data terms or chi-squared goodness-of-fit functions corresponding to the data product $\mathbf{d}$. If the data product $\mathbf{d}$ is normally distributed, these are proportional to the negative log-likelihoods that represent the log probability that the data could be observed given an underlying image $\mathbf{I}$. For data products whose distributions are not Gaussian (like closure phases and amplitudes), χ^2_D is usually an approximation to the log-likelihood. The S_R are regularizing functions which provide missing information on the image characteristics and constrain the space of possible images given the measured data. While relatively new to radio interferometry and VLBI, reconstructions using Equation 5.14 with multiple data terms and regularizers are common in optical interferometry (see e.g., Buscher, 1994; Baron et al., 2010; Thiébaut, 2013; Thiébaut & Young, 2017).

The set of "hyperparameters" α_D and β_R control the relative weighting of the different edited data and regularizer terms in the objective function (Equation 5.14). Because the location of the global minimum of $J(\mathbf{I})$ is unaffected by changes of scale, one hyperparameter can be set to unity or some other arbitrary value without changing the solution. Furthermore, interpreting the χ^2_D data terms as log-likelihoods, the data term weights α_D should ideally be determined by the number of data points of each type. For example, using the reduced χ^2 defined in Section 5.2.2, if one data term with N_1 measurements is to α_1, the remaining data terms $i > 1$ with N_i measurements should all be set as

$$\alpha_{i>1} = \alpha_i \frac{N_i}{N_1}. \tag{5.15}$$

Often, the data weights α_D are varied throughout the imaging process. In particular, heavily weighting a single data term away from the log-likelihood weighting in Equation 5.15 can aid initial convergence. The ideal weighting in Equation 5.15 can then be restored in later rounds of imaging.

In practice, the hyperparameters α_D and β_R are adjusted manually to yield reconstructions

that converge to the expected values of reduced $\chi^2 \approx 1$ (Cornwell & Evans, 1985). Recently, Akiyama et al. (2017b) instead determined hyperparameters self-consistently using cross-validation. In this method, images are reconstructed with different combinations of the hyperparameters using different data sets where a portion of the data is held in reserve. The set of hyperparameters that produces the image most compatible with the data held in reserve is then used in the final reconstruction.

5.2.2 Data terms for robust imaging

Having defined the general form of the objective function, this section details the different choices of the data χ^2 term that can be used in total intensity interferometric imaging. The simplest choice of image-data consistency metric is the reduced χ^2 of the measured visibilities. If there are N_V total measured visibilities V_j (now indexed by their position in the vector $\mathbf{V}$ of all measurements, instead of by their baseline i–j), with associated (real) thermal noise variances σ_j^2, then the reduced χ^2 is

$$\chi^2_{\text{vis}}(\mathbf{I}) = \frac{1}{2N_V} \sum_j \frac{\left|V_j - \hat{V}_j\right|^2}{\sigma_j^2}, \tag{5.16}$$

where $\hat{V}_j$ are the sampled visibilities from the Fourier transform of the trial image $\mathbf{I}$ on the same baselines as the measured V_j .

If the visibility phases are significantly corrupted by atmospheric turbulence or other phase errors, a χ^2 term that uses only the visibility amplitudes A_j (debiased by Equation 5.9) can be used:

$$\chi^2_{\text{amp}}(\mathbf{I}) = \frac{1}{N_V} \sum_j \frac{(A_j - \hat{A}_j)^2}{\sigma_j^2}, \tag{5.17}$$

where $\hat{A}_j = \left|\hat{V}_j\right|$.

Because the closure phase is robust to station-based phase errors such as those introduced by at-

mospheric turbulence, a χ^2 term defined on the bispectrum must sometimes be used instead of the complex visibility χ^2_{vis} (Equation 5.16). If N_{B} is the number of independent bispectrum measurements, and σ^2_{B} is the estimated variance on each complex bispectrum measurement (Equation 5.10), then

$$\chi^2_{\text{bispec}}(\mathbf{I}) = \frac{1}{2N_{\text{B}}} \sum_j \frac{\left| V_{\text{B},j} - \hat{V}_{\text{B},j} \right|^2}{\sigma^2_{\text{B},j}}, \tag{5.18}$$

where $\hat{V}_{\text{B},j}$ is the sampled bispectrum value corresponding to the trial image $\mathbf{I}$.

Similarly, a χ^2 term can also be defined using only the N_ψ measured closure phases (typically $N_\psi = N_{\text{B}}$, but trivial closure phases may be dropped from the fit). Defining σ^2_ψ as the estimated closure phase variances (Equation 5.11), a natural χ^2 term that automatically respects 2π phase wraps in the difference of measured and trial image closure phases ψ is

$$\begin{aligned} \chi^2_{\text{cl phase}}(\mathbf{I}) &= \frac{1}{N_\psi} \sum_j \frac{|\mathrm{e}^{\mathrm{i}\psi_j} - \mathrm{e}^{\mathrm{i}\hat{\psi}_j}|^2}{\sigma^2_{\psi,j}} \\ &= \frac{2}{N_\psi} \sum_j \frac{1 - \cos(\psi_j - \hat{\psi}_j)}{\sigma^2_{\psi,j}}, \end{aligned} \tag{5.19}$$

where the $\hat{\psi}_j$ are the sampled closure phases from the trial image on the same triangles as in the set of measurements ψ_j.

Similarly, a data term that uses only the closure amplitudes is

$$\chi^2_{\text{cl amp}} = \frac{1}{N_{\text{C}}} \sum_j \frac{(A_{\text{C},j} - \hat{A}_{\text{C},j})^2}{\sigma^2_{\text{C},j}}, \tag{5.20}$$

where there are a total of N_{C} measured independent closure amplitudes $A_{\text{C},j}$, the $\hat{A}_{\text{C},j}$ are the corresponding sampled closure amplitudes of the trial image, and the $\sigma^2_{\text{C},j}$ are the estimated variances of the measured closure amplitudes (Equation 5.12).

As discussed in Section 5.1.4, because closure amplitudes are formed from the quotient of visibility

amplitudes, the noise on the closure amplitudes (Equation 5.6) may be highly non-Gaussian. The logarithm of the closure amplitude will remain approximately Gaussian at lower SNR, so using the χ^2 of the log closure amplitudes is often a better choice in practice. In this case, the χ^2 term is

$$\chi^2_{\log \text{ cl amp}} = \frac{1}{N_\text{C}} \sum_j \frac{1}{\sigma^2_{\log \text{C},j}} \left(\log \frac{A_{\text{C},j}}{\hat{A}_{\text{C},j}} \right)^2 , \tag{5.21}$$

where the variance of the log closure amplitude $\sigma^2_{\log \text{C},j} = \sigma^2_{\text{C},j} / \left|A_{\text{C},j}\right|^2$ (Equation 5.13).

5.2.3 Regularizer terms

This section discusses the primary regularizer terms used in the tests in this chapter and implemented in the `eht-imaging` software library (Chapter 6). The first regularizer is a simple relative entropy (Frieden, 1972; Gull & Daniell, 1978; Narayan & Nityananda, 1986) which rewards pixel-to-pixel similarity to a "prior image" with pixel values P_i:

$$S_\text{MEM} = -\frac{1}{\zeta} \sum_i I_i \log \left(\frac{I_i}{P_i} \right) . \tag{5.22}$$

The summation index i runs over all $M = m \times m$ pixels in the square image. ζ is a normalization factor chosen to make the regularizing function independent of the image dimensions, total flux density, and field of view. For MEM, the units of the quantity inside the sum scale only with the total flux density f, so $\zeta = f$, where f is fixed before imaging.

In the absence of data, image entropy enforces consistency with the prior image P_i. Another reasonable choice in the absence of data is to prefer images with sparse brightness distributions. The simplest way to enforce image sparsity is by using the ℓ_1 norm as a regularizing function

(Honma et al., 2014). The ℓ_1 norm is simply the sum of the absolute pixel intensities:

$$S_{\ell_1} = -\frac{1}{\zeta}\sum_i |I_i|\,, \tag{5.23}$$

where again $\zeta = f$. The simple ℓ_1 norm in Equation 5.23 can be extended to prefer similarity to a prior image P_i or to strongly enforce sparsity in some regions of the image but not others. Furthermore, because the derivative of Equation 5.23 is not continuous, it may be preferable to use a smoothed version of the absolute value operator. These extensions to ℓ_1 regularization have been implemented in the `SMILI` imaging library (Akiyama et al. 2017a,b; Paper IV). While the simple ℓ_1 of Equation 5.23 is implemented in the `eht-imaging` library and was used in the EHT data reconstructions presented in Chapter 7, the remaining reconstructions in this chapter do not use ℓ_1.

The next regularizer is an isotropic total variation (TV) term that pushes the final image to favor pixel-to-pixel smoothness. Specifically, TV is an ℓ_2 norm on the image gradient; it favors piecewise-smooth images with flat regions separated by sharp edges (Rudin et al., 1992):

$$S_{TV} = -\frac{1}{\zeta}\sum_l\sum_m \left[(I_{l+1,m} - I_{l,m})^2 + (I_{l,m+1} - I_{l,m})^2\right]^{1/2}, \tag{5.24}$$

where in the above equation the two sums are taken over the two image dimensions and the image pixels $I_{l,m}$ are now indexed by their position (l, m) in the 2D $m \times m$ grid. Because Equation 5.24 contains finite differences between neighboring pixel intensities, the normalization factor ζ depends on the pixel size $\Delta\theta$ relative to a standard image size, here taken as the interferometer beam $\theta_{\rm beam}$. In particular, $\zeta = f(\Delta\theta/\theta_{\rm beam})$. It should be noted that the total variation in Equation 5.24 is not everywhere differentiable, so care must be taken when using it in imaging. Thiébaut & Young (2017) present a differentiable hyperbolic form of an edge-preserving smoothness regularizer (Charbonnier

et al., 1997) which approximates TV when the image is far from being smooth (i.e., when S_{TV} is large).

The reconstructions described in Section 5.4.2 use a "Total Squared Variation" (TSV) regularizer instead of TV. While still promoting image smoothness, TSV prefers smooth edges over the piecewise smooth patches favored by TV (Kuramochi et al., 2018). This property implies TSV may be more appropriate for astronomical image reconstruction. The TSV regularizer term is formed by squaring each of the terms in the sum in Equation 5.24:

$$S_{TSV} = -\frac{1}{\zeta} \sum_l \sum_m \left[(I_{l+1,m} - I_{l,m})^2 + (I_{l,m+1} - I_{l,m})^2 \right]. \tag{5.25}$$

The normalization factor for TSV is $\zeta = f^2 \left(\Delta\theta/\theta_{\text{beam}}\right)^4$.

The remaining regularizers constrain image-averaged properties. First, because closure amplitudes are independent of the normalization of the image, reconstructions made with only closure data require a constraint on the total image flux density. This constraint can be implemented as a regularizer term:

$$S_{\text{tot flux}} = -\frac{1}{\zeta} \left(\sum_i I_i - f \right)^2, \tag{5.26}$$

where the sum is again over the total M pixels in the image, and f is again the total source flux density, considered to be known a priori (e.g., by a simultaneous measurement of the source by a flux-calibrated single station). The normalization factor $\zeta = f^2$.

Next, because closure phase does not constrain the position of the image centroid, it is helpful to include a regularizing constraint to center the image in the chosen field of view:

$$S_{\text{centroid}} = -\frac{1}{\zeta} \left[\left(\sum_i I_i x_i - f\delta_x \right)^2 + \left(\sum_i I_i y_i - f\delta_y \right)^2 \right], \tag{5.27}$$

where (x_i, y_i) is the coordinate of the ith pixel and the desired image centroid position is (δ_x, δ_y). In the `eht-imaging` library, S_{centroid} pushes the image center of brightness to the frame center, $(\delta_x, \delta_y) = (0, 0)$, by default. The normalization factor for the centroid constraint is $\zeta = f^2\theta^2_{\text{beam}}$.

When only closure phases and closure amplitudes are used in the reconstruction, both the centroid and the total flux density are completely unconstrained by data. Thus, in this case almost any amount of weight on $S_{\text{tot flux}}$ and S_{centroid} should guide the final image to a centered image with the specified total flux, and the precise weighting of these terms relative to the data is not as significant in informing the final image as the relative weighting of the other regularizing terms.

The regularizers presented above are used for all of the data sets imaged in this chapter, but their relative weighting (the β_R terms in Equation 5.14) and the prior image used in S_{MEM} (Equation 5.22) are adjusted based on the data set considered. However, when comparing images produced with different data terms, the same prior image and relative regularizer weightings were used in the different reconstructions to produce fair comparisons (see Table 5.3).

5.3 "Superresolution"

Both RML and CLEAN are nonlinear methods that input some amount of prior information into the imaging process. Thus, one might reasonably expect these methods to produce some degree of image "superresolution," or the production of image features on scales less than the array nominal resolution $\theta_{\text{min}} = \lambda/b_{\text{max}}$, where b_{max} is the length of the longest baseline in the VLBI array. In the context of MEM, it is a frequently quoted result that the method has a superresolution factor of 1/4 the nominal resolution (Narayan & Nityananda, 1986). This fact is a consequence only of the analyticity of the data (not on the specific choice of prior or formulation of the "entropy" term), but the derivation of this factor requires the assumption of infinite signal-to-noise (Holdaway, 1990). In

practice, the superresolution factor may be informed by both the analyticity of the data (degraded by noise) and the choice of regularizing function(s).

This section demonstrates RML's capacity for "superresolution" by comparing RML reconstructions using only the simple entropy regularizer $S_{\rm MEM}$ (Equation 5.22) to CLEAN reconstructions from simulated Sgr A* data with thermal noise from the EHT 2017 array. For this simple test, the effects of inaccurate amplitude calibration, atmospheric phase corruption, and interstellar scattering toward Sgr A* were neglected. The RML algorithm used full complex visibilities (with calibrated phase information) via the χ^2 term in Equation 5.16. This choice, while infeasible in practice for EHT data due to phase errors, allows the RML images to be directly compared to the CLEAN reconstructions without self-calibration.

The RML and CLEAN reconstructions from the same calibrated data were convolved with a sequence of Gaussian kernels scaled from the elliptical Gaussian fitted to the Fourier transform of the (u, v) coverage (the "clean beam"). Each of these restored reconstructions was then compared to the ground truth image using the normalized root-mean-square error (NRMSE) metric. The NRMSE is a point-to-point metric that evaluates images based on pixel-to-pixel similarities rather than common large-scale features. Given two images $\mathbf{A}$ and $\mathbf{B}$ with M pixels each, the NRMSE of image $\mathbf{A}$ relative to $\mathbf{B}$ is

$$\mathrm{NRMSE}(\mathbf{A}, \mathbf{B}) = \frac{\sqrt{\sum_{i=1}^{M} (A_i - B_i)^2}}{\sqrt{\sum_{i=1}^{M} B_i^2}}. \tag{5.28}$$

In computing the NRMSE of the CLEAN reconstructions, the dirty image residuals were not added back to the convolved model. After tuning the CLEAN reconstruction parameters for this image, the total flux left in the residuals was less than 2% of the total image flux. The CLEAN reconstruction used Briggs weighting and a loop gain of 0.025, with the rest of the parameters set to the default in the algorithm's CASA implementation.[6]

[6] http://casa.nrao.edu/docs/TaskRef/clean-task.html

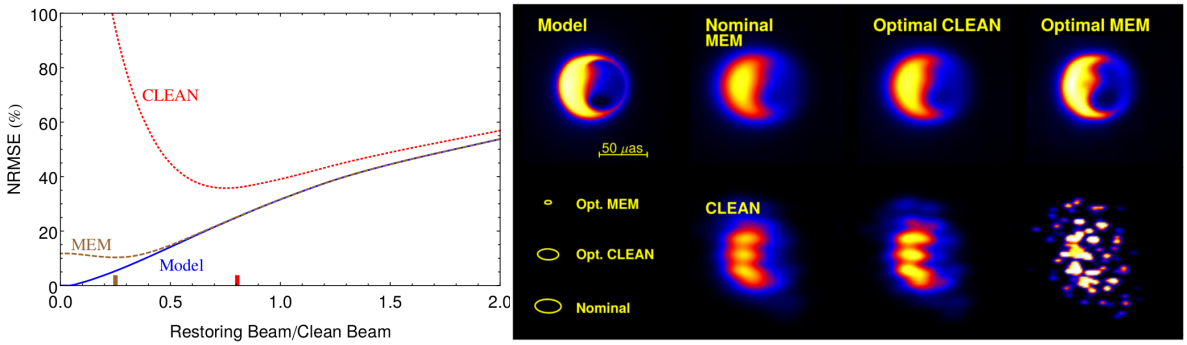


Figure 5.3: (Left) Normalized root-mean-square error (NRMSE, Equation 5.28) of RML (using an entropy regularizer, and here called MEM) and CLEAN reconstructed images as a function of the fractional restoring beam size. For comparison, the NRMSE of the blurred model image is also plotted. The reconstructed images were produced using simulated data from the EHT array; for straightforward comparison with CLEAN, realistic thermal noise was added to the simulated visibilities but gain calibration errors, random atmospheric phases, and blurring due to interstellar scattering were all neglected. The images were convolved with scaled versions of the fitted clean beam. The minimum for each NRMSE curve indicates the optimal restoring beam, which is significantly smaller for MEM (25% of nominal) than for CLEAN (78% of nominal). (Right) Example reconstructions restored with scaled beams from curves in the left panel. The center-left panels are the MEM and CLEAN reconstructions restored at the nominal resolution, with the fitted clean beam. The center-right panels show the reconstructions restored with the optimal beam for the CLEAN reconstruction and the far right panels show both reconstructions restored with the optimal MEM beam.

The results of this experiment are displayed in Figure 5.3. The left panel indicates that the RML image's curve of NRMSE with restoring beam size has a minimum at a significantly smaller beam size than the CLEAN reconstruction, demonstrating a superior ability to superresolve source structure. Furthermore, the absolute value of NRMSE from the MEM reconstruction is consistently lower than that from CLEAN for all values of the restoring beam size. Most importantly, while the CLEAN curve's NRMSE increases rapidly for restoring beams smaller than the optimal resolution, the RML image fidelity is relatively unaffected by choosing a restoring beam that is too small. Choosing a restoring beam that is too large produces an image with the same fidelity as the model blurred to that resolution. The right panel of Figure 5.3 shows the model image, the interferometer "clean" beam, and the reconstructions blurred with the nominal clean beam and the measured optimal beams. In addition to lower resolution and fidelity, the CLEAN reconstructions show prominent striping features from isolated components being restored with the restoring beam.

While Figure 5.3 demonstrates that in this case the MEM reconstruction has superior resolution

and fidelity to the CLEAN reconstruction, the optimal fractional restoring beam size for the CLEAN reconstruction is still less than unity. This result was observed in several similar reconstructions, suggesting that shrinking the restoring beam used in CLEAN reconstructions to 75% of the nominal fitted beam can enhance resolution without introducing imaging artifacts, at least on images of compact sources similar to those used in this test.

Repeating the exercise of Figure 5.3 with observations taken with increased or decreased signal-to-noise ratio resulted in NRMSE curves that were only slightly higher and lower than the curves in Figure 5.3, but shared the same form – in particular, the location of the minimum NRMSE values barely shifted. This insensitivity to additional noise is likely due to the overall high SNR of the original synthetic observations, which had an average SNR of 178 and a minimum SNR of 13. These results show that with a high average SNR, increasing or decreasing the noise by up to an order of magnitude does not significantly affect the image reconstruction. Observations with an average SNR ~ 1, on the other hand, may show a drastic change in quality with small adjustments to the noise level.

5.4 Testing RML imaging with closure quantities

This section tests the effects of using different combinations of the data terms in Section 5.2.2 in RML imaging of simulated EHT data. Observations were simulated from model images and corrupted with different amounts of uncertainty in the complex station gains $G_i e^{i\phi_i}$ (Equation 5.3). Images from these data were then reconstructed using different data term combinations and the regularizers as described in Section 5.4.2. While data terms more complicated than complex visibilities have been used in past optical (e.g., Thiébaut & Young 2017) and radio (e.g., Honma et al. 2014; Bouman et al. 2016) image reconstructions, this test (first published in Chael et al.

Table 5.1: EHT 2017 station parameters used in the imaging tests in Section 5.4.

Facility	Location	Diam. (m)	SEFD (Jy)	X (m)	Y (m)	Z (m)
JCMT	Maunakea, Hawai'i	15	6000	-5464584.7	-2493001.2	2150654.0
SMA	Maunakea, Hawai'i	7(×6)	4900	-5464555.5	-2492928.0	2150797.2
SMT	Mt. Graham, Arizona	10	5000	-1828796.2	-5054406.8	3427865.2
APEX	Atacama, Chile	12	3500	2225039.5	-5441197.6	-2479303.4
ALMA	Atacama, Chile	40(×12)	90	2225061.2	-5440057.4	-2481681.2
SPT	South Pole	10	5000	0.01	0.01	-6359609.7
LMT	Sierra Negra, Mexico	50	600	-768715.6	-5988507.1	2063354.9
PV	Pico Veleta, Spain	30	1400	5088967.8	-301681.2	3825012.2

2018b) represents the first demonstration of full closure-only imaging with no calibrated amplitude or phase information.

The imaging framework described in Section 5.2, including all of the data terms introduced in Section 5.2.2 and the regularizers in Section 5.2.3, is implemented in the `eht-imaging` software library (Chael et al., 2016, 2018b).[7] Chapter 6 discusses the structure and capabilities of the `eht-imaging` library in more detail.

5.4.1 Models and synthetic data

Simulated data were generated on EHT baselines from several 230 GHz model images at the positions of the EHT's primary science targets: Sgr A* (RA: 17h 45m 40.04s, DEC: $-29°$ $0'$ $28.12''$) and M87 (RA: 12^h 30^m $49^s.42$, DEC: $+12°$ $23'$ $28.04''$). The model images were generated by performing general relativistic ray tracing and radiative transfer on the density and temperature distributions from two previously published GRMHD simulations of hot supermassive black hole accretion disks (specifically, Mościbrodzka et al., 2016b; Gold et al., 2017). Data were also simulated from a 7 mm VLBA image of the quasar 3C 273 (Jorstad & Marscher, 2016) rescaled to a smaller FOV of 250 μas and "observed" at the sky location of Sgr A*.

[7] In particular, these tests used a version of the `eht-imaging` library from 2017, available in static form at https://zenodo.org/record/1173414

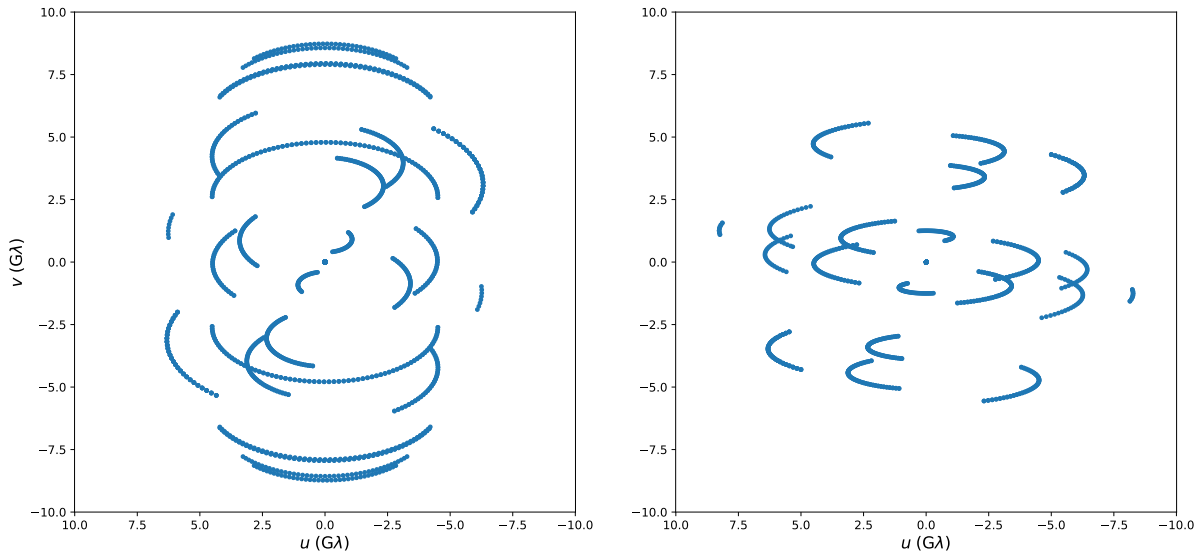


Figure 5.4: (Left) EHT 2017 (u, v) coverage for Sgr A*. The "redundant" (JCMT and APEX) stations make practically no unique contributions to the (u, v) coverage or nominal resolution aside from adding an effective zero baseline. However, these stations add closure amplitudes that are essential for closure amplitude imaging to approach the fidelity of imaging with visibility amplitudes. (Right) EHT 2017 (u, v) coverage for M87. Because the SPT cannot observe M87, the 2017 EHT has lower resolution on M87 than Sgr A*.

The EHT stations that observed in 2017 are the Atacama Large Millimeter/Submillimeter Array (ALMA), the Large Millimeter Telescope (LMT), the Submillimeter Array (SMA), the Submillimeter Telescope (SMT), the Institut de Radioastronomie Millimétrique (IRAM) telescope on Pico Veleta (PV), the IRAM Plateau de Bure Interferometer (PdB), and the South Pole Telescope (SPT). The EHT's station parameters are listed in Table 5.1.[8] In addition to the full EHT array described in Table 5.1, data were also sampled from a reduced array without the "redundant" sites JCMT and APEX. The (u, v) coverage maps for the 2017 EHT when observing Sgr A* and M87 are displayed in Figure 5.4.

In all cases, the integration time $\Delta t = 30$s and the bandwidth $\Delta\nu = 2$GHz, with scans taken every 5 minutes for a full 24 hour rotation of the Earth. The zenith opacity was set to $\tau = 0.15$ at all stations with no uncertainties in the opacity calibration. Neither the effects of refractive or diffractive interstellar scattering were included in simulated Sgr A* data (see e.g., Fish et al. 2014;

[8]Note that these parameters, in particular the SEFDs, are based on estimates made before the 2017 observations and so do not match the final numbers reported in Table 2 of Paper III.

Johnson 2016).

To test the quality of the different imaging methods on data with different levels of gain uncertainty, after adding thermal noise to the data, random gain terms were generated at seven different levels of uncertainty – 0%, 5%, 10%, 25%, 50%, 75%, and 100%. To corrupt the synthetic data with different levels of gain error, different time-dependent station-based complex gains were sampled from known underlying distributions with increasing variance. Because the atmospheric coherence time which determines the additional phase ϕ_i added at each station is much shorter than a typical observing cadence at 1.3 mm, the phases were drawn from a uniform distribution over $-\pi < \phi_i < \pi$ at each scan, independent of the uncertainty in the amplitude.

The prescription for the amplitude gain terms consisted of a random time-independent offset and a fluctuating part:

$$|G_i| = \sqrt{(1 + X_i)(1 + Y_i(t))}, \tag{5.29}$$

where X_i and Y_i are real Gaussian random variables with zero mean, but X_i is drawn only once per telescope per observation and Y_i is drawn independently at each time when (u, v) points are sampled. For simplicity, the underlying Gaussian distributions of X_i and Y_i had identical standard deviations, and this standard deviation labels the reported level of gain error (e.g., 5%, 10%, 15%). Once computed at different levels of gain error, the random sets of station-based gains were added to the ideal visibilities after adding Gaussian thermal noise according to Equation 5.3. To preserve the signal-to-noise ratio, the reported noise standard deviation terms σ_{ij} from Equation 5.8 were multiplied by the same gain factors G_i and G_j.

The source elevation angle and atmospheric opacity τ also affect the signal-to-noise at each station. The opacity attenuates the measured perfect visibility $\mathcal{V}_{ij}$ (before adding thermal noise) by a factor $\sqrt{\mathrm{e}^{-\tau_i/\sin\theta_i}\mathrm{e}^{-\tau_j/\sin\theta_j}}$, where θ_i and θ_j are the elevation angles of the source at the

Table 5.2: Initial/Prior image parameters for the imaging tests in this section.

Image	(u, v) coverage	FOV (μas)	Gaussian FWHM	Flux (Jy)(μas)
Figure 5.5	Sgr A*	135	60	2
Figure 5.6	M87	155	60	2
Figure 5.7	Sgr A*	375	25	2

Table 5.3: Imaging algorithm parameters for the tests in this section.

Round	f_{blur}	β_{MEM}	β_{TSV}	$\beta_{\text{tot flux}}$	β_{centroid}	α_1	$f_{\alpha 2}$	N_{iter}
1	N/A	1	1	100	100	100	2	50
2	0.75	1	50	50	50	100	0.75	150
3	0.5	1	100	10	10	100	0.5	200
4	0.33	1	500	1	1	100	1	200

different telescopes. This attenuation factor can be corrected by multiplying the measured visibility (including thermal noise) by its inverse using the *measured* opacity, keeping the reduced signal-to-noise constant. In general, the imperfect measurement of opacities introduces an additional source of amplitude gain error. For the purposes of this chapter, simulated data assumes the perfect measurement of opacities and sets all zenith opacities $\tau_i = 0.15$.

Note that the procedure used for this test is a simplified version of the full `eht-imaging` procedure for synthetic data generation, which includes the effects of polarimetric leakage with a Jones matrix formalism. The full `eht-imaging` procedure for synthetic data generation is described in Appendix C.

5.4.2 Imaging procedure

To aid in convergence and help the minimizer avoid local minima in the objective function, each imager was run multiple times for each dataset, substituting a version of the image produced by the previous convolved with a circular Gaussian as the next initial image. This procedure smooths out spurious high-frequency artifacts that the imager will not remove on its own given a lack of high-spatial-frequency data constraints. Each time the imager restarts, the `eht-imaging` script also adjusts the various hyperparameters α_D and β_R in Equation 5.14. The prescriptions for each data

set are presented below (in Table 5.2), but in general the approach is to generally increase the weight on the smoothness regularizer term to suppress the emergence of spurious high-frequency artifacts. The script also usually begins by weighting the closure phase data term more heavily in the reconstruction than is supported by the log-likelihood interpretation (Equation 5.15); typically, minimizing the closure phase χ^2 is the most helpful in constraining the overall image structure in early rounds of imaging. As the imager progress to later rounds, it restores the relative data term weighting to that given by Equation 5.15.

Every reconstruction used a 128×128 pixel grid. In each case, the objective function to be minimized (Equation 5.14) used one of four different data term combinations: the visibility bispectrum (Equation 5.18), visibility amplitude and closure phase (Equation 5.17 and 5.19), closure amplitude and closure phase (Equation 5.20 and 5.19), and log closure amplitude and closure phase (Equation 5.21 and 5.19). All reconstructions used the same four regularizer terms from Section 5.2.3: Maximum Entropy (Equation 5.22), Total Squared Variation (Equation 5.25), a total flux density constraint (Equation 5.26), and a centroid constraint (Equation 5.27). The initial/prior image was a circular Gaussian in all cases. The total flux densities, fields of view, and initial image Gaussian FWHMs are given in Table 5.2.

The parameters that specify the imaging procedure are listed in Table 5.3. As described in Section 5.4.2, the datasets were imaged in multiple rounds, blurring out the final image from a given round to serve as the initial image in the next. The FWHM of the circular Gaussian blurring kernel used is reported as a fraction f_{blur} of the nominal array resolution. The other imaging parameters listed in Table 5.3 include the data term and regularizer hyperparameters, α_D and β_R. For the data terms, in each case α_1 is the hyperparameter for the amplitude term (bispectrum, visibility amplitude, closure amplitude, or log closure amplitude), and α_2 is the hyperparameter for the closure phase term, if present. To capture this variation of α_2 with imaging round, α_2 is

parametrized by its ratio $f_{\alpha 2}$ with the ideal log-likelihood ratio given by Equation 5.15. That is, if there are N_1 measurements of the first (amplitude) data product and N_2 measurements of of the second (phase) data product,

$$\alpha_2 = f_{\alpha 2}\, \alpha_1\, \frac{N_2}{N_1}. \tag{5.30}$$

Finally, Table 5.3 also lists the maximum number of imager steps allowed in each round, N_{iter}.

5.4.3 Image evaluation

In evaluating the performance of the imaging algorithms with different data terms, the fidelity of the reconstructed images was assessed with the NRMSE metric introduced in Section 5.3. However, several factors complicate the simple application of Equation 5.28 in evaluating images reconstructed from closure quantities. First, often the true source image will contain fine-scale features that are at too high a resolution for any image reconstruction algorithm to capture given the longest projected baseline in the (u, v) plane. To prevent NRMSE from unduly penalizing reconstructions that successfully reconstruct the lower resolution features in the data, both the true and reconstructed images were convolved with a Gaussian kernel to blur out high-frequency structure. Since RML algorithms should provide some "super-resolution" above the scale corresponding to the longest projected baseline (Section 5.3), this kernel was chosen to have the same proportions as the interferometer "clean" beam – the Gaussian fitted to the central lobe of the Fourier transform of the (u, v) coverage – but with a beam size scaled down by a factor of $1/3$.

A second complication arises because images reconstructed without calibrated visibility phases are not sensitive to the true position of the image centroid in the field of view, so reconstructed images may be offset from the true source location. In addition, the number of pixels and field of view in the reconstructed image may be different from those in the true source image. Therefore,

when comparing images, the images were first resampled onto the same grid as the model image using cubic spline interpolation and then shifted to produce the maximal image cross-correlation before computing the NRMSE with Equation 5.28.

5.4.4 Results

The results are displayed in Figures 5.5, 5.6, and 5.7. Each Figure shows the initial model image, the initial model image blurred with a "clean" beam scaled to 1/3 of its fitted value, and the reconstructions from each method for each level of gain uncertainty, all blurred with the same beam. In the upper right, these Figures also display a plot showing the normalized root-mean-square error (Equation 5.28) for each method as a function of the level of gain error in the underlying dataset.

These results indicate that, as long as some redundant stations are included to constrain the reconstruction with "trivial" closure phases and amplitudes, closure-only imaging of EHT data can achieve fidelities nearly as good as bispectral or amplitude + closure phase imaging. As the level of amplitude gain error increases, the fidelity of the results produced using the bispectrum or visibility amplitudes drops quickly, while closure-only imaging is completely insensitive to gain error.

Figures 5.5–5.7 show that imaging with closure amplitudes directly can produce results that are slightly more faithful to the underlying image than reconstructing the image with log closure amplitudes. However, imaging with the closure amplitudes often takes much longer to converge, and it is more sensitive to the particular choices of data term weight and initial field of view. Choosing the weights and field of view incorrectly can cause the reconstruction using closure amplitudes to converge to an incorrect local minimum in its complicated energy landscape, while the log closure amplitude χ^2 term results in good images for a larger range of the parameter space.

Finally, for the narrow, high dynamic range 3C 279 image in Figure 5.7, the NRMSE was computed using the logarithm of the image. This choice results in a range of NRMSE values for the

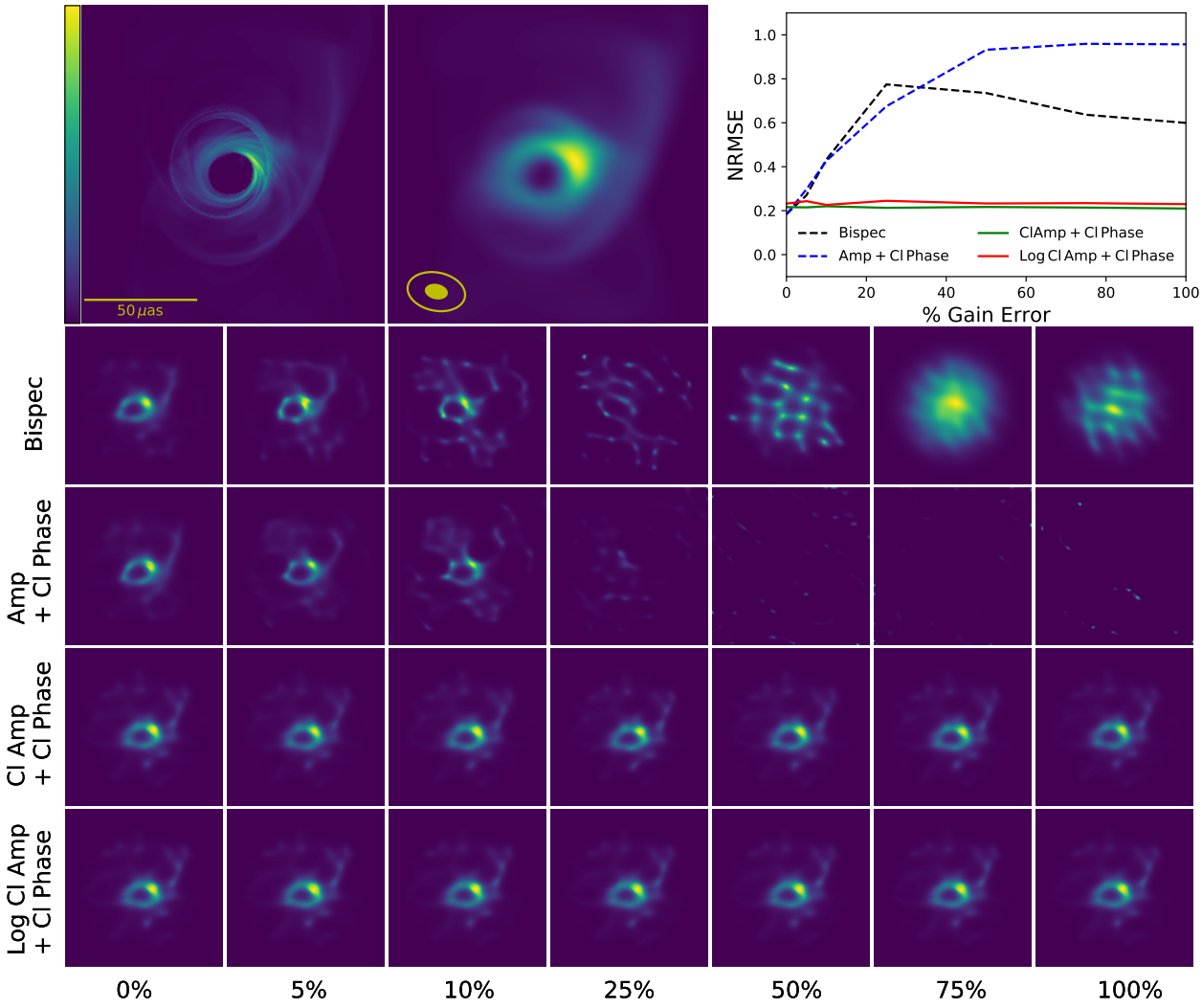


Figure 5.5: (Top left) 230 GHz image from a GRMHD simulation of Sgr A* (Gold et al., 2017). (Top middle) the same image blurred with the effective beam (solid ellipse), 1/3 the size of the fitted CLEAN beam (open ellipse). The image was observed at the sky location of Sgr A* using EHT 2017 baselines, and images were reconstructed with each method using the parameters in Table 5.3. (Top right) Curves of NRMSE (Equation 5.28) versus gain error for each reconstruction method. (Bottom) individual reconstructions from each method (y-axis) at each level of gain error (x-axis), blurred with the same beam as the model in the upper middle pane. The images and NRMSE curves show that except at the lowest levels of amplitude gain error, the closure-only results are as faithful to the model as the reconstructions that use either the bispectrum or visibility amplitudes and closure phases. Furthermore, the results of the closure-only methods are insensitive to the overall level of amplitude gain error, while the reconstructions using visibility amplitude information fail starting at the 10% level of gain error.

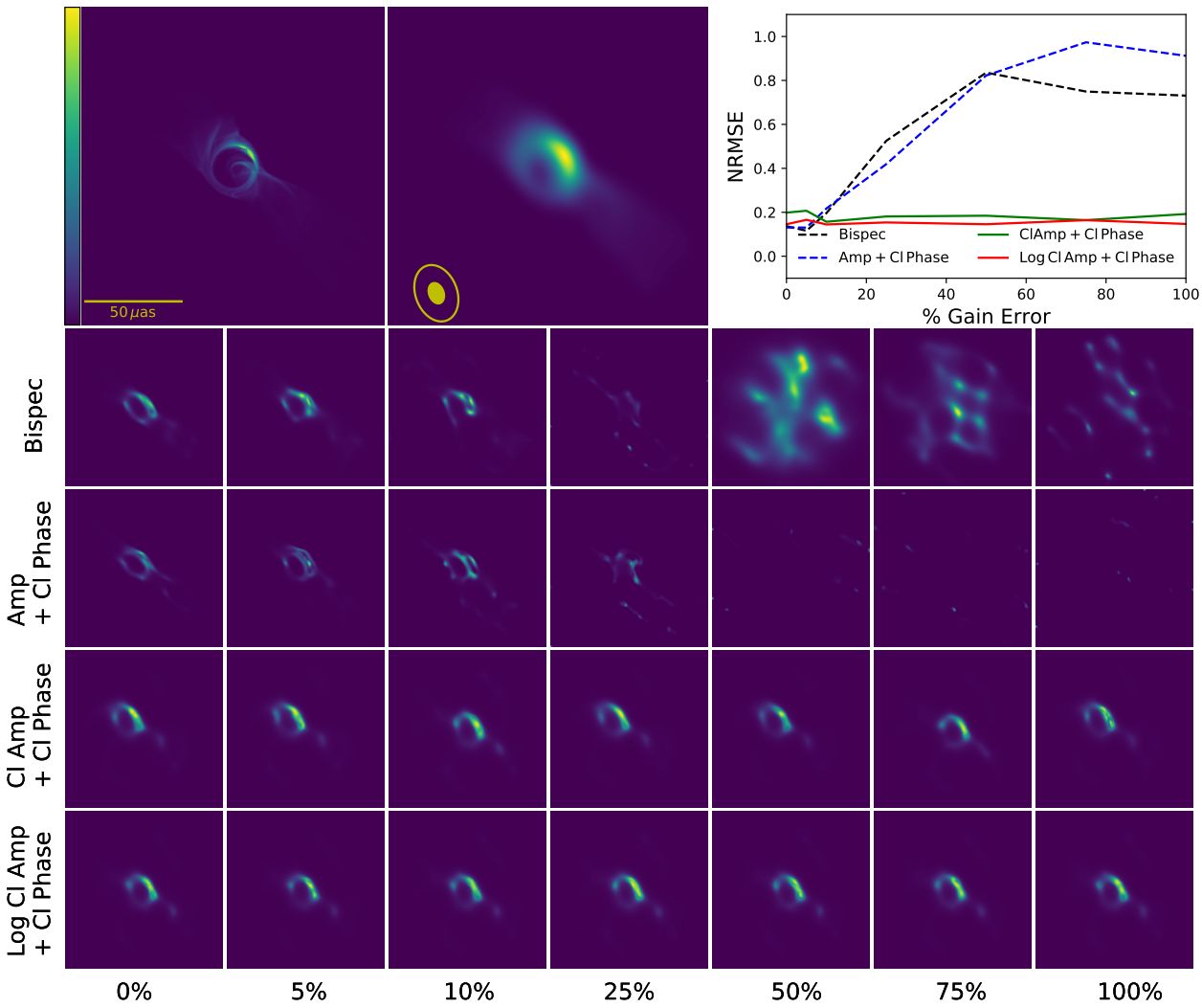


Figure 5.6: Reconstructions of a 230 GHz image from a GRMHD simulation of the M87 jet (Mościbrodzka et al., 2016b). As in the Sgr A* image in Figure 5.5, closure-only methods produce results that are as good or better than the bispectrum or visibility amplitude + closure phase methods in all but the zero gain error case, and the closure-only results are consistent at all levels of gain error. In contrast, the methods that rely on calibrated amplitudes begin to fail at the 10% level of gain error.

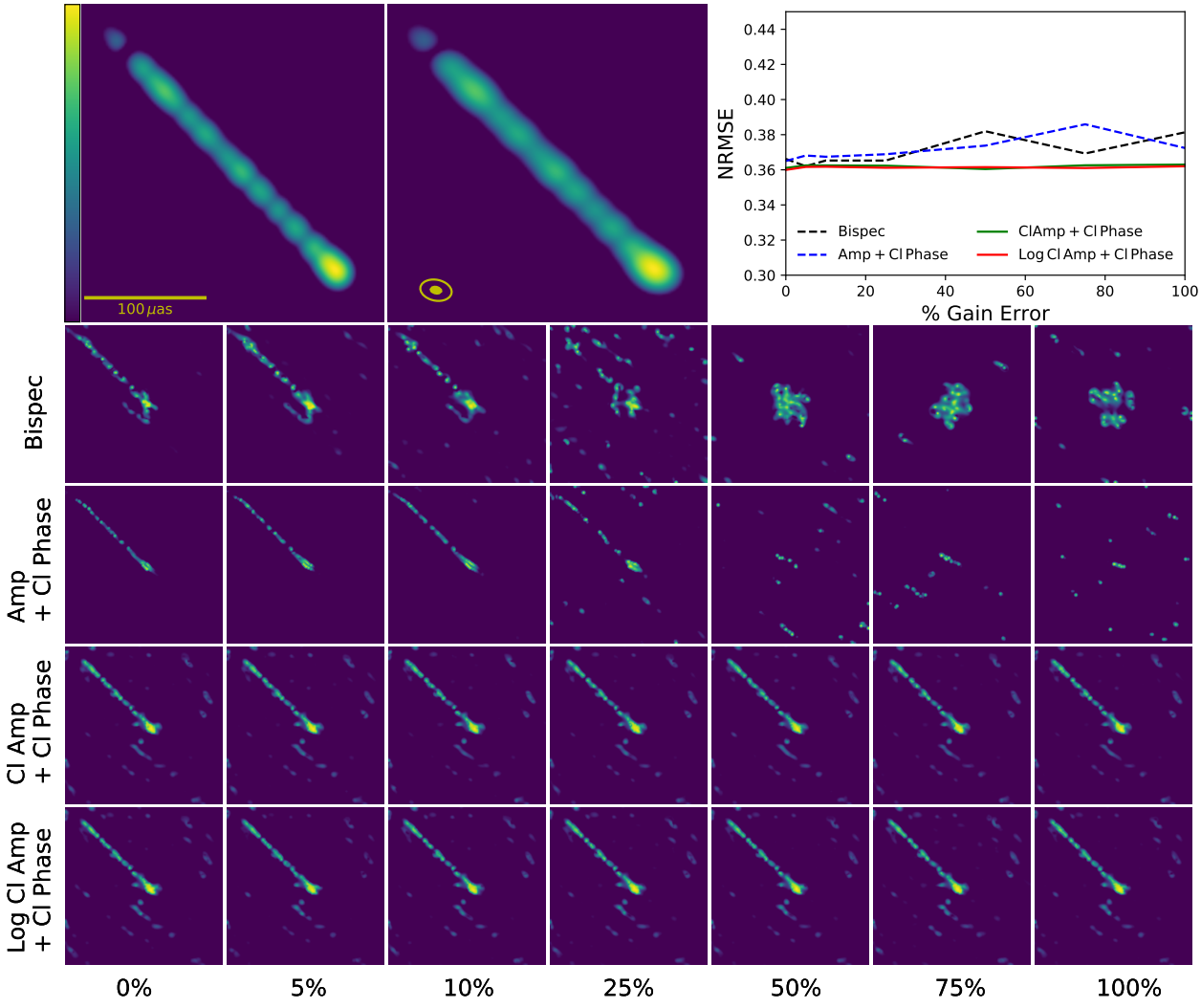


Figure 5.7: 43 GHz VLBA image of 3C 273 from Jorstad & Marscher (2016), scaled to a 250 μas field of view. Simulated data were generated using the 230 GHz EHT 2017 Sgr A* (u, v) coordinates and sensitivities (Figure 5.4). Unlike the other images in this section, this image is displayed with a log scale, and the NRMSE was computed from the log of the image. The closure-only reconstructions again capture the overall jet structure at all levels of amplitude gain error. With no gain error, imaging directly with closure amplitudes (or log closure amplitudes) instead of visibility amplitudes provides less dynamic range, as is evident from the spurious low-luminosity off-axis features in the closure-only reconstructions.

bispectrum and visibility amplitude + closure phase images that is substantially lower than those in Figures 5.5 and 5.6. However, visual inspection of the images shows that in this case, as in Figures 5.5 and 5.6, imaging methods that rely on calibrated amplitudes perform significantly worse with increasing gain error and completely fail with amplitude gain error levels >10%. In contrast, the closure-only methods have consistent performance at all levels of amplitude gain. However, the final dynamic range achieved in the closure-only reconstructions is worse than in the images produced with visibility amplitudes with zero gain error, as is evident from the spurious low-luminosity features in the closure-only reconstructions in Figure 5.7. These features parallel to the jet axis result from a local minimum of the objective function, which is invariant to overall image shifts. Since there are no data constraints on certain spatial frequencies due to sparse coverage, these Fourier components can be made large through periodic structure without increasing χ^2. Defining a masked region along the jet axis outside which the flux is zero (analogous to a CLEAN box) may help remove these features.

Figure 5.8 compares reconstructions using data from the full EHT 2017 array and the 2017 array without "redundant" stations. In both cases, closure-only methods converge to the same image for all values of systematic gain error. However, without redundant stations the results are substantially less accurate; when using a redundant stations in the dataset, the closure-only results approach the fidelity of images produced with gain-calibrated amplitudes. "Redundant" stations contribute important short baselines that combine into nontrivial closure amplitudes and act to further constrain the underlying image (Section 5.1.3). In other words, the closure-only images approach the bispectrum or amplitude + closure phase images in quality as the number of closure amplitudes increases, even if some of those closure quantities contain zero-baseline measurements from co-located stations.

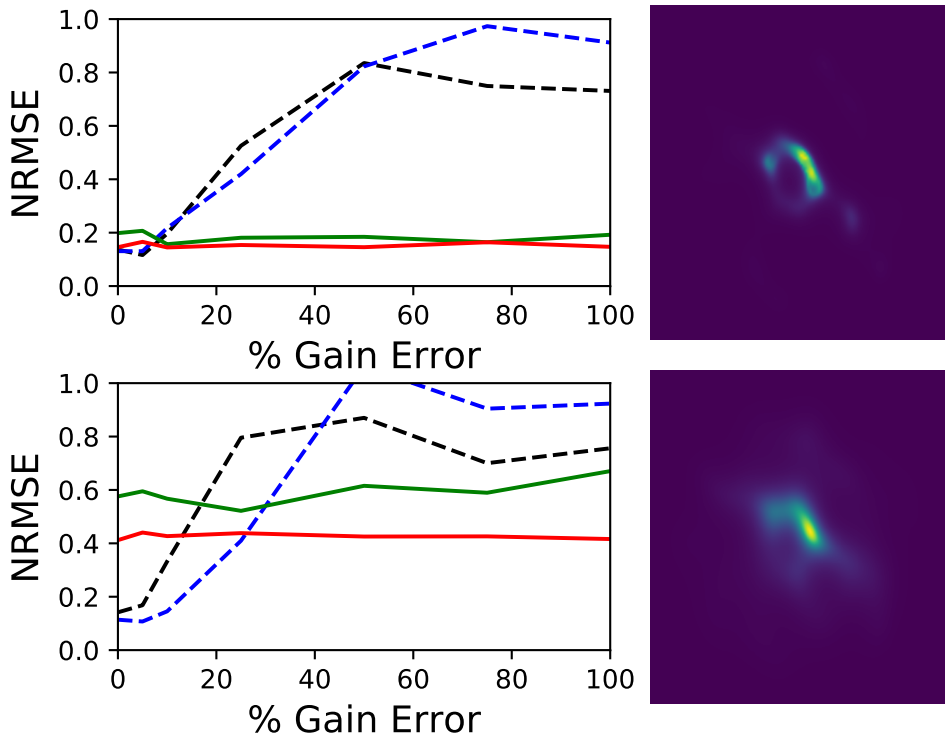


Figure 5.8: (Top) Image fidelity with the EHT 2017 array. The left panel shows NRMSE curves of image fidelity for reconstructions of the model in Figure 5.5 with different levels of gain error. The curves are styled consistently with those in Figures 5.5–5.6. The right image is the log closure amplitude + closure phase reconstruction produced at 100% gain error. (Bottom) Image fidelity with the EHT 2017 array without redundant stations (JCMT, APEX). The reconstructions from data without the redundant stations are still insensitive to different levels of gain error, but their overall fidelity is worse compared with those produced from data including these redundant stations.

5.5 VLBA and ALMA closure images

To test its performance on real observations, this section applies closure quantity imaging algorithms on millimeter-wavelength interferometric datasets from the VLBA and ALMA. In both cases, the number of visibilities and closure quantities greatly exceeds the number produced by the sparse EHT 2017 array.

Figure 5.9 compares CLEAN and closure-only RML imaging of a VLBA observation of M87 at 7 mm wavelength. In this case, and for other images with jets or narrow structure (see Figure 5.7), a major difficulty in closure-only imaging is convergence in the minimization of the objective function (Equation 5.14). When the algorithm is initialized with an uninformative image, the algorithm has

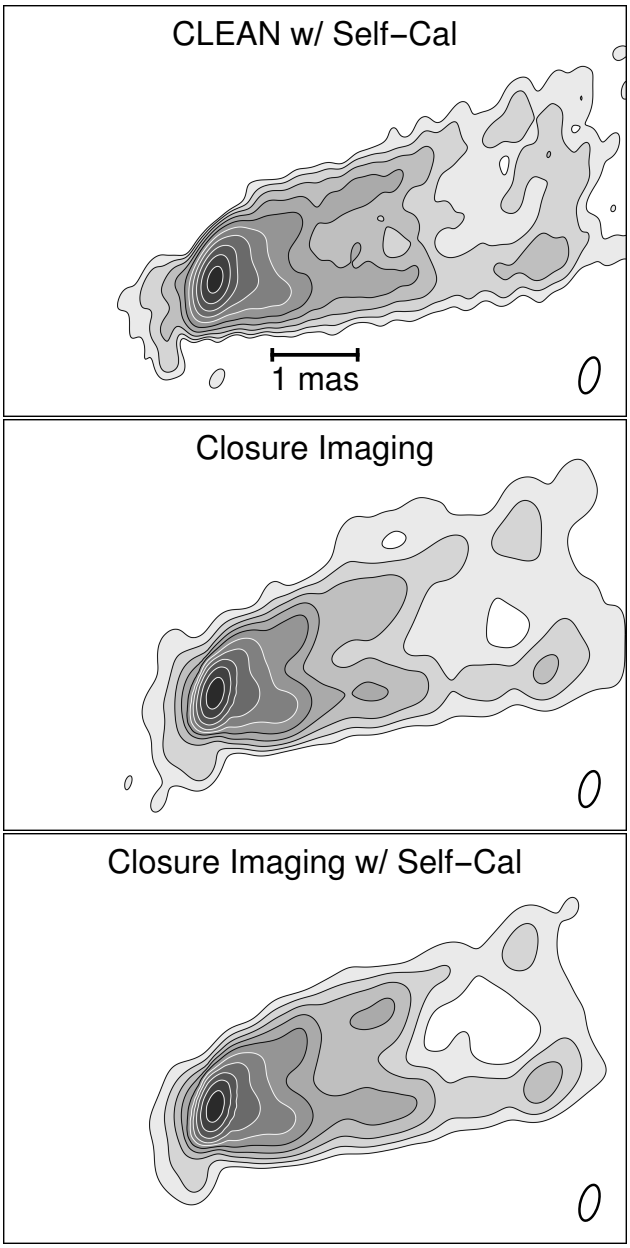


Figure 5.9: Application of closure-only imaging to a VLBA observation of M87 at 7mm wavelength observed on May 9, 2007 (for details, see Walker et al., 2016, 2018). (Top) CLEAN image made using iterative imaging and self-calibration. (Center) image reconstructed using closure-only imaging with a weak visibility amplitude constraint to aid initial convergence. (Bottom) image reconstructed using complex visibilities after self-calibrating to the closure-only image. To simplify the comparison between these approaches, each image has been convolved with the same CLEAN restoring beam and each image is rescaled to have the same total flux density as the CLEAN image. Contours in all panels are at equal levels, starting at $9.7\ \mathrm{mJy/mas^2}$ ($=1\ \mathrm{mJy/beam}$) and increasing by factors of 2.

difficulty converging to an image that has a reduced chi-squared near 1 in either the closure phases or the log closure amplitudes.

To mitigate this problem while still preserving the benefits of only using closure quantities, initially including the poorly-calibrated visibility amplitudes in the minimization with a low weight can significantly aid the initial convergence. To avoid any bias from the initial calibration, this test first assumed a "null" calibration with a single, constant SEFD (see Equation 5.8) for all sites and all times. The RML algorithm used closure quantities and these minimally-calibrated visibility amplitudes, with the visibility amplitude χ^2 term down-weighted by a factor of 10 relative to the closure quantities. The imaging script then performed another two rounds toward convergence, initializing to the previous image convolved with a circular Gaussian matching the nominal array resolution, but with visibility amplitudes this time down-weighted by a factor of 100. Finally, the last two rounds of imaging used only closure quantities and no visibility amplitudes.

Figure 5.9 compares the reconstructed image to an image reconstructed using CLEAN and iterative self-calibration (Walker et al., 2016, 2018). The self-calibrated gains from the final closure-only image are significantly different than the initialized "null" calibration solution; after normalizing to the median gain (effectively fixing the total flux density), although 50% of visibilities had residual gain corrections of less than 3%, 10% of the visibilities had residual gain corrections of more than 30%. This result justifies post-hoc the choice to use visibility amplitudes in the initial minimization steps. The majority of uncalibrated amplitudes have low error compared to the final self-calibrated set, so they are useful in aiding convergence; however, relying primarily on closure amplitudes ensures a final image that is less affected by the large outlier gain errors present on some baselines.

For comparison, the derived self-calibration solution was also applied to the data and used to produce an RML image with the self-calibrated complex visibilities (minimizing Equation 5.14 with a standard complex visibility χ^2 term, Equation 5.16). The result is displayed in the third panel of

Figure 5.9. All three methods in Figure 5.9 give results that are broadly consistent, demonstrating the potential of closure imaging to obtain images that are comparable to those obtained by multiple rounds of finely-tuned CLEAN and self-calibration from an expert user.

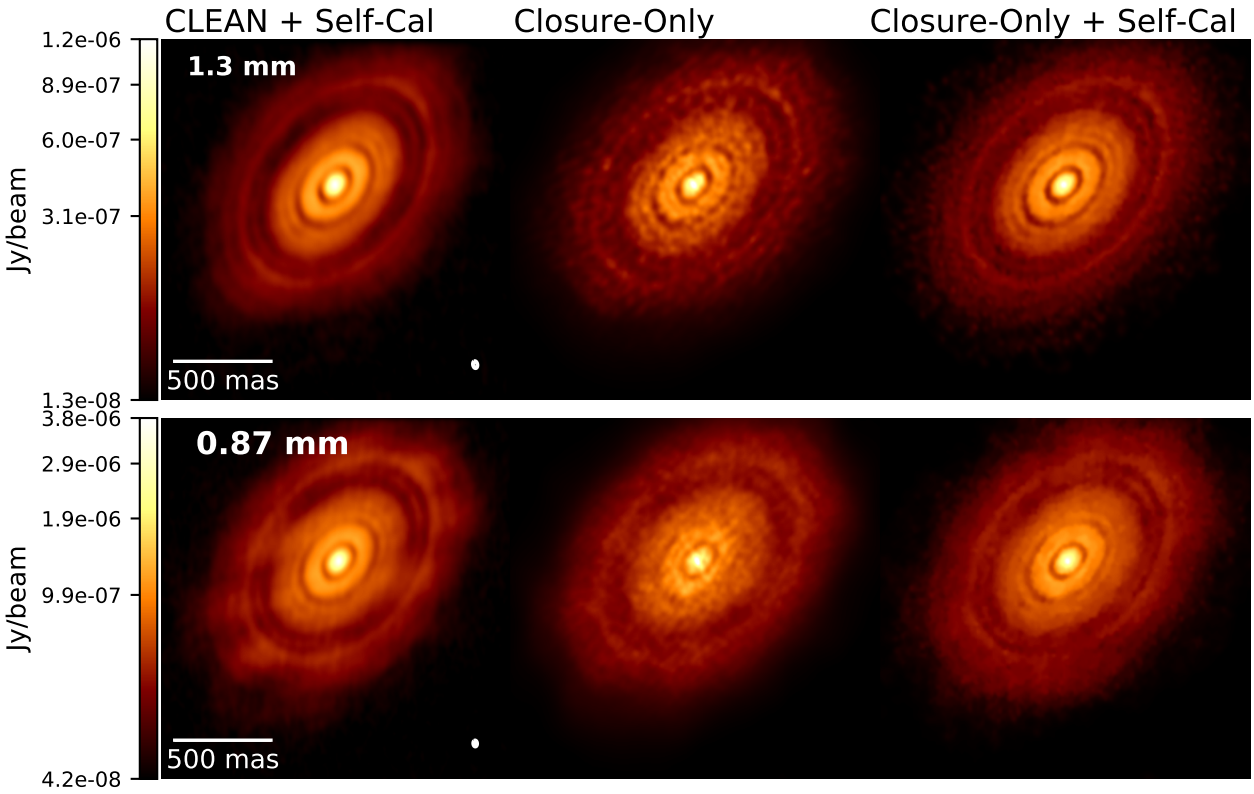


Figure 5.10: (Top) 1.3 mm band 6 ALMA image of the protoplanetary disk around HL Tau, comparing the CLEAN reconstruction from ALMA Partnership et al. (2015) with reconstructions from RML using closure quantities. The leftmost panel shows the CLEAN image with a FOV of $1.8''$. The center panel shows an image of the same data produced by directly fitting to log closure amplitudes and closure phases, with downweighted visibility amplitudes used in the initial steps to aid convergence. Closure-only imaging produces an image that is consistent with the CLEAN result, despite not using any multi-scale imaging, but the overall resolution is lower. The rightmost panel shows an image produced from complex visibilities using a strong total variation regularizer after self-calibrating the data to the center closure-only image. After self-calibration, complex visibility imaging with total variation produces a sharp image with distinct disk gaps. (Bottom) 0.87 mm band 7 ALMA images, produced using the same imaging parameters as the top 1.3 mm images. The 0.87 mm image obtained after closure-only imaging and one round of self-calibration eliminates prominent clean artifacts (dark spots) present in the original image. The 0.87 mm image is similar to recently reprocessed images using CLEAN and a modified self-calibration loop (Akiyama et al., 2016).

With 64 telescopes, ALMA has baseline coverage that much more densely fills the (u, v) plane than the EHT and VLBA observations considered above. Figure 5.10 displays CLEAN images of a 2014 ALMA observation of the HL Tau protoplanetary disk taken both at 1.3 mm and 0.87 mm

(ALMA Partnership et al., 2015) as well as RML reconstructions with only log closure amplitude and closure phases. To produce RML reconstructions of this large data set, the data were first averaged in five minute intervals. As in the 3 mm VLBA reconstructions of the M87 jet (Figure 5.9), downweighted visibility amplitudes were used to help aid convergence in the initial steps of the minimization; the poorly-calibrated amplitudes were removed from the objective function in the final runs of the imager.

Figure 5.10 demonstrates that closure imaging is able to replicate the overall structure of the published CLEAN images of HL Tau, including all of the ring gaps in the protoplanetary disk identified by the original reconstruction. However, the `eht-imaging` library does not yet include multi-scale imaging (Wakker & Schwarz, 1988; Cornwell, 2008), which was necessary to produce the detailed structure in the CLEAN image. After producing an image from closure quantities, the data were self-calibrated to the resulting image (center panel of Figure 5.10) and imaged again directly using the resulting complex visibilities (minimizing Equation 5.14 with Equation 5.16). The resulting image has a higher resolution than the closure-only image alone, with sharper and more distinct gaps apparent in the disk (right panel of Figure 5.10). Furthermore, particularly in the 0.87 mm image, the final reconstruction lacks the prominent periodic dark spots present in the CLEAN image. These artifacts are likely caused by prominent dirty beam sidelobes resulting from amplitude miscalibration; they were also ameliorated in recently reprocessed CLEAN + self-calibration images by Akiyama et al. (2016).

5.6 Discussion

The results in Section 5.4.4 and Section 5.5 demonstrate that closure amplitudes and phases can be directly used in interferometric imaging to produce images that are insensitive to phase and

amplitude calibration errors. Traditional self-calibration and imaging loops require many iterations of CLEAN imaging and fitting complex gains to visibilities. These loops contain many tunable parameters, including the choice of initial source model, the strategy for independent or concurrent calibration of amplitude and phase gains, the CLEAN convergence criterion, the choice of taper and weighting for the CLEAN visibilities, and the scales and regions to clean in each CLEAN iteration.

Closure-only imaging does not remove all tunable parameters from the model, but imaging with closure quantities alone necessarily produces results that are less biased by calibration assumptions. Images from closure imaging can stand on their own as minimal assumption estimates of the source structure; alternatively, results from closure-only imaging may be used as a well-motivated self-calibration model or initial source image for other imaging pipelines using calibrated data. On the ALMA and VLBA datasets in Section 5.5, just one round of self-calibration to an image produced with closure quantities can be used to produce smooth high-resolution images that match the best iterative, multi-scale CLEAN + self-calibration results.

The most significant challenge in closure-only imaging is a difficulty in the early convergence and a tendency to quickly get stuck in wildly incorrect local minima. Counterintuitively, this tendency seems to be more of a problem in datasets with more interferometric baselines. This limitation may arise because the energy landscape represented by the closure amplitude terms (Equation 5.20) becomes increasingly complicated with more correlated closure data. When using simulated data from the sparse EHT array (Section 5.4.4), using closure quantities alone with a reasonable Gaussian prior and several imaging iterations is enough to guide the algorithm to converge on a reasonable image. However, imaging the real datasets from ALMA and the VLBA in Section 5.5 with closure quantities alone and an uninformative initial model was difficult. For these cases, adding a weak data constraint using uncalibrated visibility amplitudes (Equation 5.17) helped guide the minimizer to the region of a good local minimum. This constraint can be as

low as 1-10% of the closure amplitude χ^2 term and still produce excellent results. As the imaging proceeds, the weak amplitude constraint is removed, allowing the final image to be only guided by the closure amplitudes and phases. Given the robustness of the results in Figures 5.9 and 5.10 to different choices of regularizer and data weights, there is significant promise for this method to eventually allow for unsupervised closure imaging that can blindly produce a calibration-free image from decent initial data without user intervention.

A general characteristic of the closure-only images in Section 5.4.4 and Section 5.5 is their tendency to avoid high-frequency artifacts when highly converged. By removing spurious features from CLEAN images, closure methods could thus be useful in providing maximally conservative results. However, note the CLEAN reconstruction in Figure 5.9 recovers more extended structure along the jet than the closure-only solution, and the closure-only images of HL Tau in FIgure 5.10 show much less fine structure than the original CLEAN images. The additional features in the CLEAN images could be a result of the multi-scale approach used in both cases (Wakker & Schwarz, 1988; Cornwell, 2008), but a simpler explanation is just that the CLEAN reconstructions have access to more information in their self-calibrated data sets. However, errors in the self-calibration solution introduce artifacts in the CLEAN results, such as the depressions seen in the 0.87 mm image of HL Tau in Figure 5.10. Self-calibrating to a robust closure-only model recovers the reliable high-spatial-frequency data in the complex visibilities without introducing calibration errors; a process of closure-only imaging plus self-calibration can thus result in images that are both more reliable and higher resolution than the traditional CLEAN reconstructions. Going forward, implementing multi-scale RML approaches will further improve the imaging methods presented in this chapter.

The `eht-imaging` library used in this chapter provides flexible framework where images can be easily produced from the same data set using different data and regularizer terms. `eht-imaging`

can also be used to self-calibrate data, to generate synthetic data from images with realistic thermal error and calibration uncertainties, and for the general plotting, analysis, and comparison of interferometric data. Within this framework, it is easy to experiment with different arbitrary combinations of data terms and implement new imaging methods, such as polarimetric imaging (Chael et al., 2016), imaging in the presence of refractive scattering (Johnson, 2016), and producing continuous movies from multi-epoch observations (Johnson et al., 2017). The `eht-imaging` library is described in detail in Chapter 6.

5.7 Summary and conclusions

This chapter presented a framework for interferometric imaging using regularized maximum likelihood with arbitrary data products. This framework is implemented in the software library `eht-imaging` (Chapter 6). This work builds on decades of past work in applying regularized maximum likelihood approaches to interferometric imaging, and is in particular inspired by the simultaneous minimization of multiple data terms pioneered in optical interferometric imaging (e.g., Thiébaut 2013; Thiébaut & Young 2017). This chapter extends that framework by imaging data directly with closure amplitudes (or their logarithms) for the first time, rather than relying on amplitude self-calibration.

With closure-only imaging, self-calibration can be bypassed entirely, producing an image that will contain minimal calibration assumptions and will not depend on the choice of initial self-calibration model or other assumptions made in the self-calibration loop. Section 5.4.4 showed that this strategy performs well on simulated EHT data of Sgr A* and M87. Images produced using only closure quantities have consistent fidelity at all levels of amplitude gain or miscalibration. Furthermore, when redundant stations are included in the array, the overall fidelity of the results

approaches that of images made with perfectly calibrated data using conventional algorithms.

Section 5.5 demonstrates that closure imaging can also produce high quality images for VLBA and ALMA datasets at millimeter wavelengths, giving results that are of comparable quality to expert reconstructions with multi-scale CLEAN and self-calibration. Results from closure-only imaging can also be used to self-calibrate data and initialize additional imaging with the self-calibrated complex visibilities. For the HL Tau ALMA datasets considered in Figure 5.10, just one round of self-calibration and complex visibility imaging after closure-only imaging produced refined results with fewer suspicious features than the CLEAN reconstructions.

Techniques involving calibration-insensitive closure quantities like those presented in this paper can help push interferometric imaging to more and more challenging regimes, including higher frequencies than the EHT's current 230 GHz operating frequency. While applicable to all interferometric astronomical data, these techniques are especially valuable at millimeter and sub-millimeter wavelengths, where calibration uncertainties are a large and variable component of the error budget. Adding more data terms, this method can be easily generalized to produce polarimetric images (Chael et al., 2016; Akiyama et al., 2017a), spectral index maps, simultaneous multi-band images, scattering mitigation (Johnson, 2016), and dynamical movies of multi-epoch data (Bouman et al., 2018; Johnson et al., 2017).

Page intentionally left blank

Some text in section 6.5 was previously published in *ApJ* 857 (2018), 1, 23 (A. Chael, M. Johnson, K. Bouman, L. Blackburn, K. Akiyama, and R. Narayan).

6

The `eht-imaging` library

The Regularized Maximum Likelihood (RML) approach toward interferometric imaging described in Chapter 5 is a flexible framework. The basic idea behind RML imaging – finding the image that minimizes an objective function (Equation 5.14) that is a weighted sum of data consistency and regularizer terms – allows for creativity and experimentation in the imaging process, as imagers can design new data terms and regularizers to meet the needs of their particular problem.

RML-type methods have a long history in radio interferometry, particularly in the form of the Maximum Entropy Method (e.g., Cornwell & Evans, 1985; Narayan & Nityananda, 1986; Briggs, 1995; Holdaway & Wardle, 1990), but they have typically been implemented as a routine in a larger interferometric data analysis package focused on CLEAN imaging (e.g., AIPS or CASA). While RML algorithms in optical interferometry have more modern implementations, they have tended to be implemented one-by-one in individual pieces of software (see e.g., Thiébaut & Young, 2017, Table 1), making it difficult to compare different imaging strategies on the same data. The unique challenges of EHT data – including the lack of phase coherence at mm-wavelengths, extremely sparse (u, v) coverage, interstellar scattering, and intraday time variability of Sgr A* – have encouraged the development of new RML methods and imaging techniques. These methods have primarily

been collected and developed in two open-source software libraries: `eht-imaging` (Chael et al., 2016, 2018b, 2019a)[1] and SMILI (Akiyama et al., 2017a,b).[2]

The `eht-imaging` software library (often referred to as `ehtim`) is a flexible Python environment where RML algorithms can be used, experimented with, and developed.[3] `ehtim` was originally developed out of software for experiments with polarimetric EHT imaging (Chael et al., 2016). Since 2016, it has become one of the primary software libraries in the calibration, analysis, and imaging of EHT data. In addition to powering one of the three imaging pipelines that produced the first EHT images of the black hole shadow in M87 (Paper IV), functions developed in `ehtim` were used in calibrating the 2017 EHT data (Paper III) and in generating synthetic data from GRMHD simulations and model fits (Paper V; Paper VI). Furthermore, it has served as a testbed for novel imaging algorithms, including closure-only imaging (Chael et al., 2018b), the mitigation of refractive interstellar scattering (Johnson, 2016) and the dynamical reconstruction of evolving sources (Johnson et al., 2017; Bouman et al., 2018). While `ehtim` was designed with the EHT in mind, it is usable with any interferometric data set, including the VLBA and ALMA data imaged in Chapter 5. As of April 2019, the `ehtim` library has been used in the analyses of 18 published papers in the astronomical literature (including the first series of EHT results in 2019).

This chapter is not intended to be documentation for the `eht-imaging` code; the functions, methods, and classes described below are far from an exhaustive accounting of the capabilities of `ehtim`.[4] Instead, this chapter provides an overview of the code structure, primary classes, methods, and scripts that were developed in `ehtim` and used to help produce the first images of the M87 black hole shadow from EHT data (Paper IV).

[1] https://github.com/achael/eht-imaging
[2] https://github.com/astrosmili/smili
[3] `ehtim` is compatible with both Python 3 and Python 2.7
[4] The `eht-imaging` documentation can be found at https://achael.github.io/eht-imaging/

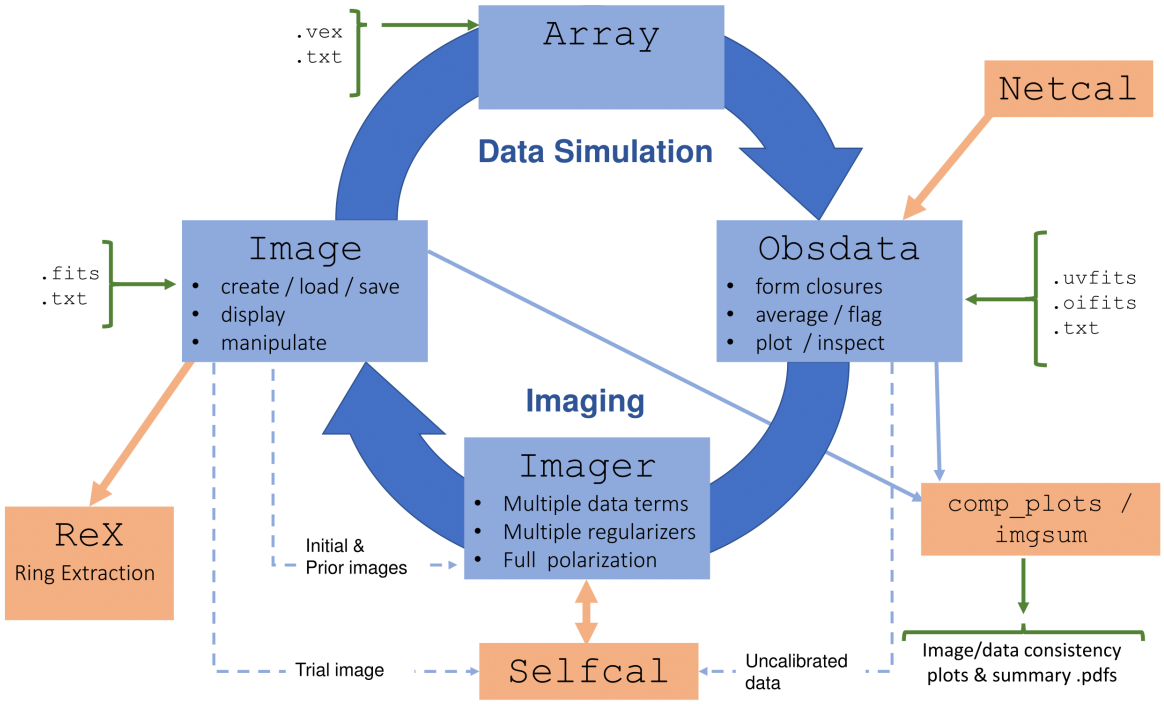


Figure 6.1: The structure of the `eht-imaging` software library.

6.1 Code outline

`eht-imaging` is a library of software; even in only a few years of development, it has accrued functions and capabilities that span a wide variety of tasks in interferometric data analysis. The core of the code, however, is divided into two main tasks; data simulation and imaging. In data simulation, an instance of the `Image` class is combined with an `Array` to produce data stored in an `Obsdata` object. In imaging, an RML algorithm defined in an instance of the `Imager` class acts on an `Obsdata` dataset (and an initial `Image`) to produce a reconstruction `Image`. This primary cycle of data simulation and imaging is summarized in Figure 6.1. The four main classes are discussed in Sections 6.3–6.5.

In addition to these primary classes and functionality, an `ehtim` user will typically encounter several other classes and functions in simulating, processing, and imaging interferometric data (Section 6.6). For instance, the `const_def` module defines many helpful physical constants and conversion factors used throughout the code. The `Movie` class extends the `Image` to time series of

frames (e.g., from a GRMHD simulation) that can then be mock observed in a single observation with the underlying time evolution. The `Vex` class can read in a VLBI .vex schedule file so data can be simulated on exactly the same baselines as a real observation. The `Caltable` class contains information about a time-series of complex station gains used to calibrate an `Obsdata` object. A `Caltable` can be derived by calibrating data to an image with the `calibrating.self_cal` module, or by enforcing network consistency if redundant sites are present using `calibrating.network_cal`. The `comp_plots` module in `ehtim` contains a variety of functions to compare data sets and images with different plots, and the `comparisons` module has tools for comparing images to each other via metrics including NRMSE (Equation 5.28) and normalized cross-correlation. In designing an imaging or data analysis script in `ehtim`, a user might alternate between functions from many of these modules and classes. For instance, as in CLEAN, imaging with `ehtim` is often aided by performing alternate rounds of imaging and self-calibration. One strategy for combining these functions into an imaging script (used for imaging M87 with EHT data in Paper IV) is presented in Appendix E.

6.2 The `Image` class

An instance of the `Image` class contains information about an astronomical image and various methods for manipulating the image and generating synthetic data. Many `Image` objects only contain total intensity information, but the class can contain information about all four polarizations, either in a Stokes (I, Q, U, V) or circular cross-hand (RR, LL, RL, LR) basis. The polarization basis is stored in the `Image.polrep` attribute, which can be set to either `'stokes'` or `'circ'`.

6.2.1 Image representation and metadata

`Image` instances can be rectangular with dimensions $m \times n$.[5] The number of pixels n in the y-dimension (north-south) is stored as `Image.ydim`, and similarly the number of pixels m in the x, east-west direction is stored in `Image.xdim`. The pixels are all square with a linear size $\Delta\theta$ available as the attribute `Image.psize`, stored in radians.

The primary image data (typically total intensity, or Stokes I, and always in units of Jy/pixel) is stored as a one-dimensional `numpy` array (Walt et al., 2011) $\mathbf{I}$ of length $M = m \times n$ accessed with the `Image.imvec` attribute. However, any polarization image can be designated as the primary image and accessed via `Image.imvec`; the label that determines the "primary" polarization that is used by default is `Image.pol_prim`. The various polarization images can all be accessed through their own one-dimensional arrays (e.g., `Image.qvec` or `Image.rrvec`). `Image` stores the data for all polarizations in a hidden dictionary; attributes like `Image.imvec` use the Python `@property` feature to access the underlying data structure in a way that keeps the data consistent regardless of which polarimetric representation is used at a given moment.

While images are specified by 1D arrays of length M, these arrays actually represent *continuous* images. Following Bouman et al. (2016), each array of pixel intensities is taken to represent a continuous function formed by convolving a comb of Dirac delta functions with a pixel "pulse" function (this function is also available as an attribute, `Image.pulse`). When imaging, this continuous image representation multiplies the visibilities of the discrete Dirac comb array by a taper given by the Fourier transform of the pulse function. The pulse function removes spurious high-frequency structure introduced by the regular pixel spacing from the Fourier transform. The default pulse function is a triangular pulse (`trianglePulse2D`) with width $2\Delta\theta$, where Δ is the image pixel spac-

[5] In practice, most images are square with $m = n$. Many functions throughout `ehtim` have not yet been fully debugged with rectangular images!

ing. Other pulses can also be used, including a rectangular pulse (`rectPulse2D`), a circular Gaussian (`GaussPulse2D`), and a cubic spline (`cubicPulse2D`).

Every image also carries with it the metadata necessary to simulate interferometric data from the image. These include the source right ascension (`Image.ra`) and declination (`Image.dec`), the image frequency in Hz (`Image.rf`), the observing MJD and epoch in hours (`Image.mjd` and `Image.time`), and the source name (`Image.source`).

6.2.2 Loading or creating an image

Most instances of an `Image` class are loaded from a data file, but occasionally (e.g., when constructing a Gaussian source model) they are created using a variety of builder functions implemented in the `Image` class itself. `ehtim` can read in images from standard FITS format with the function `ehtim.image.load_fits`. The data can be read in with any pulse function and in either polarimeteric representation.[6] The `Image.save_fits` method exports an `Image` as a FITS file. Images can also be saved and loaded in a custom ASCII format using the `Image.save_txt` method and the `ehtim.image.load_txt` function.

Building an image from scratch starts with defining an empty image frame. This task can be accomplished for a square image with the `ehtim.image.make_empty` function, which takes the number of pixels, image field of view, and source metadata as arguments. Once an empty image has been created, a variety of methods can be used to add structures to the image. For instance, `Image.add_flat` adds a flat background brightness; `Image.add_tophat` adds a constant-brightness disk of a specified radius, and `Image.add_gauss` adds an elliptical Gaussian to an image. A polarization field can also be created – either constant (`Image.add_const_pol`) or random with a specified position

[6] In addition to loading in images from the primary HDU of a FITS file, the user can choose to read in a list of CLEAN components and deposit them on the grid without convolving them with a beam by using the `aipscc=True` flag. Experience has shown this flag to be very important.

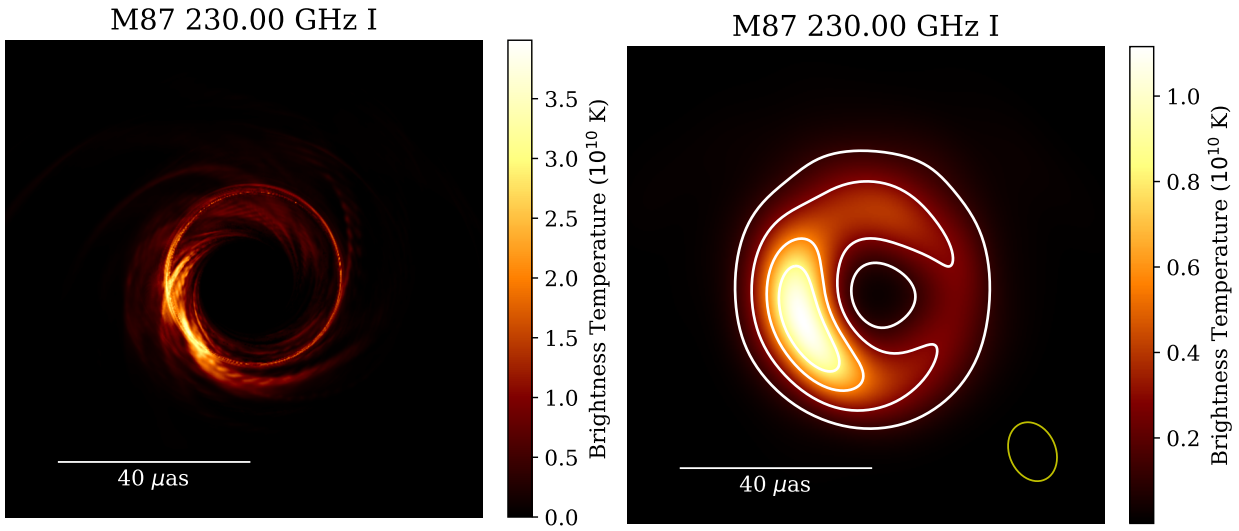


Figure 6.2: Examples of display functions in the `Image` class. (Left) Output of the `Image.display` method after loading a 230 GHz raytraced image from the M87 simulation `R17` described in Chapter 4 (Chael et al., 2019b). (Right) Output of the `Image.contour` method on the same image after convolution with a Gaussian kernel with the same orientation as the fitted CLEAN beam to the EHT 2017 (u, v) coverage on M87, scaled down in size by a factor of 1/2 (Paper III).

angle correlation length (`Image.add_random_pol`).

6.2.3 Image methods

After they are loaded or created, `Images` have a variety of useful methods for obtaining image parameters, image manipulation, and display. For instance, methods like `Image.total_flux`, `Image.lin_polfrac`, and `Image.centroid` provide the image total flux density, integrated linear polarization fraction, and centroid position, respectively.

A critical operation in the imaging process is blurring an image by convolving it with a Gaussian kernel. This step can be done in `ehtim` with the `Image.blur_gauss` method. This method convolves the underlying image with an elliptical Gaussian beam defined by three parameters; the major axis FWHM, the minor axis FWHM, and the position angle east of north (all in radians). In addition to blurring, other operations implemented in `ehtim` for manipulating an image include shifting by a certain number of pixels (`Image.shift`), rotating the image counterclockwise (`Image.rotate`), and

thresholding the image to remove low-brightness noise (`Image.threshold`). The class also contains two methods for interpolating an image onto a differently sized grid. The `regrid_image` method uses linear or cubic interpolation to resample the pixels directly onto a new grid of arbitrary field of view and resolution; in contrast, the `Image.resample_square` method directly samples the continuous image representation defined by the pulse function to a new resolution with the same field of view, so that the underlying continuous image is preserved.

The `Image` class also has two methods to display the image using `matplotlib` (Hunter, 2007) and export it to a pdf file (Figure 6.2). The `Image.display` method can be used to display any individual polarization image or a vector field of polarization ticks; this method allows for enough flexibility to design near-publication-quality plots with only a few command-line options (e.g. the choice of color map, the choice of linear, logarithmic, or gamma intensity scale, the choice of color bar units, the choice of axis ticks or a scale bar, and the choice of how to plot the interferometer beam). Similarly, the `Image.contour` method allows the user to easily create high-quality contour plots of the image.

6.2.4 Generating synthetic data

The most important method of the `Image` class is `Image.observe`, which takes an `Array` object (Section 6.3) and produces synthetic visibilities in an `Obsdata` object. The many options in the call to `Image.observe` allow the user to decide exactly how the data will be sampled, and how they are corrupted with phase, amplitude, or polarimetric miscalibration terms. For instance, the following command generates simulated data from an `Image` instance `im` using an `Array` instance `eht` (observing with a bandwidth `bw`= 2 GHz, integration time `tint`= 60 s, and with new scans every `tadv`=600 s) over a full 24-hour rotation of the Earth:

```
obs_sim = im.observe(eht, tint, tadv, 0, 24, bw,
                     add_th_noise=True, ampcal=False, phasecal=False
                     ttype='nfft')
```

This specific observation includes both thermal noise (`add_th_noise=True`) and additional station-based amplitude and phase errors (`ampcal=False, phasecal=False`).

`Image.observe` calls the `Array.obsdata` object to generate (u, v) points from an `Array` operating at a frequency `Image.rf` as the Earth rotates between a UTC time `tstart` and `tstop`. The points are sampled every `tadv` seconds in this interval.[7] Thermal noise sampled from a Gaussian distribution can be added to the visibilities (if `add_th_noise=True`); the standard deviation of the Gaussian thermal noise is determined by Equation 5.8 using the bandwidth `bw` and integration time `tint` specified. Once thermal noise is added, various flags (including `phasecal` and `ampcal`) determine how the simulated data are mis-calibrated. This miscalibration can take the form of station-based errors (Equation 5.3), but it can also use a full Jones formalism including polarimetric leakage and parallactic angle rotation by setting `jones=True`. Appendix C discusses the full Jones formalism implemented in `ehtim` in more detail.

The simplest way `ehtim` can generate visibilities is with a direct time Fourier transform, or DTFT. For N observed visibilities, the corresponding visibilities $\mathbf{V}$ of the image vector $\mathbf{I}$ are $\mathbf{V} = \mathbf{F}\mathbf{I}$, where $\mathbf{F}$ is an $N \times M$ matrix with entries

$$F_{ij} = \mathrm{e}^{-2\pi \mathrm{i}(u_i x_j + v_i y_j)}. \tag{6.1}$$

As in the integral van Cittert-Zernike theorem (Equation 5.1), (x_j, y_j) are the angular coordinates (in radians) of the jth pixel, and (u_i, v_i) are the angular frequencies of the ith visibility measurement. For the continuous image representations used in `ehtim`, the F_{ij} entries are also multiplied by the

[7] In addition to the continuous sampling in `Image.observe`, the method `Image.observe_vex` can be used to generate data on a realistic VLBI schedule, read from a .vex file.

Fourier transform of the pulse function convolution kernel, `Image.pulse`, sampled at (u_i, v_i).

While the direct-time Fourier transform (DTFT) represented by Equation 6.1 is often the fastest way to compute trial visibilities for sparse arrays observing with narrow fields of view, for large images or large numbers of visibilities, the DTFT is slow and prohibitively expensive in terms of computer memory. In this regime, `ehtim` uses the Nonequispaced Fast Fourier transform C library (NFFT: Keiner et al. 2009) accessed via the Python `pyNFFT` wrapper.[8] NFFT takes the Fast Fourier Transform (FFT) of the trial image and interpolates the result to the irregularly sampled (u, v) points, producing a highly accurate approximation of the exact DTFT. The choice of Fourier transform scheme is specified in `Image.observe` with the `ttype` flag.

6.3 The `Array` class

The `Array` class is relatively simple. It contains a `numpy` record array (`Array.tarr`) that lists the telescopes in an interferometric array and their important properties. For each telescope, this data table stores its label, x, y, z position in Cartesian geocentric coordinates, SEFDs both for right and left circular polarizations, complex d-terms, and field rotation parameters (see Appendix C).

The primary method of the `Array` class is `Array.obsdata`. This method takes in a source sky position (RA and dec), a telescope radio frequency and bandwidth, and an observing cadence between a start and stop time in order to generate (u, v) points from earth rotation and produce an empty `Obsdata` object. A call to `Array.obsdata` is the first call made in `Image.observe`.

An `Array` can be loaded from a text file with the `ehtim.array.load_txt` function. In addition to VLBI arrays that rotate with the Earth, Palumbo et al. (2019) added the ability to track potential orbiting stations in an `Array` object using orbital two-line elements (TLEs).

[8] https://pypi.Python.org/pypi/pyNFFT

6.4 The `Obsdata` class

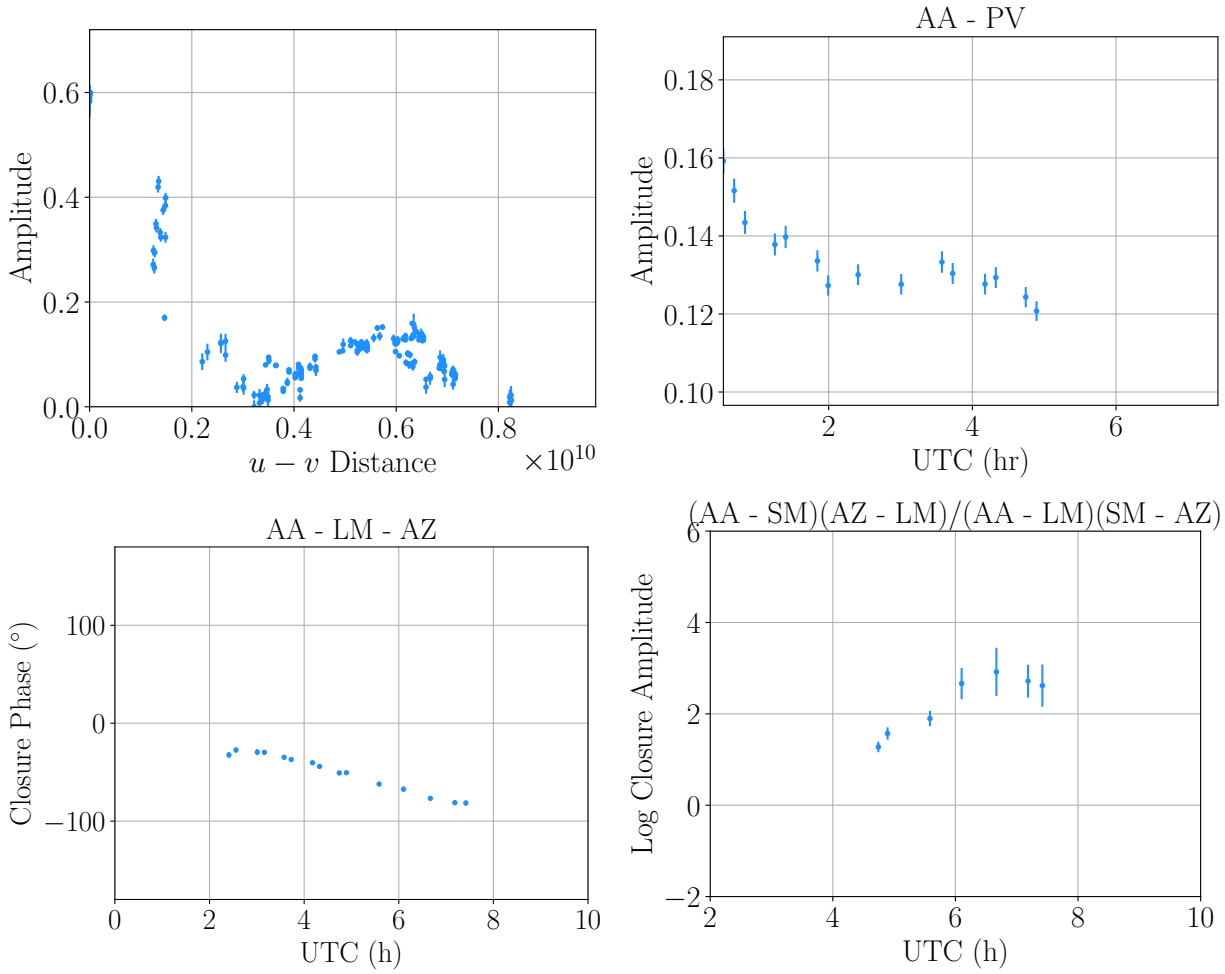


Figure 6.3: Examples of the output of the built-in plotting methods in the `Obsdata` class applied to scan-averaged data from the April 11, 2017 EHT observations of M87 (Paper III). (Top left) visibility amplitudes plotted versus baseline distance, $(u^2 + v^2)^{1/2}$. (Top right) Visibility amplitudes vs UTC time on the ALMA-PV baseline (AA-PV using the EHT 2017 station codes). (Bottom left) closure phase plotted vs UT time on the ALMA-LMT-SMT (AA-LM-AZ) triangle. (Bottom right) log closure amplitudes vs UT time on the ALMA-SMA-SMT-LMT (AA-SM-AZ-LM) quadrangle.

Every `Obsdata` instance stores a data table in the form of a `numpy` record array `Obsdata.data` that contains information on the observing and integration times, telescopes, (u, v) coordinates, complex visibilities, and thermal noise on all four polarizations across a full interferometric observation. Each `Obsdata` object also contains a telescope array (`Obsdata.tarr`) identical to that found in the `Array`. In addition to keeping the telescope data (positions, SEFDs) accessible, the order of telescopes in

this array is critical in determining how closure quantities are generated (Section 6.4.3).

In addition to the underlying data, an `Obsdata` object contains certain metadata specifying the observation, including the source name (`Obsdata.source`), sky coordinates (`Obsdata.ra, Obsdata.dec`), observing day and start/stop times in hours (`Obsdata.mjd, Obsdata.tstart, Obsdata.tstop`), and the observing frequency and bandwidth (`Obsdata.rf, Obsdata.bw`). Like an `Image`, an `Obsdata` instance can either be represented in a Stokes or circular polarization basis, determined by the `Obsdata.polrep` attribute. An `Obsdata` with a given `polrep` can be switched to the other representation with the `Obsdata.switch_polrep` method. Regardless of the representation, any polarimetric data product can be extracted from an `Obsdata` instance with the `Obsdata.unpack` method (Section 6.4.2).

6.4.1 Loading or creating an observation

Typically, `Obsdata` instances are created from a call to `Image.observe` (Section 6.2.4) or by loading data from a .uvfits file, one standard format for interferometric data exchange.[9] Data can be read from uvfits files using `ehtim.obsdata.load_uvfits` in either a circular polarization (`polrep='circ'`) or Stokes (`polrep = 'stokes'`) basis. Currently, the `eht-imaging` library only supports frequency-averaged data. When loading in data from a multi-channel .uvfits file, the data may be averaged in frequency, or a particular channel and IF can be selected. In addition to standard .uvfits files, data can also be saved and loaded in a custom ASCII format using `Obsdata.save_txt` and `ehtim.obsdata.load_txt`. Total intensity visibility data can also be saved or loaded from the .oifits optical interferometry standard via the `Obsdata.save_oifits` method and `ehtim.obsdata.load_oifits` function.

[9]The .uvfits standard is defined in AIPS memo 117: ftp://ftp.aoc.nrao.edu/pub/software/aips/TEXT/PUBL/AIPSMEM117.PS

6.4.2 Accessing and editing visibility data

In general, data should not be accessed directly from the `Obsdata.data` table. Instead, they should be accessed through class methods. The most general of these methods is `Obsdata.unpack`. The `unpack` method can extract one or several data products (e.g., integration time or Stokes I visibility) from the data table; it can also derive new data products from the base data types (e.g., the source elevation angle at a site $\theta_{\rm el}$ or the fractional polarization on a baseline $\check{m}$). The full list of the 72 quantities that can be accessed with `Obsdata.unpack` is available in `ehtim` as the constant `ehtim.FIELDS`.

Using `Obsdata.unpack` allows the `Obsdata` object to store only the minimal set of data it needs; for instance, `Obsdata.unpack` intelligently computes conjugate visibilities on the baseline $(-u, -v)$ if requested, though only data on one baseline ordering is stored in the data table. More generally, using `Obsdata.unpack` to read the data ensures that the underlying data are not corrupted, overwritten, or made inconsistent when accessing the table. In addition to `Obsdata.unpack`, which extracts the specified data on all baselines throughout an observation, the method `Obsdata.unpack_bl` allows the user to isolate a certain data product on an individual baseline over time.

In the reverse scenario, where the user wishes to edit, flag, or scale data in an existing object, it is always safer to create a new object than to edit the data in `Obsdata.data` in place. There are a variety of helper functions already defined in `ehtim` for data editing. For flagging, these methods include `Obsdata.flag_uvdist` (to flag short or long baselines) , `Obsdata.flag_sites` (to flag individual stations) `Obsdata.flag_UTrange` (to flag data in a given time range), `Obsdata.flag_anomalous` to flag large outliers, and `Obsdata.flag_low_snr` (to flag low SNR data points). Other methods for manipulating data that are relevant to imaging include `Obsdata.taper` and `Obsdata.inverse_taper`, which multiply or divide the visibilities by a Fourier transformed Gaussian circular kernel to degrade

or enhance the resolution of the data, and `Obsdata.rescale_zbl`, which rescales short baselines that may contain contributions from extended structure invisible to longer baselines so that they are consistent with the compact flux density. Rescaling short baselines is particularly important for the EHT, where intra-site baselines (ALMA-APEX; JCMT-SMA) are many orders of magnitude shorter than the VLBI baselines.

The fundamental data products in an `Obsdata` object are the complex visibilities. Typically, closure quantities are generated from the underlying visibilities on-the-fly when needed in imaging and analysis. This method ensures that all closure data are consistent with the underlying observation. However, it is occasionally useful to manipulate the closure quantities independently. For instance, it may be useful to manually flag certain closure triangles, or average closure phases on a different timescale than visibility amplitudes in imaging. For this purpose, an `Obsdata` object can carry attributes, including `Obsdata.cphase, Obsdata.camp, Obsdata.logcamp`, which contain pre-computed tables of closure quantities. These attributes can be set manually or generated with the class methods `Obsdata.add_cphase, Obsdata.add_camp` with specified averaging and flagging schemes. If pre-computed closure quantities are present in an `Obsdata` instance, plotting and imaging routines elsewhere in the code will automatically use these tables instead of recomputing the closure quantities from the visibility data.

6.4.3 Generating closure quantities

.

A core functionality of the `Obsdata` class is the ability to generate sets of closure quantities (closure phases, closure amplitudes, log closure amplitudes) for use in imaging and model fitting to the data. These quantities are typically generated on-the-fly prior to imaging to ensure they are consistent with the underlying data in the `Obsdata.data` table.

The `Obsdata.c_phases` method returns a record array of the chosen closure phases on all specified triangles. It can compute closure phases for any polarization (with the `vtype` argument), and via the `count` argument, this method will either return a maximal or minimal set of closure phases (see Section 5.1.2). The maximal set includes closure phases computed on every triangle formed by telescopes observing at a given time stamp. In contrast, the minimal set includes only a subset of $\frac{(N_s-1)(N_s-2)}{2}$ triangles at any instant needed to reconstruct the closure phase on any other triangle. The algorithm used by `ehtim` for choosing this minimal set prioritizes closure triangles that include a reference station (TMS; Blackburn et al. 2019); this reference station in `ehtim` is always the first listed station in the `Obsdata.tarr` array. Typically, this reference is set to the station with the highest SNR; however, it can be randomized with the `Obsdata.reorder_tarr_random` method.

Similarly, the `Obsdata.c_amplitudes` method returns a record array of closure amplitudes or log closure amplitudes (depending on the `ctype` parameter) for any polarization. Like `Obsdata.c_phases`, `Obsdata.c_amplitudes` can return either an array of all closure amplitudes on all quadrangles, or it can return a minimal set needed to recover all the closure amplitude data. The minimal set algorithm for closure amplitudes is also from Blackburn et al. (2019); similarly to the closure phase minimal set algorithm, it relies on the reference station specified by the telescope array.

6.4.4 Plotting

Finally, several built-in methods of `Obsdata` provide easy interfaces to plot data products against each other and across time (Figure 6.3 shows several example plots generated with these methods). `Obstata.plotall` is a method that allows for any two derived data products in the observation (from the list in `ehtim.FIELDS`) to be plotted against each other in a scatter plot. `Obsdata.plot_bl` plots individual quantities on a baseline as a function of time, and `Obsdata.plot_cphase` and `Obsdata.plot_camp` each produce quick plots of the closure phases and amplitudes (from any polariza-

tion) on a specified quadrangle or triangle.

All the `Obsdata` plotting methods use underlying `matplotlib` methods and return an `Axes` object; these plots can be easily customized to produce professional quality plots or combined with other `Axes` in a larger figure. The `ehtim.plotting.comp_plots` module makes it possible to quickly overplot information from multiple `Obsdata` objects or to compute corresponding information from an `Image`. For example, with the `plotall_compare` function of the `comp_plots` module, the user can quickly overplot visibility amplitudes (or phases, polarimetric ratios, etc.) from a given observation on the same axis as the corresponding noise-free quantities computed from several trial image reconstructions.

6.5 The `Imager` class

The `Imager` class in `ehtim` defines an algorithm for RML imaging through a data set, initial image, and an objective function with given regularizer and data term weights (Equation 5.14). The data is passed in with an `Obsdata` object, and the initial image and prior image (used in MEM regularization) are `Image` objects.[10] The data and regularizer terms used in the RML algorithm and their associated hyperparameter weights are set with Python dictionaries.

For instance, suppose an `ehtim` user wanted to produce an image from a data set in the `Obsdata` object `obs`. They set up an initial Gaussian image with the appropriate size and field of view using the `Image` construction methods described in Section 6.2; this Gaussian will also be used as the prior image for MEM. Then they define the objective function (Equation 5.14); they use visibility amplitude and closure phase χ^2 terms, with data term weights $\alpha_{\text{amp}} = 1$, $\alpha_{\text{cl phase}} = 10$. The regularizer terms will be a standard maximum entropy regularizer (called 'simple' in `ehtim`) with a weight $\beta_{\text{MEM}} = 1$ and a centroid constraint $\beta_{\text{centroid}} = 100$ to keep the source in the frame. The

[10] Critically, these `Images` and the `Obsdata` must have the same source metadata.

total flux density of the source is 1 Jy. They would set up the imager with the following command:

```
imgr = ehtim.Imager(obs, init_gauss, prior_im=init_gauss, flux=1.0,
                    data_term={'amp':1, 'cphase':10},
                    reg_term={'simple':1, 'cm':100}
                   )
```

Once initialized, the data can be accessed and changed with the attribute `Obsdata.obs_next`, and the initial image/prior can be changed with the `Obsdata.init_next` and `Obsdata.prior_next` attributes. Similarly, the regularizer terms and data terms used can be accessed and changed through the attributes `Imager.dat_term_next` and `Imager.reg_term_next`. All of the data terms in Section 5.2.2 and all of the regularizers in Section 5.2.3 are included in the `Imager` class, as well as several additional regularizers (e.g., several implementations of a constraint on image compactness).

Several additional parameters are important in specifying an imaging routine. These include the maximum number of iterations allowed in the imager (`Imager.maxit_next`), and the convergence criterion on the change in the objective function and objective function gradients that will terminate the imager before the maximum number of iterations is reached (`Imager.stop_next`). The type of Fourier transform algorithm used can also be specified (`Imager.ttype_next`, see Section 6.5.1), and the user can also choose whether or not to perform the imager on a log-space image in order to avoid negative pixels by setting `Imager.transform_next='log'` (Section 6.5.3).

6.5.1 Running the `Imager`

Once the data, initial image, χ^2 terms, regularizers, and additional parameters are defined in an `Imager` object, the optimization can be run with the command `Imager.make_image_I`. The `Imager` can run in the background and provide diagnostics only when it converges or reaches the maximum number of iterations; alternatively, the `Imager` can update the user and display the current image and χ^2 values in real time with every step toward a minimum of the objective function.

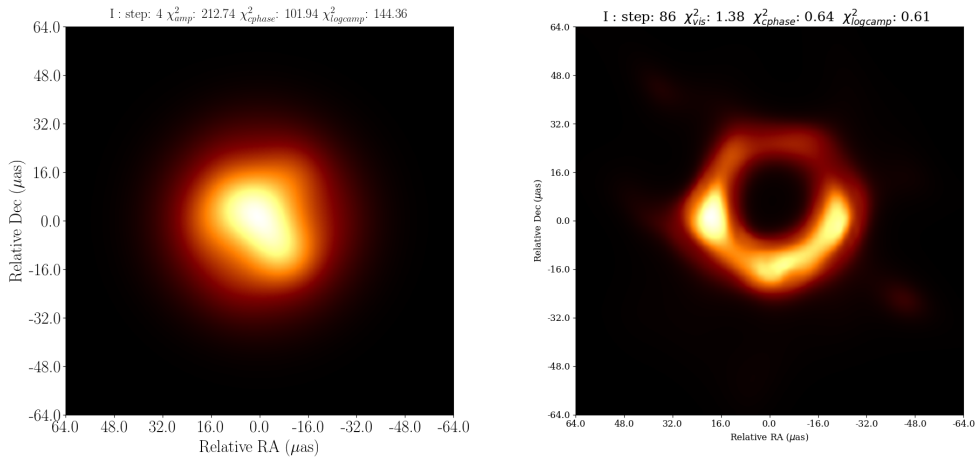


Figure 6.4: Example snapshots of the interactive display while running `Imager.make_image_I` with the flag `show_updates=True` on EHT 2017 observations of M87 (Paper III,Paper IV). (Left) Only four steps into the imaging process, the image still resembles the circular Gaussian initial image and the χ^2 terms on all data products are high. (Right) later in the imaging process, the image nearly fits the data and has converged to a ring-like structure.

This feedback is enabled by setting `show_updates = True` in the call to `Imager.make_image_I`. Figure 6.4 shows two snapshots of an example `Imager` in progress after calling `Imager.make_image_I` with `show_updates=True`. While this real time updating of the `Imager`'s progress slows down the imaging process substantially, it is useful for providing feedback on how a certain set of parameter choices affects the image reconstruction. For instance, it is particularly helpful to diagnose if the imager is sent into a false local minima early in the optimization. The resulting image from a call to `Imager.make_image_I` is accessed by calling the method `Imager.out_last`.

The `Imager` stores precomputed data and gradient information and can run an imaging algorithm quickly multiple times in a row using different initial images (e.g., iteratively blurring and re-imaging from the blurred output image for convergence). When data terms, regularizer weights, or even the underlying data are changed through the attributes described above, the `Imager` notes this change in the history and recomputes the pre-computed data products and gradient terms. The `Imager` object contains a full history of all imaging steps run through this interface. Even if a user is

imaging experimentally in the command line, the exact sequence of commands can be recovered from the built-in history to produce an automated imaging script. As much as possible in a given imaging session, users should update existing `Imager` objects rather than re-initializing the `Imager` at different stages in the imaging process.

In addition to the total intensity imaging discussed in this thesis, the `Imager` class also implements individual imaging of the four Stokes parameters, and the simultaneous imaging algorithm for the polarization ratio and polarimetric position angle described in Chael et al. (2016). It also implements the simultaneous imaging + scattering deconvolution algorithm of Johnson (2016).

On small data sets or on reconstructions with a small field of view, direct-time Fourier transforms are sufficient to compute the data terms in Section 5.2.2. For larger data sets, the DTFT matrix F_{ij} (Equation 6.1) becomes prohibitively large to store in memory and prohibitively slow at extracting visibilities from the trial image at each step. In this regime, the `Imager` can use NFFTs instead by setting the `ttype='nfft'` flag in initializing the `Imager` instance, or by setting `Imager.ttype_next='nfft'` after the instance has been created.

To find a minimum of the objective function, the `Imager.make_image_I` method uses the Limited-Memory BFGS algorithm (Byrd et al., 1995) as implemented in the `scipy` package (Jones et al., 2001; Oliphant, 2007). The L-BFGS algorithm is a quasi-Newton gradient descent method. While L-BFGS can compute the gradient at each step numerically, it is much more efficient to specify the gradient analytically. All of the data and regularizer terms implemented in `ehtim` have analytic forms of the gradient defined (Section 6.5.2 and Appendix D). These analytic gradients are used by default, but the `Imager` can be told to use numerical approximations to the gradient with the `grads=False` flag in the call to `Imager.make_image_I`.

6.5.2 Data term gradients

When using gradient descent algorithms to minimize the objective function (Equation 5.14), providing an analytic expression for the gradient of the objective function with respect to the image pixel values greatly increases the speed of the algorithm by bypassing the expensive step of estimating gradients numerically. When using a DTFT, the number of computations to evaluate the gradient of a χ^2 term numerically via finite differences is roughly $\mathcal{O}(M^2 \times N)$, where M is the total number of image pixels and N is the number of measurements. When using an FFT or NFFT, the scaling is roughly $\mathcal{O}(M \times (M \log M + N))$. In contrast, when using analytic gradients, the corresponding scalings for DTFT and FFT are $\mathcal{O}(M \times N)$ and $\mathcal{O}(M \log M + N)$, respectively.

The gradient of the simplest χ^2 term, using complex visibilities (Equation 5.16), is

$$\frac{\partial}{\partial I_i}\chi^2_{\text{vis}} = -\frac{1}{N}\sum_j \text{Re}\left[F^\dagger_{ij}\left(\frac{V_j - \hat{V}_j}{\sigma_j^2}\right)\right]. \tag{6.2}$$

Equation 6.2 indicates that the gradient is proportional to the adjoint Fourier transform of the data residuals. While more complicated (and more difficult to derive), the gradients for the other χ^2 terms given in Section 5.2.2 take similar forms. The gradients for all the total intensity imaging data terms implemented in `ehtim` are presented in Appendix D. The gradients of the regularizer terms are easier to derive, and are also presented in Appendix D for completeness.

6.5.3 Image transformation

When imaging, the gradient of the objective function may drive image pixels to negative values, breaking the assumption of image positivity that underlies VLBI imaging. To avoid this issue, the

`Imager` class can ensure a positive brightness in each pixel by performing a change of variables

$$I_i = \exp \xi_i \ \ , \ \ -\infty < \xi_i < \infty. \tag{6.3}$$

When imaging in the log intensity domain ξ_i, the gradients in Section D must be multiplied by $\exp \xi_i$. The final returned image (`Imager.out_last`) is always stored in the original units (Jy/pixel). The `Imager` can also be run without transforming the pixel intensities to ensure positivity; this is done by setting `transform=False` when initializing the imager, or `Imager.transform_next=False` in later stages.

Even without any other regularization, this positivity constraint acts as an effective regularizer that removes many local minima in the χ^2 landscape the minimizer might otherwise fall into. For instance, the "dirty image", or direct inverse Fourier Transform of the measured visibilities with all unmeasured visibilities set to zero, will typically have negative pixel values. Figure 7 of Paper IV shows that the positivity constraint, combined with a restricted FOV, can regularize the reconstruction enough to produce accurate images of geometric models from synthetic EHT data.

6.6 Other routines

In addition to the primary image classes described above, the flexible nature of the `eht-imaging` framework has encouraged the development of many additional software tools for analyzing and manipulating EHT and other interferometric data. This section discusses only two of these additional functions; a full description of all classes and functions in the `eht-imaging` code is available in the documentation.[11]

[11]The `ehtim` documentation is dynamically updated from docstrings embedded in the code using Sphinx: http://www.sphinx-doc.org/en/master/.

6.6.1 Self-calibration

Typically in VLBI imaging, rounds of imaging the data (either by running CLEAN or minimizing the objective function in RML) are alternated with rounds of self-calibration where the station gains and phases are adjusted to match the observed complex visibilities to the solved-for structure as best as possible. While closure-only imaging allows for images that fit the most robust data products to be generated without any calibration (Chapter 5), including at least one self-calibration step can substantially improve the results of this method (Section 5.5). Furthermore, the station amplitude gain terms are often at least weakly constrained a priori; in this case, it is often preferable to include visibility amplitudes (with a weak hyperparameter) in the imaging objective function along with closure amplitudes, before refining the gain solution with self-calibration.

The `eht-imaging` library implements self-calibration via the `Caltable` class and the `ehtim.selfcal` function. For example, to self-calibrate the data in `obs` to the final image produced by the `Imager` `imgr`, an `ehtim` user could run the commands:

```
caltab = ehtim.selfcal(obs_static, imgr.out_last(),
                       gain_tol=0.1, caltable=True)
obs_sc = caltab.applycal(obs_static, interp='nearest')
```

A `Caltable` instance (`caltab` in the above code snippet) stores an array of two complex station gains (one per polarization) as a function of time for each telescope in an array. `Caltable` objects are associated with a particular `Obsdata` instance. If the `Caltable` and `Obsdata` metadata match, the gain solution in the `Caltable` can be applied to the `Obsdata` with the `Caltable.applycal` method. This method interpolates the complex gains stored in the `Caltable`, multiplies the `Obsdata` visibilities according to the convention in Equation 5.3 (e.g. $V_{12} \rightarrow G_1 e^{i\phi_1} G_2 e^{-i\phi_2} V_{12}$), and returns the resulting calibrated data set as a new `Obsdata` instance.

The `ehtim.selfcal` function produces a `Caltable` by finding the set of complex gains $g_a = G_a e^{i\phi_a}$

that, when applied to a set of complex visibilities in an `Obsdata` object, produce calibrated visibilities that best match the model visibilities from the supplied `Image`. As is standard in interferometry, the gains can be derived with a specified solution interval such that all the data points within a given time window are forced to receive the same gain correction. The self-calibration algorithm finds the set of complex gains $\{g\}$ that minimizes the objective function:

$$J_g\left(\{g_a\}\right) = \chi^2_g\left(\{V\},\{\hat{V}\},\{g\}\right) + S_g\left(\{g\}\right). \tag{6.4}$$

The goodness-of-fit χ^2_g function for the gains is a function of the set of measured visibilities $\{V\}$, the set of model visibilities sampled on the same baselines $\{\hat{V}\}$, and the set of complex, time-variable station gains $\{g\}$. It is

$$\chi^2_g\left(\{V\},\{\hat{V}\},\{g\}\right) = \sum_i \frac{1}{\sigma_i^2}\left|\hat{V}_i - g_a(t)g_b^*(t)V_i\right|^2, \tag{6.5}$$

where the visibility V_i is on a baseline formed by stations a and b, and $\hat{V}_i$ is the model image visibility, both at time t. The `selfcal` function also imposes a prior term on the gain so as to prevent the algorithm from over-fitting the data with gains that are too different from unity (i.e., perfect a priori calibration). The prior term S_g is

$$S_g\left(\{g\}\right) = \sum_{\{g\}} \log\frac{|g|^2}{T^2}, \tag{6.6}$$

where T (set by the `gain_tol` flag) is a fractional tolerance on the allowed gain deviation from unity, and the sum is over the full set of all station gains at all times t. For instance, setting $T = 0.1$ would push the algorithm to find gain solutions that are within 10% of 1.

In addition to the complex self-calibration described above, the `selfcal` function can also derive

solutions for only the phase corrections ϕ_a or only the amplitude corrections G_a. The user can set which type of self-calibration is run by changing the `method` flag to `'both'`,`'amp'` or `'phase'` when running the `selfcal` function. As in CLEAN imaging, self-calibrating the visibility phases before calibrating the amplitudes is usually a good choice to help the imaging process converge. Finally, depending on whether the `Caltable` flag is set to `True` or `False`, the `selfcal` function will either return the resulting table of station gains in a `Caltable` object, or it will go ahead and apply those gains directly to the data and return a new `Obsdata` object.

In addition to the `ehtim.selfcal` function, `ehtim` also contains the function `ehtim.netcal` for network calibration of the visibility without any source image. The network calibration algorithm (Johnson et al., 2015) enforces the constraint that visibilities to co-located sites be identical and that the visibility amplitude on trivial baselines between co-located sites be equal to a specified total flux density. The network calibration routine in `ehtim` was the final step of the data calibration pipeline for the 2017 EHT data described in Paper III.

6.6.2 Diagnostic summary plots

After a final image is produced from an `ehtim` imaging script, it is helpful to gather a set of diagnostic statistics and plots on how well that image fits the data used in the imaging process. The `ehtimaging` library contains many functions to compute these (e.g., visibility χ^2 terms, self-calibrated gains, plots of closure phase vs time), but it can be difficult to manually check each statistic every time an imaging algorithm is run.

The `ehtim.imgsum` function computes a full set of diagnostic statistics and plots to assess the fit quality of an image when compared to data. For instance, to run the function on a final `Image` instance `im_out`, a final self-calibrated `Obsdata` instance `obs_sc`, and an initial dataset `obs_original`: the user would call

```
ehtim.imgsum(im_out, obs_sc, obs_original, 'imgsum.pdf')
```

This function will produce a .pdf file called `imgsum.pdf` that contains the final image and a full set of diagnostic information. Figure 6.5 shows an example of the first page of one of these summary .pdfs, generated from running the sample imaging script in Appendix E on scan-averaged 2017 EHT data. The first page of a summary file contains a plot of the image, the image blurred to the nominal array resolution, χ^2 statistics on all the quantities described in Section 5.2.2, and closure phase and log closure amplitude goodness-of-fit statistics broken down on individual triangles and quadrangles in the minimal sets used in imaging. An example second page is shown in Figure 6.6; this page shows a plot of the visibility amplitudes from the model and self-calibrated data, the visibility amplitude χ^2 statistics broken down by baseline, and the derived self-calibration solution. The remaining plots in an image summary sheet show the amplitudes, closure phases, and log closure amplitudes on individual baselines, triangles, and quadrangles. Investigating these plots and the individual statistics in the summary .pdf file can provide quantitative information about how well an image fits the data; it can also guide the user to which amplitudes, closure amplitudes, or closure phases are most critical in driving the fit or which of these data products is not being well fit by the current imaging approach.

6.7 Summary

The `eht-imaging` library is a new, open-source, comprehensive software suite that allows for the easy inspection, analysis, calibration, and imaging of both real and simulated interferometric data. This chapter presents the fundamentals of the main classes and methods developed in `ehtim` over the last two years. The development of `ehtim` proceeded in parallel with advances in techniques for imaging EHT data sets. While the code was originally developed for polarimetric imaging (Chael et al., 2016), it was reorganized into a modular framework to enable imagers to easily swap out

data terms and regularizers during the development of the closure-only imaging method (Chael et al., 2018b). As the 2017 EHT data were prepared and imaged over 2017 and 2018 (Paper III; Paper IV), new ideas for dealing with this particularly challenging data set expanded the code's functionality and reliability even further.

Despite the length of the full `ehtim` code (68,554 lines of Python code at the time of this writing) and its many capabilities, the organization of `ehtim`'s of main functionality into four main classes – `Image`, `Array`, `Obsdata`, and `Imager`– means that learning the basic concepts behind the code is relatively straightforward. The modular nature of the code makes it easy to build off of previous class methods and example scripts and create new functions, classes, and methods to attack new challenges. Thanks to the active development of many contributors (10 unique contributors responsible for $>$ 1500 total `git` commits as of 2019), this chapter has only scratched the surface of the tools available in `ehtim`. An incomplete list of other capabilities of `ehtim` developed by EHT collaborators includes: imaging polarimetric data and solving for polarimetric gain and leakage terms, simulating data from movies of time-variable sources, reconstructing time-variable data into movies with special regularizing functions to control how images change over time, sophisticated averaging and filtering of visibility and closure data, new regularizers to constrain image large-scale structure without short (u, v) spacings, and simulating data on baselines to antennas orbiting the Earth.

Over the past two years, `ehtim` has evolved into a critical component of the analysis pipeline of the full global EHT collaboration. It is actively used and developed by collaboration members around the world. Results from `ehtim` have contributed to 18 peer-reviewed publications over the last two years, including the first images of a black hole shadow from the EHT (Paper III; Paper IV; Paper V; Paper VI). Appendix E presents a full imaging script very similar to the one used in Paper IV to produce images of the black hole shadow in M87 from EHT data. This script has

been adapted with only minor changes from the script used to produce the fiducial M87 images in Paper IV. Despite containing multiple rounds of imaging and self-calibration, the script is relatively short and relies only on `ehtim` concepts presented in this chapter. The imaging and characterization of the M87 black hole shadow using `ehtim` is described in Chapter 7.

Summary Sheet for M87 on MJD 57854

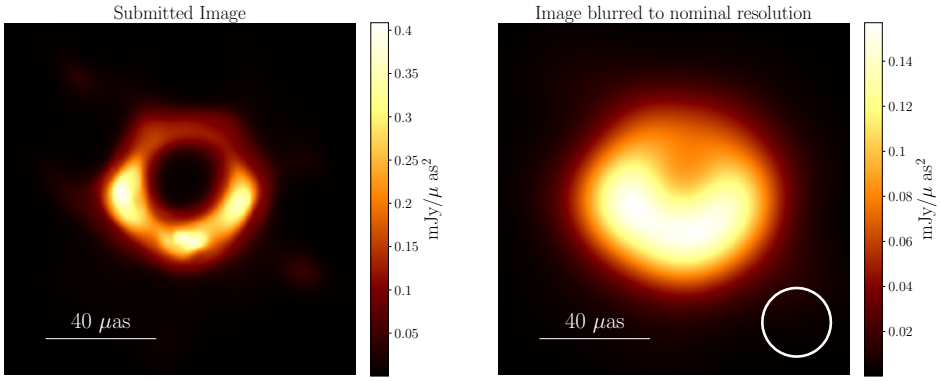


Image statistics

Source:	M87	χ^2_{vis}	0.61	(0.61)
MJD:	57854	χ^2_{amp}	0.47	(0.47)
FREQ:	227 GHz	χ^2_{cphase}	0.65	(0.65)
FOV:	128.0 μas	$\chi^2_{logcamp}$	0.61	(0.61)
FLUX:	0.51 Jy	χ^2_{camp}	0.68	(0.68)

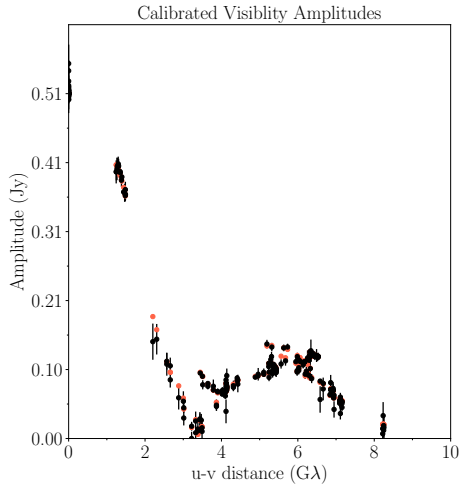
Closure phase statistics

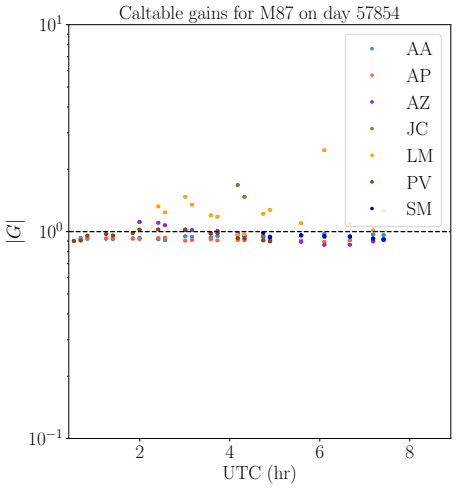
Triangle	N_{tri}	χ^2_{tri}/N_{tri}	χ^2_{tri}/N_{tot}
AA-LM-JC	9	1.1	0.144
AA-AZ-SM	7	1.3	0.125
AA-LM-AZ	15	0.5	0.097
AA-AZ-JC	9	0.7	0.084
AA-PV-LM	8	0.6	0.072
AA-PV-AZ	9	0.4	0.056
AA-PV-JC	4	0.8	0.048
AA-PV-SM	2	0.4	0.012
AA-LM-SM	7	0.1	0.007

Log Closure amplitude statistics

Quadrangle	N_{quad}	χ^2_{quad}/N_{quad}	χ^2_{quad}/N_{tot}
AA-AP-LM-AZ	14	0.9	0.106
AA-AP-AZ-LM	14	0.8	0.102
AA-AP-AZ-PV	9	0.8	0.067
AA-JC-AZ-AP	8	0.8	0.056
AA-JC-AP-SM	6	1.0	0.054
AA-JC-SM-LM	7	0.8	0.053
AA-AP-PV-AZ	9	0.5	0.043
AA-AZ-PV-LM	8	0.6	0.041
AA-SM-AP-LM	6	0.7	0.038
AA-JC-AZ-SM	7	0.5	0.034
AA-SM-AZ-LM	7	0.5	0.034
AA-AP-PV-LM	8	0.5	0.033
AA-AZ-LM-PV	8	0.4	0.031
AA-SM-AZ-AP	6	0.6	0.030
AA-JC-AP-PV	4	0.7	0.024
AA-JC-LM-AZ	9	0.3	0.022
AA-SM-LM-AZ	7	0.3	0.017
AA-JC-PV-LM	4	0.4	0.013
AA-SM-AP-PV	2	0.1	0.001
AA-JC-SM-PV	2	0.1	0.001
AA-SM-PV-LM	2	0.0	0.000

Figure 6.5: Example of the first page of an image summary sheet automatically generated by `ehtim.imgsum` on a reconstruction of 2017 EHT data. The first page of the summary sheet displays the image and the image blurred to the array nominal resolution, presents statistics of the image fit to the data, and breaks down the image-data goodness of fit on individual closure triangles and quadrangles.





Visibility amplitude statistics

Baseline	N_{amp}	χ^2_{amp}/N_{amp}	χ^2_{amp}/N_{total}
AP-JC	8	1.5	0.055
LM-AP	14	0.8	0.053
SM-JC	7	1.4	0.046
AZ-JC	9	1.1	0.045
AZ-AP	15	0.6	0.040
LM-SM	7	0.9	0.029
PV-AZ	9	0.6	0.025
LM-JC	9	0.5	0.021
LM-AZ	15	0.3	0.019
AA-LM	15	0.3	0.018
AA-JC	9	0.4	0.017
AP-SM	6	0.6	0.016
AA-SM	7	0.5	0.015
PV-SM	2	1.4	0.013
PV-LM	8	0.3	0.012
AA-AP	20	0.1	0.012
AA-AZ	16	0.1	0.009
PV-AP	14	0.1	0.008
AZ-SM	7	0.2	0.008
AA-PV	15	0.1	0.008
PV-JC	4	0.2	0.004

Station gain statistics

Site	Median Gain	Percent diff.
LM	1.22	22
AP	0.91	9
AA	0.94	6
JC	0.94	6
SM	0.95	5
PV	0.96	4
AZ	0.96	4

Figure 6.6: Example of the second page of an image summary sheet automatically generated by the `ehtim.imgsum` function on a reconstruction of 2017 EHT data. The second page of the summary sheet displays self-calibrated visibility amplitudes (black) and the model visibility amplitudes (orange) in the upper left, the self-calibrated amplitude gain solutions as a function of UTC time for each station in the upper right, and statistics of the fit to the visibility amplitudes on each baseline.

Page intentionally left blank

Text in this chapter was previously published in *ApJL* 875 (2019), L4 (The Event Horizon Telescope Collaboration et al.)

7

Measuring the supermassive black hole shadow in M87

In April 2017, the Event Horizon Telescope (EHT) observed the supermassive black hole in M87 at 230 GHz with a full array at five geographic sites for the first time. These sites were ALMA and APEX in Chile, the SMA and JCMT on Maunakea in Hawai'i, the SMT in Arizona, the IRAM 30-m telescope in Spain, and the LMT in Mexico. This historic observation was only made possible by years of technical development, outfitting the planet-spanning array with the advanced digital backends, masers, and data recorders that allowed every EHT site to record data at 230 GHz with 2 GHz of bandwidth (see Paper II). After extensive effort in correlating, reducing, and calibrating

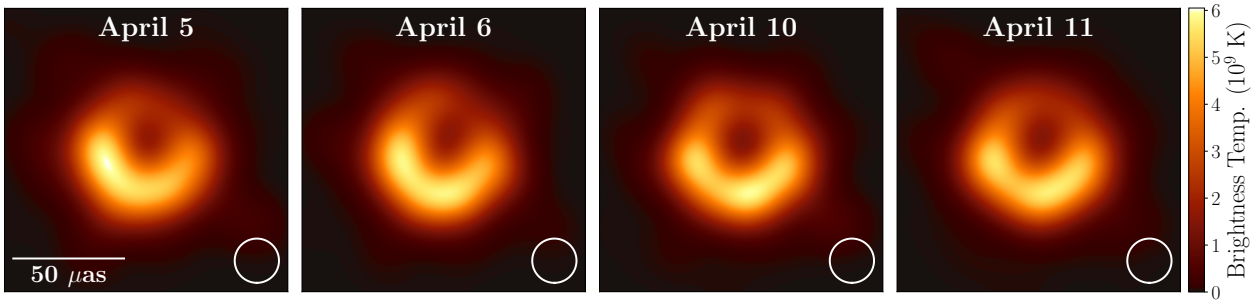


Figure 7.1: Maximally conservative images of M87 from the EHT on each of the four observing days in 2017 (Figure 15 of Paper IV). The indicated beam is 20 μas, which corresponds to the CLEAN beam used in the DIFMAP reconstructions of these data.

the data from all seven telescopes (Paper III), the first images of the black hole shadow in M87 were generated from these data (Paper IV).

After summarizing the imaging process – from initial blind imaging to automated surveys of the imaging parameter space – and discussing the main features of the resulting images in Section 7.1, Sections 7.2–7.6 present the full feature extraction and ring-fitting analysis presented in Section 9, Appendix G, and Appendix I of Paper IV. These sections present the method implemented in `eht-imaging` for identifying rings in images of M87 and identifying their features (Section 7.2), the method's results on tests with synthetic data reconstructions (Section 7.3), and the ring diameters and other parameters measured from the fiducial M87 images from three independent imaging pipelines (Section 7.4). Section 7.5 notes some potential biases in the measured parameters – notably the ring diameter and width – that arise from the finite resolution of the EHT reconstructions. Section 7.6 presents unwrapped radial and angular profiles produced from these ring fits and uses these profiles to discuss features in the ring-like structure observed in M87 at 230 GHz. Finally, Section 7.7 presents a simplified version of the analysis developed in Paper VI to measure the mass of the central supermassive black hole in M87 from the 2017 EHT images produced with the `eht-imaging` library.

7.1 EHT images of M87

The EHT's first 230 GHz images of M87 were subjected to an exhaustive process of validation and cross-checks designed to remove human bias from the imaging process as much as possible and to determine which features in the final images are robust to imaging choices. Paper IV used a two-stage imaging process. In the first stage, four teams of experts imaged the EHT data in isolation; prevented from exchanging information with each other for over a month, these teams worked to

understand the data and produce reliable images using a variety of imaging methods and software (see Figure 4 of Paper IV). Once these first, blind images were shown to be consistent, Paper IV proceeded to systematically explore the imaging choices available in three software pipelines – (1) the traditional CLEAN algorithm implemented in `DIFMAP` (Shepherd, 1997), (2) the `eht-imaging` library described in Chapter 6 (Chael et al., 2016, 2018b), and (3) the `SMILI` sparse imaging library (Akiyama et al., 2017a,b). After designing minimal template scripts for each method, an exhaustive survey of the parameter search available to each imaging method was conducted. This parameter search generated images both from M87 data and synthetic data sets for $\approx 10^4$ parameter combinations. Every parameter combination was then ranked based on its performance in reconstructing a suite of synthetic datasets with the same (u, v) coverage as the April 2017 M87 observations, but with distinct underlying emission structures. Only combinations of imaging parameters that could reconstruct the four distinct synthetic sources in the training set and distinguish between them (e.g., between a filled disk and a jet) were included in the final "Top Set" of best imaging parameters for each method. From each Top Set, the *best* performing parameter combination was denoted as the "fiducial parameters."

Figure 7.2 shows the results of reconstructing images from the EHT 2017 M87 data with the fiducial parameters of all three imaging pipelines. All twelve images in Figure 7.2 (three methods, four days) are consistent in producing an asymmetric ring of $\approx 40\mu$as diameter, brighter in the south than the north. However, the images produced by the different methods are not identical. For instance, the `DIFMAP` images are restored with a $20\,\mu$as FWHM Gaussian beam, limiting their resolution when compared with the RML methods (`eht-imaging` and `SMILI`). While the structure in the images from the two RML pipelines is in general consistent, `eht-imaging` and `SMILI` produce images with different apparent azimuthal structure than `DIFMAP`.

It is not surprising that, given the sparse nature of even the best EHT data, different imaging

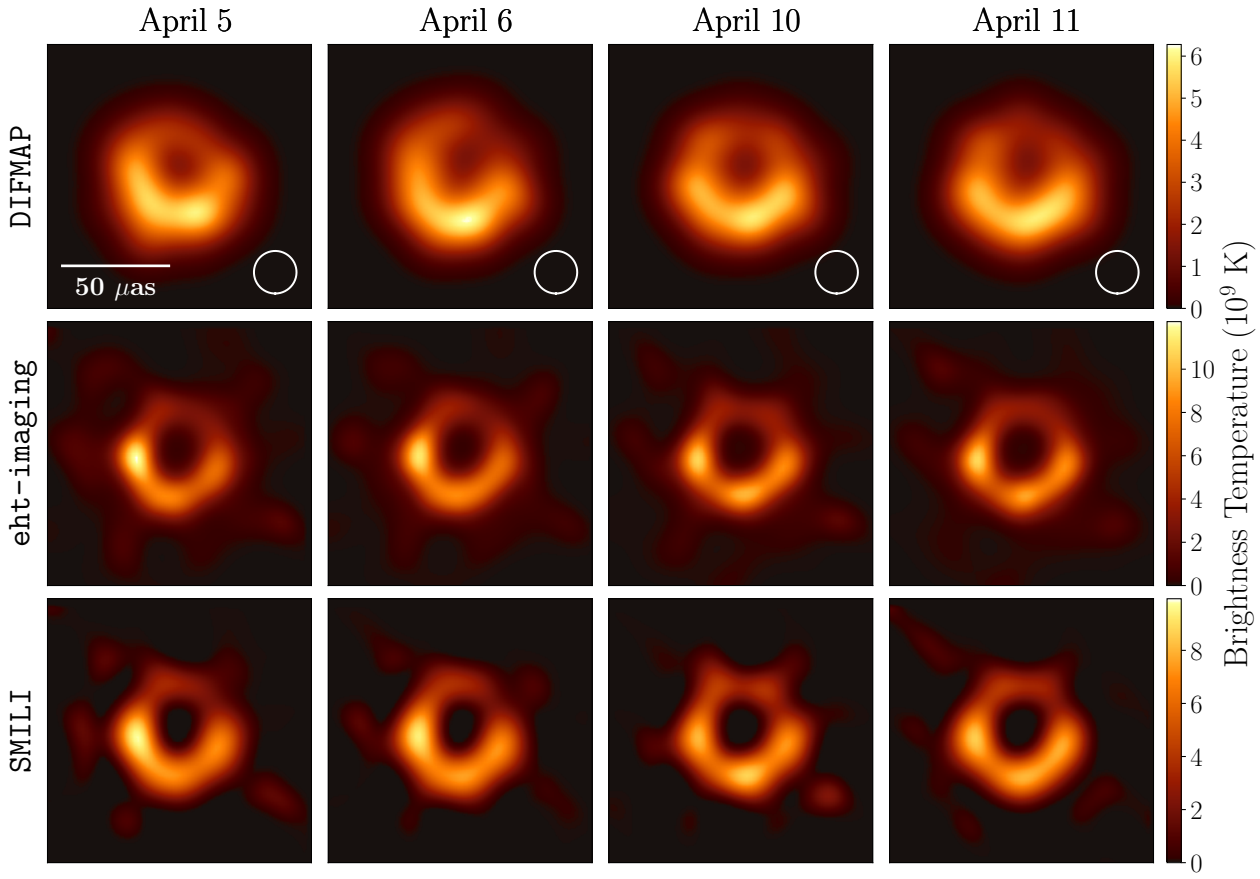


Figure 7.2: Fiducial images of M87 on all four observed days from each of the three imaging pipelines (Figure 11 of Paper IV). CLEAN images (from `DIFMAP`) are shown after convolution with a 20 μas beam; `eht-imaging` and `SMILI` results have no restoring beam applied.

methods would produce different structure, particularly on scales smaller than the ≈20 μas CLEAN beam. The final, maximally conservative images in Figure 7.1 were generated by averaging together the fiducial images from all three imaging pipelines after restoring each to a common resolution set by the beam-convolved `DIFMAP` reconstruction.

On all four days of observation, these EHT images of M87's central engine show a characteristic ring of diameter $\sim$ 40 μas, consistent with the shadow of a supermassive black hole with mass $M = 6.5 \pm 0.7 \times 10^9\, M_\odot$ at a distance of 16.8 Mpc (Paper VI). The ring is brighter in the south on all four days, consistent with a clockwise sense of fluid rotation near the horizon (Chapter 3). Assuming the jet is powered by the Blandford & Znajek (1977) mechanism, this asymmetry indicates that the black hole spin is oriented away from Earth (Paper V). In the six days between the first EHT

observation on April 5 and the last on April 11, the structure in the core of M87 evolved, changing the observed closure phases and visibility amplitudes (Paper III). The final images in Figure 7.1 show a *counterclockwise* shift in brightness along the ring between the first and last observations. While intriguing, this variation should not be interpreted as reflecting any underlying physical motion in this direction. In the simulations of the M87 accretion flow presented in Chapter 3, for instance, the accretion flow and jet rotate clockwise, but similar counterclockwise apparent motion can arise on ∼week timescales from variations in the brightness of certain image features (see Figure 3.12).

The maximally conservative images in Figure 7.1 provide a visual indication of what structures in the images are guaranteed to be robust to choices in the imaging procedure. To further assess the consistency of the fiducial images (Figure 7.2) with each other and determine which specific image features are most reliable, it is helpful to measure certain parameters that characterize asymmetric rings directly from the images and compare the results across days and imaging method. These parameters are the ring diameter d, the width w, the orientation angle η, the asymmetry A, and the fractional central brightness f_C.

7.2 Ring parameter definitions

The ring parameter estimation code described in this Section and used in the analysis of EHT images in Paper IV and Paper VI is implemented as the `REx` (Ring Extractor) module in `eht-imaging`.[1] The motivating idea behind this method is that, to robustly identify a ring in a given reconstructed image, one should search for the central point from which radial profiles are peaked at a similar distance. That is, from a candidate ring center position (x, y), `REx` samples a linearly interpolated image equally in azimuthal angle θ between 0 and 360° and in radius r between 0

[1]This section was originally published as Section 9.1 of Paper IV

and 50 μas to obtain a transformed image: $I(r, \theta; x, y)$. Then, for each radial profile at fixed θ, the code identifies the distance $r_{\rm pk}$ at which the angular profile assumes its peak brightness. The ring radius at a given position, $\bar{r}_{\rm pk}(x, y)$ is defined as the mean of these peak distances:

$$r_{\rm pk}(\theta; x, y) = \text{argmax}_r \left[I(r, \theta; x, y) \right], \tag{7.1}$$

$$\bar{r}_{\rm pk}(x, y) = \langle r_{\rm pk}(\theta; x, y) \rangle_{\theta \in [0, 2\pi]}.$$

To estimate an associated uncertainty in $\bar{r}_{\rm pk}$, `REx` uses the standard deviation $\sigma_{\bar{r}}(x, y)$ of the $r_{\rm pk}(\theta; x, y)$ values.

To find the ring center (x_0, y_0), the code searches over (x, y) and identifies the position that minimizes the normalized radial peak dispersion:

$$(x_0, y_0) = \text{argmin} \left[\frac{\sigma_{\bar{r}}(x, y)}{\bar{r}_{\rm pk}(x, y)} \right]_{(x,y)}. \tag{7.2}$$

The measured diameter d is then twice $\bar{r}_{\rm pk}$ measured from the identified center (x_0, y_0),

$$d = 2\bar{r}_{\rm pk}(x_0, y_0), \tag{7.3}$$

and the associated uncertainty σ_d in the diameter is

$$\sigma_d = 2\sigma_{\bar{r}}(x_0, y_0). \tag{7.4}$$

Although `REx` was designed to specifically search for circular features, note that σ_d/d can be interpreted as a measure of the circularity of the identified ring-like feature (Paper VI).

To avoid spurious detections when searching for the center location (Equation 7.2), `REx` first

blurs the image with a $2\,\mu$as FWHM Gaussian (approximately the pixel size of the original M87 reconstructions) and thresholds the search image below 5% of the peak brightness. The range of allowed diameters is also restricted to 10–100 μas. However, the original (unconvolved and unthresholded) image is used for all subsequent analysis.

The ring width w is determined by measuring the FWHM of each radial slice at constant θ and taking the mean. To avoid bias in the measurement from the resampled image $I(r,\theta)$ having a nonzero floor value outside the ring, it is important to subtract the value $I_{\rm floor} = \langle I(r_{\rm max} = 50\,\mu{\rm as}, \theta)\rangle_\theta$ from each radial profile before computing the FWHM:

$$w = \langle {\rm FWHM}\,[I(r,\theta) - I_{\rm floor}]\rangle\,. \tag{7.5}$$

The uncertainty σ_w is computed from the standard deviation of the set of FWHMs. Note that the measured width is dependent upon both the intrinsic width of the source and the finite resolution of the array. Thus, w should be viewed only as an upper limit on the intrinsic ring width, and it is biased upward by the application of a restoring beam (e.g., for `DIFMAP` reconstructions). This bias is explored further in Section 7.5.

The ring orientation angle η (measured east of north) is computed from the individual angular profiles $I(r,\theta)$ at fixed r. To compute η, `REx` finds the argument of the first angular mode ($m = 1$) of the angular profile at each radius and then takes the overall orientation angle η as the circular mean of these angles over the ring width; from $r_{\rm in} = (d - w)/2$ to $r_{\rm out} = (d + w)/2$. That is,

$$\eta = \left\langle {\rm Arg}\left[\int_0^{2\pi} I(\theta) e^{{\rm i}\theta}{\rm d}\theta\right]\right\rangle_{r\in[r_{\rm in}, r_{\rm out}]}\,. \tag{7.6}$$

Similarly, the associated uncertainty σ_η is the circular standard deviation of the angle measurements across the ring width, for $r \in [r_{\rm in}, r_{\rm out}]$.

The degree of azimuthal asymmetry in a ring is determined from the normalized amplitude of the first angular mode for radii between $r_{\rm in}$ and $r_{\rm out}$. That is,

$$A = \left\langle \frac{\left| \int_0^{2\pi} I(\theta) e^{i\theta} d\theta \right|}{\int_0^{2\pi} I(\theta) d\theta} \right\rangle_{r \in [r_{\rm in}, r_{\rm out}]} . \tag{7.7}$$

The associated uncertainty σ_A is the standard deviation of the asymmetry at each $r \in [r_{\rm in}, r_{\rm out}]$. The asymmetry A takes values in the range from 0 to 1, with 0 corresponding to perfect azimuthal symmetry and 1 corresponding to a delta function concentrating all of the flux density at a single orientation angle. For instance, the simple crescent models used as synthetic sources in Paper IV have angular brightness profiles $I(\theta) \propto 1 + 2A\cos(\theta - \eta)$, where $0 \leq A \leq 1/2$.

The last parameter is the ring fractional central brightness (or inverse contrast ratio) $f_{\rm C}$. This quantity is computed as the ratio of the mean brightness interior to the ring to the mean brightness around the ring. To define the interior brightness, the code averages over a disk of radius 5 μas centered in the ring center. That is,

$$f_{\rm C} = \frac{\langle I(r,\theta) \rangle_{\theta, r \in [0, 5\mu{\rm as}]}}{\langle I(d/2, \theta) \rangle_{\theta}} . \tag{7.8}$$

This statistic has an extremely large scatter across the Top Sets, primarily because the interior brightness can become arbitrarily low in RML imaging. Thus, the imaging methods explored in Paper IV only securely identify an upper limit on $f_{\rm C}$.

7.3 Tests with synthetic data reconstructions

Paper IV tested the analysis methods described in Section 7.2 on image reconstructions of synthetic data from a crescent model with asymmetry parameter $A = 0.23$ oriented at $\eta = 150°$, and on a

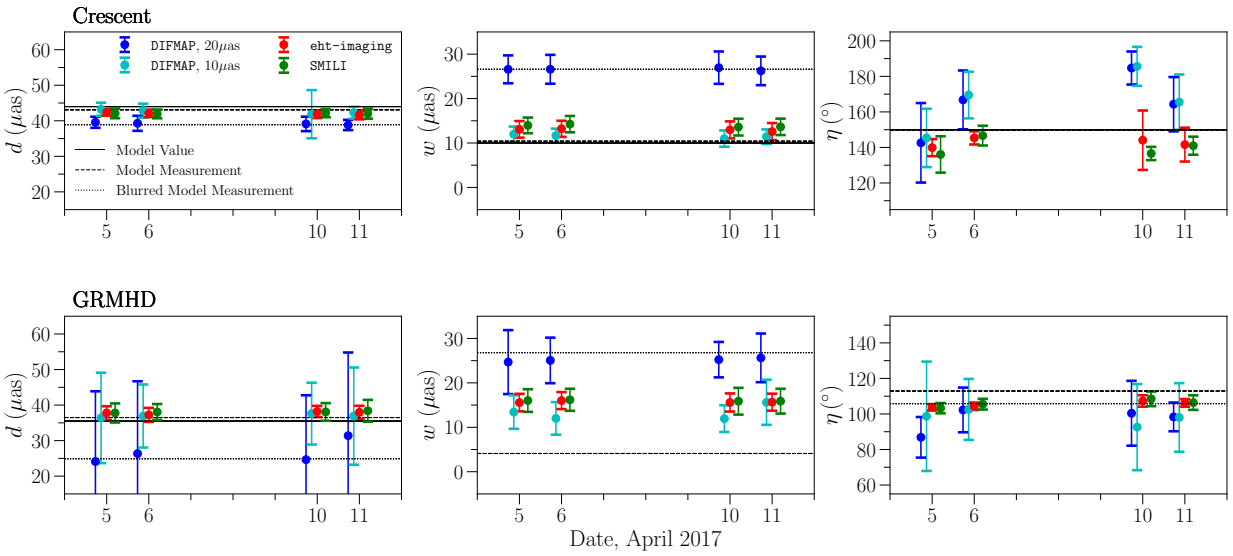


Figure 7.3: Measurements of ring features on fiducial reconstructions of a crescent model (top row) and a GRMHD simulation snapshot (bottom row) from images reconstructed with the fiducial parameters for three imaging pipelines (Figure 24 of Paper IV). From left to right, panels show the measured ring diameter d, width w, and orientation angle η. The DIFMAP results are shown for images restored with both a 10 μas (cyan) and a 20 μas (blue) Gaussian beam. The eht-imaging (red) and SMILI (green) results are shown for the unblurred images. Solid lines indicate the ground truth values of the three quantities in the crescent model, and the photon ring diameter of the GRMHD simulation. Dashed lines indicate the measured values from the ground truth images, and dotted lines indicate the measured values from the ground truth images convolved with a 20 μas FWHM circular Gaussian beam. The error bars are computed as the quadrature sum of the measurement uncertainty from the fiducial images (Section 7.2) and the median absolute deviation of the parameter estimates across the Top Set.

GRMHD simulation image from Paper V.[2] For both synthetic sources, the ring features defined in Section 7.2 were measured from every image reconstructed with all the parameter combinations in the M87 Top Sets (see Table 3 of Paper IV).

The results of this analysis for the diameter (d), width (w), and orientation angle (η) are displayed in Figure 7.3. The points shown correspond to the measured quantities from the fiducial reconstructions. These measurements have two distinct sources of error: the intrinsic measurement uncertainty on each quantity from a single image, and the uncertainty in the quantity from varying the imaging parameters across the Top Set. These two sources of error are combined in the error bars in Figure 7.3 by taking the quadrature sum of the measurement uncertainty from the fiducial image and the median absolute deviation of the parameter estimates across the Top Set.

[2]This section was originally published as Section 9.2 of Paper IV.

In both simulated data tests, the ring diameter d is the most accurately recovered quantity; however, the diameter measurement is correlated with the ring width and is biased downward by several microarcseconds when the image is blurred (see Section 7.5). In the simulated crescent, the diameter of the unblurred model is $44\,\mu$as; the value measured using the same approach on the ground truth image is pushed down to $43\,\mu$as because of the Gaussian convolution. Taking the median across all days, the diameters extracted from the fiducial `DIFMAP` crescent model reconstructions have a median value of $40 \pm 2\,\mu$as when restored with a $20\,\mu$as FWHM beam. When they are instead restored with a smaller $10\,\mu$as beam, the `DIFMAP` results become more accurate (compared to the model values), with a measured diameter of $43 \pm 2\,\mu$as. The RML crescent model reconstructions measure a median diameter of $42 \pm 1\,\mu$as.

The GRMHD image (Paper V) has a lensed photon ring with a diameter of $9.79M/D \approx 35.5\,\mu$as;[3] `REx` recovers a value of 36.5 μas from the ground truth image, indicating that even with a perfect image reconstruction, fine-scale substructure and extended flux in an image can bias the ring diameter measurement away from the photon ring value (see Paper VI). The ring diameter measured from the `eht-imaging` and `SMILI` GRMHD fiducial reconstructions (Figure 7.3) has a median value of $38 \pm 2\,\mu$as across all four observed days. However, for this data set, the `DIFMAP` results are strongly dependent upon the chosen restoring beam. The `DIFMAP` reconstructions restored with a $20\,\mu$as beam lack prominent rings, leading to a poor recovery of the ring diameter ($26 \pm 20\,\mu$as). When blurred with a smaller 10 μas beam, the `DIFMAP` results align with the RML methods but have larger uncertainties due to larger scatter across the Top Set ($37 \pm 11\,\mu$as).

As shown in Figure 7.3, the measured orientation angles from the crescent model reconstructions using the RML imaging pipelines have a median value of $141° \pm 5°$, moderately discrepant from the true value of 150°. The values from the `DIFMAP` images are even more discrepant and are

[3]The simulation used an assumed black hole mass $M = 6.2 \times 10^9\,\mathrm{M}_\odot$ and dimensionless spin $a_* = 0.9375$, observed at a distance $D = 16.9\,\mathrm{Mpc}$ and 17° inclination (Paper V).

unstable among days. The GRMHD image has no defined a priori orientation angle; the measured orientation angles from all pipelines are more stable than for the simple crescent model, and they are consistent with the value measured from the blurred simulation image (105°). The results of both the GRMHD and crescent model reconstructions indicate that this procedure may underestimate uncertainties on η.

Both the crescent and GRMHD models have intrinsic widths that are much narrower than the resolution of the EHT. As expected from applying the 20 μas restoring beam, the DIFMAP reconstructions give a larger measurement of ring width than the RML reconstructions. When a 10 μas FWHM beam is used to restore the DIFMAP images, the measured widths align with the RML results. Nonetheless, all reconstructions have widths that are systematically biased upward from the true values, and the extracted ring widths from image domain fitting can at best be viewed as upper limits.

Similarly, the brightness depression contrast ratio $f_{\rm C}$ is highly sensitive to the image resolution and particular imaging choices. In particular, the Top Sets from the RML imaging pipelines have an extremely large scatter in $f_{\rm C}$. When blurred to the same 20 μas resolution, both RML and CLEAN methods give measurements of $f_{\rm C}$ that are consistently in the range 0.2–0.5 (Section 7.5). Consequently, the current reconstructions of horizon-scale structure in M87 can only determine an upper limit for $f_{\rm C}$.

For all three imaging pipelines, the scatter in ring diameters across the Top Set (typically $\lesssim 1\,\mu$as) is subdominant to the intrinsic measurement uncertainty estimated from a single image (typically 1-2 μas). Thus, choices made in the imaging process do not significantly affect the measured ring diameter from these models. In contrast, the other measured features have error budgets which are more evenly divided between intrinsic uncertainty in a single image and the scatter across the Top Set.

Table 7.1: Diameter d, width w, orientation angle η, asymmetry A, and floor-to-ring contrast ratio $f_{\rm C}$ measured from the fiducial M87 images for each day from each imaging pipeline (Table 7 of Paper IV).

	d (μas)	w (μas)	η ($^\circ$)	A	$f_{\rm C}$
`DIFMAP`					
Apr 5	37.2 ± 2.4	28.2 ± 2.9	163.8 ± 6.5	0.21 ± 0.03	5×10^{-1}
Apr 6	40.1 ± 7.4	28.6 ± 3.0	162.1 ± 9.7	0.24 ± 0.08	4×10^{-1}
Apr 10	40.2 ± 1.7	27.5 ± 3.1	175.8 ± 9.8	0.20 ± 0.04	4×10^{-1}
Apr 11	40.7 ± 2.6	29.0 ± 3.0	173.3 ± 4.8	0.23 ± 0.04	5×10^{-1}
`eht-imaging`					
Apr 5	39.3 ± 1.6	16.2 ± 2.0	148.3 ± 4.8	0.25 ± 0.02	8×10^{-2}
Apr 6	39.6 ± 1.8	16.2 ± 1.7	151.1 ± 8.6	0.24 ± 0.02	6×10^{-2}
Apr 10	40.7 ± 1.6	15.7 ± 2.0	171.2 ± 6.9	0.23 ± 0.03	4×10^{-2}
Apr 11	41.0 ± 1.4	15.5 ± 1.8	168.0 ± 6.9	0.20 ± 0.02	4×10^{-2}
`SMILI`					
Apr 5	40.5 ± 1.9	16.1 ± 2.1	154.2 ± 7.1	0.27 ± 0.03	7×10^{-5}
Apr 6	40.9 ± 2.4	16.1 ± 2.1	151.7 ± 8.2	0.25 ± 0.02	2×10^{-4}
Apr 10	42.0 ± 1.8	15.7 ± 2.4	170.6 ± 5.5	0.21 ± 0.03	4×10^{-6}
Apr 11	42.3 ± 1.6	15.6 ± 2.2	167.6 ± 2.8	0.22 ± 0.03	6×10^{-6}

7.4 Results for M87 EHT images

Table 7.1 lists the values of all ring parameters measured for the fiducial EHT 2017 M87 reconstructions for each day from the three parameter surveys.[4] As in the previous Section, for the fiducial images, the uncertainties are computed by adding the scatter in the measured quantities across the Top Sets in quadrature to the intrinsic measurement uncertainty.

Figure 7.4 plots the measured diameter, width, and orientation angle from each method over all four days. Across all days, the `DIFMAP` reconstructions recover a median diameter of $40 \pm 3\,\mu$as when restored with a $20\,\mu$as beam and $44 \pm 5\,\mu$as when restored with a $10\,\mu$as beam. The RML methods recover a median diameter of $41 \pm 2\,\mu$as. For each imaging pipeline, there is slight upward trend in the diameter over time; it increases by $\approx 2\,\mu$as from the first to the last day of the observing campaign. However, this trend is well within the estimated uncertainty of the diameter

[4]This section was originally published as Section 9.3 of Paper IV.

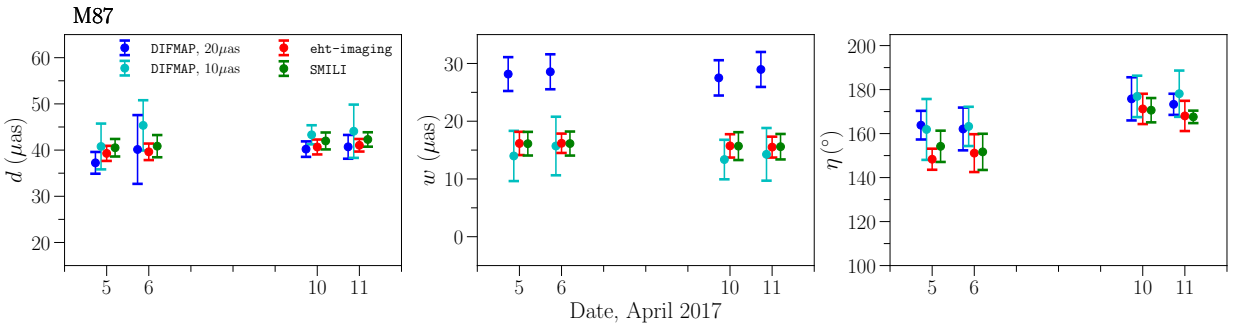


Figure 7.4: Measured ring properties on the fiducial images of M87 produced with all three imaging pipelines (Figure 25 of Paper IV). From left to right, panels show the measured ring diameter d, width w, and orientation angle η. The `DIFMAP` results are shown for images restored with both a 10 μas (cyan) and a 20 μas (blue) Gaussian beam. The `eht-imaging` (red) and `SMILI` (green) results are shown for the unblurred images. The three imaging pipelines produce consistent measurements of the ring diameter across all days. The measured orientation angles indicate a modest shift between April 5/6 and 10/11. The error bars are computed in the same manner as in Figure 7.3.

measurements. The measured ring diameters are consistent among all imaging pipelines and largely unchanged from the first April 11 images produced from early-release engineering data (Section 5 of Paper IV). However, the other parameters are less consistent from these first images to the final fiducial images selected by the parameter survey, indicating more sensitivity in these parameters to the data quality and the imaging method.

As in the synthetic data results presented in Figure 7.3, the beam-convolved `DIFMAP` reconstructions produce larger measured widths than the RML methods. The width measurements become consistent when the `DIFMAP` images are restored with a smaller 10 μas beam, but the ring width remains limited by the resolution of the EHT. Thus, from these imaging results, one can only firmly conclude that the ring width is $\leq 20\,\mu$as.

In the reconstructions from all three imaging methods, there is a counterclockwise trend in the orientation angle from April 5 to 11, consistent with the apparent shift in brightness along the ring. However, this $\approx 20°$ counterclockwise shift could be the result of spurious azimuthal structure introduced in the imaging process. The tests on synthetic data indicate that the method presented in this section may underestimate orientation angle uncertainties (see Figure 7.3). Even if this shift in angle is physical, it does *not* necessarily indicate continuous motion or a flow direction associated

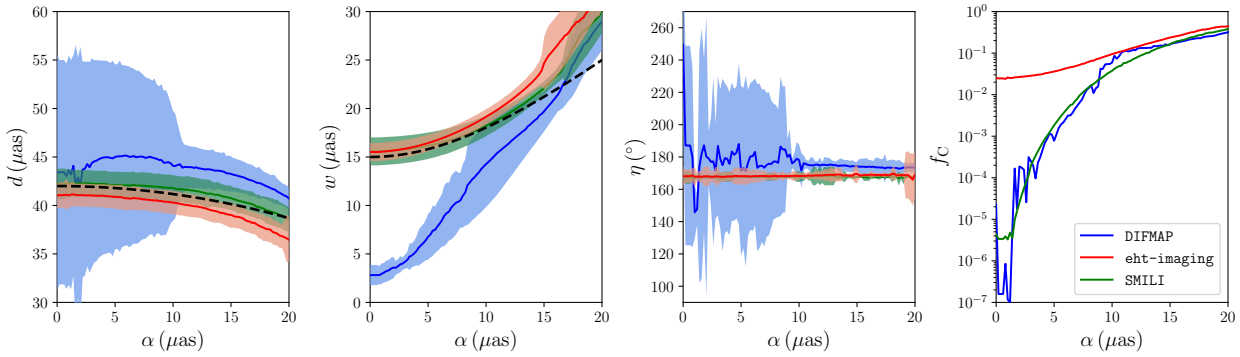


Figure 7.5: From left to right, the measured diameter d, width w, orientation angle η, and central brightness ratio f_C from the April 11 fiducial images blurred with circular Gaussian kernels of increasing FWHM α (Figure 35 of Paper IV). The solid lines indicate the measured value, and the shaded regions give the 1σ uncertainty as defined in Section 7.2 (these uncertainties do not include a contribution from scatter across the Top Set). The dashed line on the first panel shows the prediction of Equation 7.11 for the measured diameter d_meas as a function of the blur kernel assuming $d_\mathrm{true} = 42\,\mu$as. The dashed line in the second panel shows the FWHM of a $15\,\mu$as 1D Gaussian convolved with the kernel of FWHM α (Equation 7.12). Because the DIFMAP images are fundamentally composed of point sources, the measurements from these images become highly uncertain when $\alpha \lesssim 10\,\mu$as.

with the black hole accretion flow in M87, and it is opposite to the inferred rotation direction for the large-scale jet (Walker et al., 2018, see Chapter 3).

7.5 Finite resolution bias on ring parameters

In addition to uncertainties from thermal noise, systematic noise, and algorithmic imaging assumptions, estimated image properties are necessarily limited by the image resolution.[5] Image structure at scales finer than the diffraction limited resolution can bias properties such as the magnitude and location of the maximum image brightness.

As a simple example, a thin ring has a finite resolution bias in both width and diameter. If an infinitesimally thin ring of diameter d has the form:

$$I_{\delta\mathrm{ring}}(r,\theta;d) = \frac{1}{\pi d}\delta(r - d/2), \tag{7.9}$$

[5]This section was originally published as Appendix G of Paper IV.

a ring of finite thickness can be generated by convolving this image with a circular Gaussian kernel of FWHM α:

$$I_{\rm ring}(r,\theta;d,\alpha) = \frac{4\ln 2}{\pi^2\alpha^2 d}\int r'\,{\rm d}r'{\rm d}\theta'\delta\left(r'-d/2\right)\exp\left[-\frac{4\ln 2}{\alpha^2}\left(r^2+r'^2-2rr'\cos\theta'\right)\right]. \qquad (7.10)$$

Paper IV, Appendix G shows that in the limit $\alpha \ll d$, only keeping terms to leading order in r/d and α/d, the FWHM of the blurred ring is approximately equal to α. Furthermore, the angular diameter of peak brightness of the convolved ring is

$$d_{\rm true} \approx \frac{d_{\rm meas}}{1-\frac{1}{4\ln 2}\left(\frac{\alpha}{d_{\rm meas}}\right)^2}, \qquad (7.11)$$

where the approximation is accurate to leading order in $\alpha/d_{\rm meas}$. As a concrete example, an infinitesimally thin ring with true diameter $d = 44\,\mu$as and width $\alpha = 15\,\mu$as will have its measured diameter biased downward by $\approx 2\,\mu$as.

More generally, convolving a ring that has a Gaussian profile of intrinsic FWHM $w_{\rm true}$ with a circular Gaussian kernel with FWHM α gives an effective width

$$w_{\rm meas} \approx \sqrt{w_{\rm true}^2+\alpha^2}, \qquad (7.12)$$

and the diameter bias of a blurred finite width ring is still given by Equation 7.11 to first order.

While the simple Gaussian convolution assumed in this example only crudely approximates the effects of finite resolution on reconstructed images, it indicates that for images dominated by a thin ring, estimated diameters (widths) will be biased downward (upward) by the finite resolution of an image reconstruction. For the fiducial image measurements in Section 7.4, the diameter bias from this finite-resolution effect would be maximum in the case of an infinitesimal intrinsic ring, with

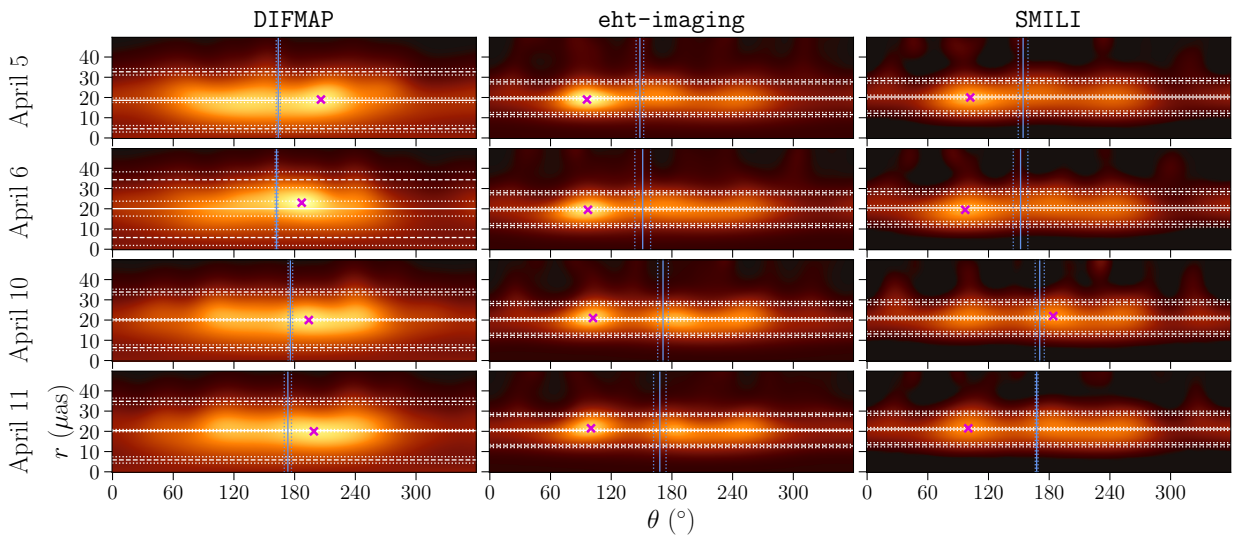


Figure 7.6: Unwrapped ring profiles of the fiducial images from April 5–11 (top to bottom) and for the three imaging pipelines (left to right) (Figure 27 of Paper IV). The estimated radius $d/2$ is shown with a horizontal line, with dotted lines denoting the associated uncertainty (Section 7.2). Horizontal dashed lines at $(d \pm w)/2$ show the measured ring width. Vertical blue lines give the orientation angle η and its uncertainty. The magenta cross marks the peak brightness in each reconstruction.

α then approximately corresponding to the measured width. Thus, the finite-resolution diameter bias is at most a few μas.

Figure 7.5 explores the dependence of several estimated ring parameters on the image angular resolution. Using the fiducial M87 images on April 11 from all three imaging pipelines, Figure 7.5 shows the ring parameters computed after convolving each image with a circular Gaussian kernel of FWHM α in the range $0 < \alpha < 20\,\mu$as. The DIFMAP images are fundamentally composed of point sources ("CLEAN" components). When these point sources are restored with a Gaussian beam of FWHM $\alpha \lesssim 10\,\mu$as, the CLEAN components still appear in the convolved image as individual point sources. As a result, the feature extraction methods of Section 7.4, which assume a smooth ring structure, have large uncertainties when extracting ring parameters from these images. This uncertainty is particularly apparent in the orientation angle measurement. The RML images, in contrast, have a finite width and smooth structure even at $\alpha = 0$. As a result, their parameters vary smoothly and have similar uncertainties at all values of α.

The leftmost panel of Figure 7.5 shows that the dependence of the measured diameter on α closely follows the bias predicted by Equation 7.11. The fiducial images have a scatter of approximately 1 μas across the different imaging methods but an uncertainty of $\approx$3–4 μas across the different values of α. Thus, when extracting physical parameters from the measured diameter, it is important to take into consideration the additional bias and uncertainty induced by this effect.

For small values of the restoring beam α, Figure 7.5 shows that the measured width w of the RML reconstructions follows the prediction of Equation 7.12. For larger values of $\alpha > 15\,\mu$as, the kernel size approaches the ring radius, and higher order effects become important (i.e., contributions from the opposite side of the ring in the convolved width). Because it is intrinsically built of point sources that are *not* confined to a δ-function in radius, the `DIFMAP` image does not follow this simple prediction for the increase in width with blurring kernel size, and it instead increases more rapidly to converge with the RML result at $\alpha \approx 20\,\mu$as.

For the RML reconstructions (and for `DIFMAP` with $\alpha > 10\,\mu$as), the measured orientation angle is relatively unaffected by the Gaussian convolution. In contrast, of all the parameters defined in Section 7.2, the fractional central brightness $f_{\rm C}$ between the average ring center brightness and the rim varies the most with resolution. In the absence of a restoring beam, both `SMILI` and `DIFMAP` produce rings with practically zero brightness in the ring center; as a result of this near-zero floor, $f_{\rm C}$ is extremely small ($f_{\rm C} < 10^{-5}$). As convolution with a finite Gaussian kernel fills in the center of the ring, $f_{\rm C}$ increases rapidly with α by several orders of magnitude. In contrast, because they include inverse-tapering of the initial visibility data and a final blurring with a 5 μas Gaussian (Paper V, Section 6), the `eht-imaging` fiducial reconstructions always have a non-zero central brightness. All three imaging methods give $f_{\rm C} \approx 0.3$ for $\alpha = 20\,\mu$as; this value represents an upper bound on $f_{\rm C}$ at the most conservative image resolution.

7.6 Radial and azimuthal profiles

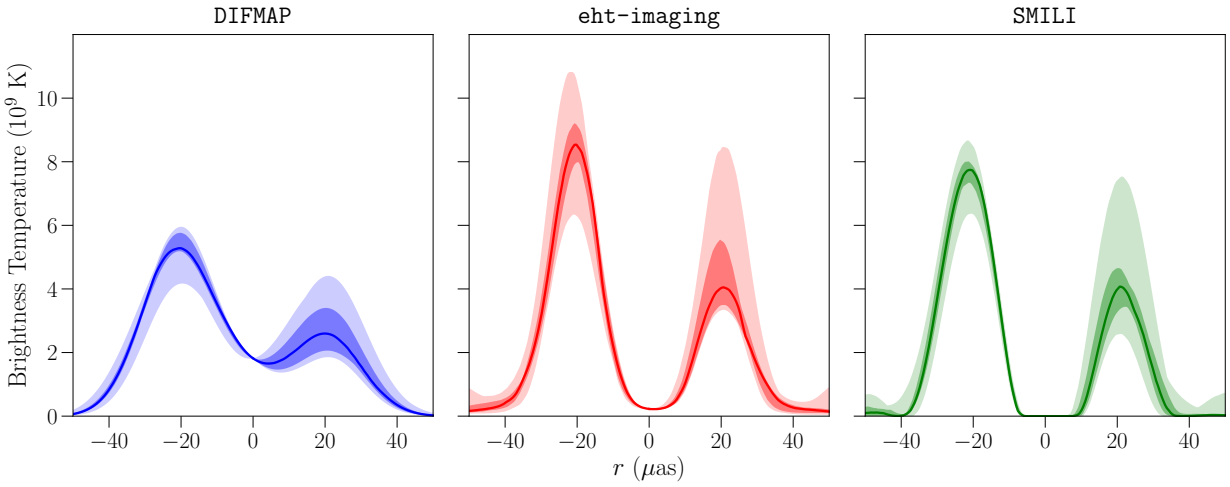


Figure 7.7: One-dimensional radial brightness profiles of the three fiducial M87 images on April 11 (Figure 37 of Paper IV). For each image, radial profiles in the semicircle centered on η are plotted with negative values of r, and radial profiles through the opposing semicircle centered on $\eta + 180°$ are plotted with positive r. The solid curves show the median profile over the corresponding semicircle, the darker band shows the 25th to 75th percentile range, and the lighter band shows the full range of profiles in the fiducial images.

Identifying a ring in the fiducial images allows the brightness distribution around the M87 shadow to be "unwrapped" and displayed in $I(r, \theta)$ space.[6] Figure 7.6 shows these unwrapped ring profiles of the fiducial images for all pipelines and days. In addition to the larger width of the DIFMAP reconstructed rings, the difference in position angle of the peak brightness is evident in these unwrapped profiles. Figure 7.6 also indicates visually that the measured brightness-weighted position angle η (Table 7.1) is more consistent than the angle of peak brightness across different reconstruction pipelines. η shows counterclockwise evolution between April 5 and April 11 in the fiducial images for all three methods.

Figure 7.7 shows radial profiles taken across the rings identified in the three April 11 fiducial images, with the DIFMAP image restored by the nominal 20 μas beam. For each image, the ring was divided in two by the line perpendicular to the measured orientation angle. On each half-ring,

[6] This section was originally published as Appendix I of Paper IV.

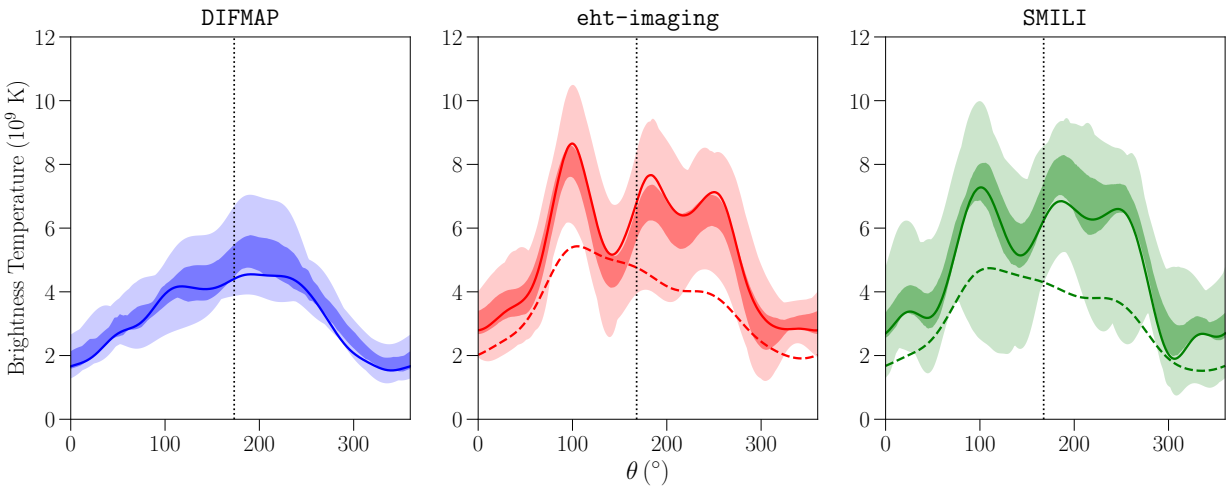


Figure 7.8: One-dimensional angular profiles of M87 on April 11 from the three imaging methods (Figure 38 of Paper IV). For each method, the solid line shows the angular profile obtained from the fiducial image, the darker band shows the 25th to 75th percentile range across the Top Set, and the lighter band shows the full Top Set range. For the RML methods, the dashed line shows the angular profile from the fiducial image blurred to the resolution of the DIFMAP image. The "knot" features in the unblurred eht-imaging and SMILI reconstructions show significant variation across the Top Set, suggesting that these features are sensitive to imaging parameter choices.

the median angular profile is denoted by solid lines in Figure 7.7), and the 25–75% and 0–100% percentile ranges of variation are shown as bands around the median. Figure 7.7 shows that the peak-to-peak ring diameters measured from the three imaging methods are broadly consistent, as indicated by the diameter measurements plotted in Figure 7.4. The overall sense of ring asymmetry recovered from each imaging pipeline is also consistent; the ring is always brighter in the south.

However, the shapes of the median radial profiles differ among the different imaging methods due to the different assumptions made in the imaging process. The DIFMAP reconstructions in Figure 7.7 are restored with a 20 μas beam, so they produce wider, shallower radial profiles. In contrast, the SMILI images tend to zero out low brightness regions, and the fiducial images have a near-zero brightness depression in the center of the ring. The eht-imaging results are higher-resolution than the DIFMAP results, but they have a less dramatic brightness depression that the SMILI images as a result of the choice to include a 5 μas maximum resolution in the eht-imaging script, as well as the use of maximum entropy regularization.

Figure 7.8 shows the angular profiles along the ring from the three imaging pipelines on April 11 data. While Figure 7.7 showed variability in the radial profiles from the individual fiducial images, Figure 7.8 considers variability in the azimuthal structure across the entire Top Set. In each panel, the solid line represents the fiducial value of the angular profile, and the bands show variability across the Top Set.

The median `DIFMAP` angular profile in Figure 7.8 is smooth and shallow, and it is distinct from those of the RML methods due to its 20 μas restoring beam. The dashed lines in the `eht-imaging` and `SMILI` panels show the angular profiles from the fiducial image blurred to match the `DIFMAP` resolution. When blurred, the angular profiles from the `eht-imaging` and `SMILI` images better match the broad `DIFMAP` profile, but they still differ in the measured orientation angle and the position of the brightest location on the ring.

The angular profiles for the unblurred `eht-imaging` and `SMILI` reconstructions are similar, with the 0.1 Jy difference in total flux density in these reconstructions manifesting in an overall lower profile for the `SMILI` reconstruction. The "knot" features in the `eht-imaging` and `SMILI` reconstructions show significant variation across the Top Sets. This variability suggests that these ring features are highly sensitive to imaging parameter choices, and that they are potentially artifacts of the limited (u, v) coverage. In general, the azimuthal structure in the reconstructed images from all three imaging pipelines is more variable than the radial structure, making measurements of the orientation angle intrinsically more uncertain than measurements of the ring diameter.

7.7 Weighing a black hole from an image

Measurements of the diameter of the ring feature in the first EHT images of M87 consistently find the diameter $d \approx 40\,\mu$as (Table 7.1). The lensed photon ring of a supermassive black hole has an

angular diameter in the range $d \approx (5.0 \pm 0.2)\theta_g$, depending on its inclination and spin (Bardeen et al., 1972). The angular size of a gravitational radius is

$$\theta_g = \frac{GM}{c^2 D}. \tag{7.13}$$

For a distance $D = 16.8$ Mpc to M87 (chosen based on three recent studies summarized in Paper VI, Appendix I), the ring in the EHT images is thus consistent with the lensed photon ring or shadow boundary of a $\sim 6.5 \times 10^9 M_\odot$ supermassive black hole (Paper IV).

To more rigorously measure the mass of the black hole from the ring size requires evaluating biases and uncertainties from several sources including the finite image resolution bias explored in Section 7.5, image structure outside the photon ring, the black hole spin, and the distance to M87. For the first EHT results, this analysis was performed in detail in Paper VI for measurements from both geometric model fits and image reconstructions. The following section presents a simplified version of the Paper VI analysis as applied only to the `eht-imaging` results reported in this chapter (Table 7.1).

In particular, the image domain analysis in Paper VI, Section 7 measures the black hole mass by finding the factor α (not to be confused with the restoring beam size in Section 7.5) that relates the measured diameters in Table 7.1 with the gravitational radius angular size θ_g:

$$d = \alpha \theta_g \tag{7.14}$$

If the measured ring feature in these images were to correspond exactly to the lensed photon ring diameter, α would be in the range $9.6\theta_g < \alpha < 10.4\theta_g$ (Paper VI). However, in realistic GRMHD simulations (including the two-temperature M87 simulations of Chapter 4), the measured ring diameters from images with degraded resolution do not correspond perfectly to the photon ring

diameter that is known a priori. Two effects contribute to this discrepancy. First, the diameter-width bias explored in Section 7.5 tends to reduce the ring diameter if the image is analyzed with less-than-perfect resolution. Second, as seen in GRMHD reconstructions, not all emission from the near-horizon region is expected to be concentrated in the photon ring. Some emission from the surrounding accretion flow and jet will contribute to a nonzero brightness both inside and outside the ring; depending on where this extra brightness lives in the image plane, it can bias the ring diameter inwards or outward.

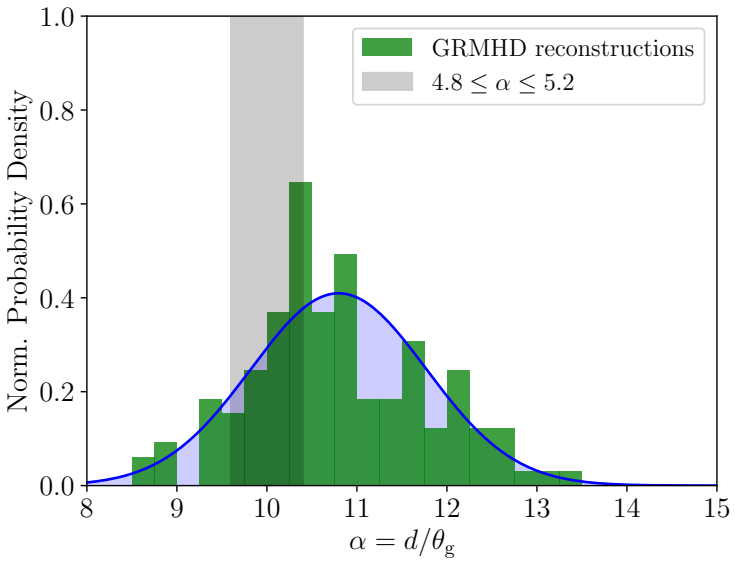


Figure 7.9: Histogram of REx measurements of the scale factor $\alpha = d/\theta_g$ from `eht-imaging` reconstructions (using the M87 fiducial script) of the J+ calibration data set of GRMHD snapshot images, as described in Section 7 of Paper VI. The blue curve is the best-fit normal distribution to the data. Note that this histogram contains outlier points from poor image reconstructions that are not included in the analysis in Paper VI, Section 7.

To measure θ_g from the REx measurements of the diameter d, Paper VI first measured α and its associated uncertainty from a suite of 103 snapshot images ("J+") spanning a wide range of GRMHD simulations with different physical parameters (e.g., black hole spin, magnetic flux on the horizon, post-processing electron temperature prescription) taken from the image library presented in Paper V. Synthetic data were generated from these images with `eht-imaging` (following the procedure described in Paper IV and Appendix C with realistic thermal noise, station gain

errors, and polarization leakage) on the (u, v) points corresponding to the four days of EHT 2017 observations. Images of all the J+ GRMHD data sets were then reconstructed using the fiducial imaging scripts from each pipeline and the ring diameters were measured using `REx`. While Paper VI considers results from all three imaging pipelines, the remainder of this section focuses only on the `eht-imaging` results.

Figure 7.9 shows a histogram of the resulting measurements of α from the full set of GRMHD snapshot images, as well as a best-fit normal distribution to the histogram. The median measurement of α from this set is $10.71 \pm 0.59\,\mu$as, where the quoted uncertainty is the median absolute deviation of the set of α values. Fitting a normal distribution to the histogram measures a mean and standard deviation or α of

$$\alpha = 10.80 \pm 0.97. \tag{7.15}$$

The central value of α in this set is biased upward from the range of lensed photon ring diameters produced in the Kerr metric ($9.6\theta_{\rm g}$–$10.4\theta_{\rm g}$). This bias is primarily the result of emission in the GRMHD simulation from outside the photon ring contributing to the brightness around the reconstructed ring (Paper V; Paper VI).[7]

When converting a diameter measurement d to a measurement of $\theta_{\rm g}$ for images of M87's supermassive black hole, the "theory error", or the scatter in α from the diameter measurement from different GRMHD snapshots dominates the intrinsic uncertainty in the diameter measurement from the feature extraction method reported in Table 7.1. Taking April 11 as a representative day, the

[7]Neither of these measurements of α is exactly the same as the value quoted in Paper VI, Table 6 for `eht-imaging`. They derive the measurement of α and its associated error budget with fits to generalized λ distributions. Furthermore, the analysis in this section includes outlier points from poor image reconstructions that are dropped in Paper VI. The aim in this section to to provide a slightly less rigorous analysis than the full Paper VI procedure that still captures the main features of the GRMHD calibration.

fiducial image measurement of d from Table 7.1 is

$$d = 41.0 \pm 1.4\,\mu\text{as} \quad (\texttt{eht-imaging}\text{ April 11 fiducial reconstruction}), \tag{7.16}$$

where, as in Table 7.1, the uncertainty includes contributions from both the imaging method and the scatter in d across the Top Set. Including both the measurement uncertainty and the uncertainty in α gives a measurement of the gravitational angular scale of M87:

$$\theta_{\rm g} = d/\alpha = 3.80 \pm 0.36. \tag{7.17}$$

The estimate of $\theta_{\rm g}$ in Equation 7.17 is very close to the value reported in Table 6 of Paper VI, $\theta_{\rm g} = 3.79^{+0.42}_{-0.37}\,\mu$as, derived from the average diameter measured from the `eht-imaging` Top Sets on all four days.

Finally, having measured d from the April 11 `eht-imaging` fiducial reconstruction of M87, and calibrating to a set of GRMHD reconstruction measurements made using the same script to find $\theta_{\rm g}$, it is possible to estimate the black hole mass $M = c^2 D \theta_{\rm g}/G$. The distance to M87 used in Paper VI is $D = 16.8 \pm 0.8$ Mpc. Thus,[8]

$$M = (6.47 \pm 0.62) \times 10^9\,M_\odot. \tag{7.18}$$

For comparison, the final mass measurement in Paper VI derived three independent analysis pipelines – geometric model fitting, image feature extraction with multiple pipelines, and direct GRMHD model fitting – is $(6.5 \pm 0.7) \times 10^9\,M_\odot$, in good agreement with this section's less careful analysis. The EHT's measurement of the mass of the supermassive black hole in M87 results

[8] The factor $\frac{c^2}{G\,1\,\mu\text{as}} = 1.0128 \times 10^8\,M_\odot$.

from directly imaging the emission from the near-horizon region. It is in excellent agreement with the Gebhardt et al. (2011) measurement (scaled for this distance) of $6.14^{+1.07}_{-0.62} \times 10^9\,M_\odot$ from stellar dynamics, but it is inconsistent with the gas dynamical measurement of Walsh et al. (2013) ($3.45^{+0.85}_{-0.26} \times 10^9\,M_\odot$).

7.8 Summary and conclusions

The April 2017 Event Horizon Telescope observations were the first to produce an image of the "shadow" of a supermassive black hole. The lensed photon ring and surrounding emission apparent in the 230 GHz EHT images of M87 (Figure 7.1) is the product of synchrotron emission in the black hole accretion flow and jet only a few gravitational radii from the event horizon, including from the photon orbit at $1 - 3\,r_g$.[9] The black hole mass inferred from these images is $(6.5 \pm 0.7) \times 10^9\,M_\odot$, implying that M87 is one of the most massive black holes in the observable universe.

The `eht-imaging` library (Chapter 6) played a crucial role in producing these first images of a black hole shadow. `eht-imaging` was used both in the initial, blind imaging of the M87 data and as one of three imaging pipelines in a systematic parameter search exploring a large range of choices in imaging parameters and their effects on the final image. Top Sets of best-performing parameters were selected from these surveys systematically based on their performance on synthetic data sets, and the final results from all three pipelines were consistent. At 230 GHz, the image of M87 is dominated by a ring of diameter $d \approx 40\,\mu$as, with a width $w \lesssim 15\,\mu$as, and with enhanced brightness toward the south (Paper IV). These characteristic image features are consistent with simulated images from a wide range of GRMHD simulations of a hot, thick accretion flow around a black hole of mass $M \approx 6.5 \times 10^9\,M_\odot$, with a nonzero black hole spin pointed away from the Earth. Notably, they are consistent with the predictions from the high-spinning, MAD, two-temperature

[9] For a spin $a = 1$ and $a = 0$, respectively.

simulations of the M87 accretion flow presented in Chapter 3. To launch the powerful jet observed in VLBI images at lower frequencies, magnetic fields in this accretion flow must be extracting rotational energy from the black hole by the Blandford-Znajek mechanism, and the overall sense of jet rotation at lower frequencies and the observed asymmetry in the ring observed by the EHT imply that the black hole spin vector is oriented anti-parallel to the line of sight (Paper V).

These images are the first direct probes of the inner accretion flow and jet launching region of a black hole; they represent only a glimpse of the potential science opened up by the EHT. Future EHT observations of M87 (including processing and imaging of the 2018 data already collected) will refine the measurement of the black hole mass and enable more precise tests of simulations of M87 with different assumptions about the accretion and plasma physics, such as those presented in Chapter 4 (Chael et al., 2019b). Polarimetric and multi-frequency images of the 2017 and 2018 M87 data will provide more stringent constraints on the origin of the emission in the accretion flow, the magnetic field strength, and the temperature of the emitting electrons. Repeated observations will further constrain time evolution on day-to-day timescales such as that seen across the week of observation in 2017 (Paper IV), potentially connecting these observations to the motion of magnetic field lines around the black hole event horizon. Eventually, with a combination of an expanded array and more advanced imaging techniques, future EHT images with increased dynamic range will be able to image the extended jet at the launching point in addition to the bright photon ring. These observations will enable further tests of the jet launching physics and energy extraction from the black hole and directly connect the energy delivered out of the entire galaxy at kpc scales to the black hole horizon.

Text in this Appendix was previously published in *MNRAS* 478 (2018), 4, pp 5209-5229 (A. Chael, M. Rowan, R. Narayan, M. Johnson, and L. Sironi), and in *MNRAS* 486 (2019), 2, pp 2873-2895 (A. Chael, R. Narayan, and M. Johnson)

Appendix A

Observational data

A.1 Sgr A*

The observed SED of Sgr A* presented in Chapter 2 (Figure 2.4) are mostly the same as plotted in the spectra in Ressler et al. (2017) with some additions.

Radio and millimeter points are from Falcke et al. (1998) over the range 1.46–235.6 GHz, from An et al. (2005) over the range 0.33–42.9 GHz, from Bower et al. (2015) in the range 1.6–352.6 GHz, and from Liu et al. (2016a) and Liu et al. (2016b) in the interval 93–709 GHz and at 492 GHz, respectively. 230 GHz measurements of the total flux density using the Event Horizon Telescope (EHT) are taken from Doeleman et al. (2008) and Johnson et al. (2015).

Infrared upper limits are from Cotera et al. (1999) in the range 8.7–24.5 μm, from Genzel & Eckart (1999) at 2.2 μm, and from Schödel et al. (2007) at 8.6 μm. Genzel et al. (2003) provide infrared flux density measurements for both Sgr A*'s quiescent state and flares at 1.76, 2.16, and 3.76 μm. Schödel et al. (2011) provide quiescent state measurements at 2.1, 3.8, and 4.8 μm, and Witzel et al. (2012) report a quiescent value at 2.2 μm.

The observed range of X-ray flare luminosities over the range 2-10 keV is reported in Neilsen et al. (2013), and the measurement of the quiescent X-ray luminosity is taken from Baganoff et al.

(2003). As Neilsen et al. (2013) note that only about 10% of the X-ray quiescent luminosity is produced in the inner accretion flow, Figure 4.9 plots the range between 10% and 100% of the Baganoff et al. (2003) measurement as the lower shaded band.

Simple estimates of the root-mean-square (RMS) variability in the 230 GHz light curve are plotted as 20 and 40% bands in Figure 2.5. Marrone et al. (2008), Yusef-Zadeh et al. (2009), and Bower et al. (2015) all report a value of roughly 20% RMS variability relative to the mean. Finally, the 230 GHz Sgr A* image size estimate in the E-W direction is from Event Horizon Telescope data reported in Doeleman et al. (2008) and Johnson et al. (2015).

A.2 M87

Comparing the simulation SEDs of Chapter 3 with observations of M87 requires some care. Because the total radio flux density along the extended jet of M87 is comparable to that of the significantly brighter but compact region near the black hole (the "core"), a meaningful comparison requires excising jet contributions that are outside the simulated domain. The data points in Chapter 3 (Figure 3.7) are based on Table 1 of Prieto et al. (2016), which compiles total flux density measurements from radio to X-ray from M87 in its quiescent state, using only measurements that achieve at least $0.4''$ resolution in order to securely exclude emission from the brightest jet knot, HST-1.

Prieto et al. (2016) have also compiled measurements of the total flux density of the most compact component identified by VLBI observations (their Table 4). However, these latter measurements have some notable limitations. For example, at 86 GHz, Prieto et al. (2016) include the value $S_{\rm L} = 0.16 \pm 0.07\,{\rm Jy}$ measured by Lee et al. (2008) as the measured flux density on the longest baseline for observations with the Coordinated (Global) Millimeter VLBI Array (CMVA/GMVA). This approach is problematic because M87's core is resolved at 86 GHz (Kim et al., 2018) and

Table A.1: The total and compact radio spectrum of M87 from recent VLBI observations.

Frequency [GHz]	Total Flux Density [Jy]	Core Flux Density [Jy]
15.4	2.2 ± 0.3	1.3 ± 0.1^a
22	2.1 ± 0.1	1.2 ± 0.1^b
43.1	1.6 ± 0.4	0.7 ± 0.2^c
86.3	1.1 ± 0.5	0.8 ± 0.4^d
230.0	2.05 ± 0.15	0.98 ± 0.05^e

[a] 19 MOJAVE observations from 2001-2011 (Lister et al., 2018).
[b] 10 KaVA & 3 VLBA (24 GHz) observations from 2013-2014 (Hada et al., 2017).
[c] 50 VLBA observations from 1999-2016 (Table 3 in Walker et al., 2018).
[d] 5 GMVA observations analyzed by Kim et al. (2018).
[e] 2 EHT observations: Doeleman et al. (2012) & Akiyama et al. (2015).

because interference among compact components can significantly affect the correlated flux density on a single baseline. At 22 GHz, Prieto et al. (2016) include the compact flux density value 0.35 Jy reported by Junor & Biretta (1995). However, during the observing epochs considered in Junor & Biretta (1995), the total flux density of M87 was only ∼1.1 Jy, which is significantly lower than the values measured more recently with the Very Long Baseline Array (VLBA) and the KVN/VERA Array (KaVA) (Hada et al., 2017). The Junor & Biretta (1995) measurement is also lower than the values measured at 15 GHz and 43 GHz since 2000 (Lister et al., 2018; Walker et al., 2018).

Because the simulation spectra in Chapter 3 are normalized to have a total flux density at 230 GHz that matches EHT measurements taken in 2009 and 2012 (Doeleman et al., 2012; Akiyama et al., 2015), Table A.1 provides updated estimates of the total and compact flux density of M87 from 15–230 GHz. For each row in Table A.1, the flux density of the compact component was estimated from the peak flux density of a beam convolved VLBI image at that frequency. This procedure gives a direct comparison between simulated images and reconstructed images from VLBI. Note that the total flux density measured with VLBI may still contain significant contributions from outside the raytracing domain of the M87 simulations in Chapter 3, especially at centimeter wavelengths.

Page intentionally left blank

Text in this Appendix was previously published in *MNRAS* 470, 2, pp 2367-2386 (A. Chael, R. Narayan, and A. Sądowski), *MNRAS* 478, 4, pp 5209-5229 (A. Chael, M. Rowan, R. Narayan, M. Johnson, and L. Sironi).

Appendix B

Simulation initial conditions

The simulation grid used in Chapters 2-4 is defined by a mapping that takes code coordinates x_1, x_2, x_3 to standard Kerr-Schild coordinates (r, θ, ϕ) in the Kerr metric (Gammie et al., 2003). This coordinate mapping is exponential in r and concentrates grid cells near the equator. The chosen functional form also naturally 'cylindrifies' grid cells somewhat at small radii closer to the poles, expanding them laterally at small radii so that the coordinates in the inner region are more cylindrical than spherical. This cylindrification speeds up the simulation by limiting the time step constraint imposed by the Courant condition (Tchekhovskoy et al., 2011).

The KORAL grid is defined by the equations:

$$
\begin{aligned}
r &= \mathrm{e}^{x_1} + r_0, \\
\theta &= \frac{\pi}{2}\left\{1 + \tan\left[\pi h_0 \left((1-2x_2)\left(\frac{2^p(y_2-y_1)}{(\mathrm{e}^{x_1}+r_0)^p} + y_1\right) + \right.\right.\right. \\
&\qquad \left.\left.\left. \left(x_2 - \frac{1}{2}\right)\right)\right] \cot\left[\frac{\pi h_0}{2}\right]\right\}, \\
\phi &= x_3.
\end{aligned}
\tag{B.1}
$$

The parameter $r_0 < 0$ changes the grid spacing near the origin, with a smaller $|r_0|$ placing more cells near the inner boundary $r_{\rm min}$. Increasing the parameter $h > 0$ concentrates cells toward the

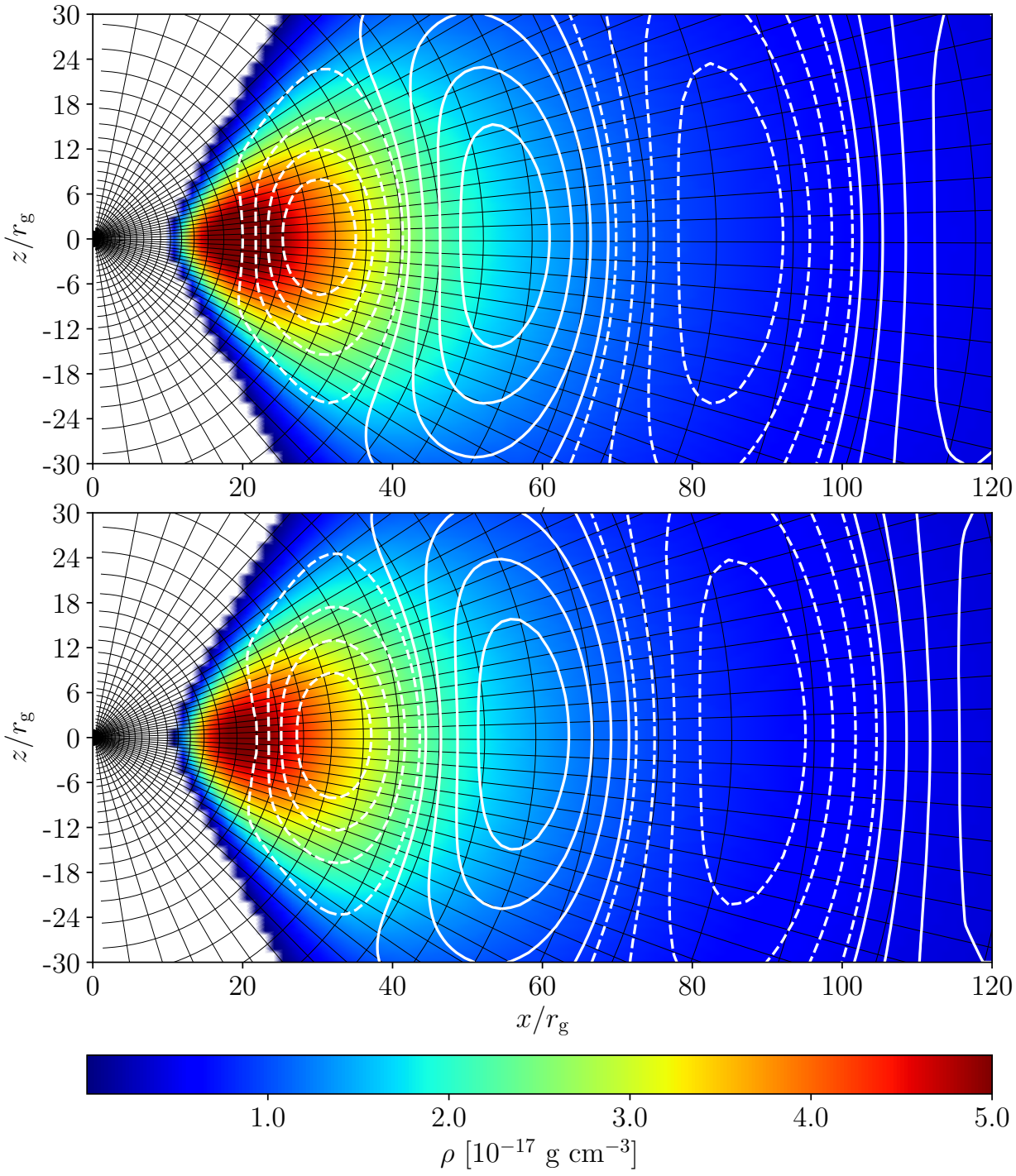


Figure B.1: Coordinate grid and density of the initial torii for the four Sgr A* simulations in Chapter 2. The top panel shows the initial torus for the two spin zero models R-Lo and H-Lo, and the bottom panel shows the torus for the two spin 0.9375 models R-Hi and H-Hi. White contours indicate the dipolar magnetic field lines.

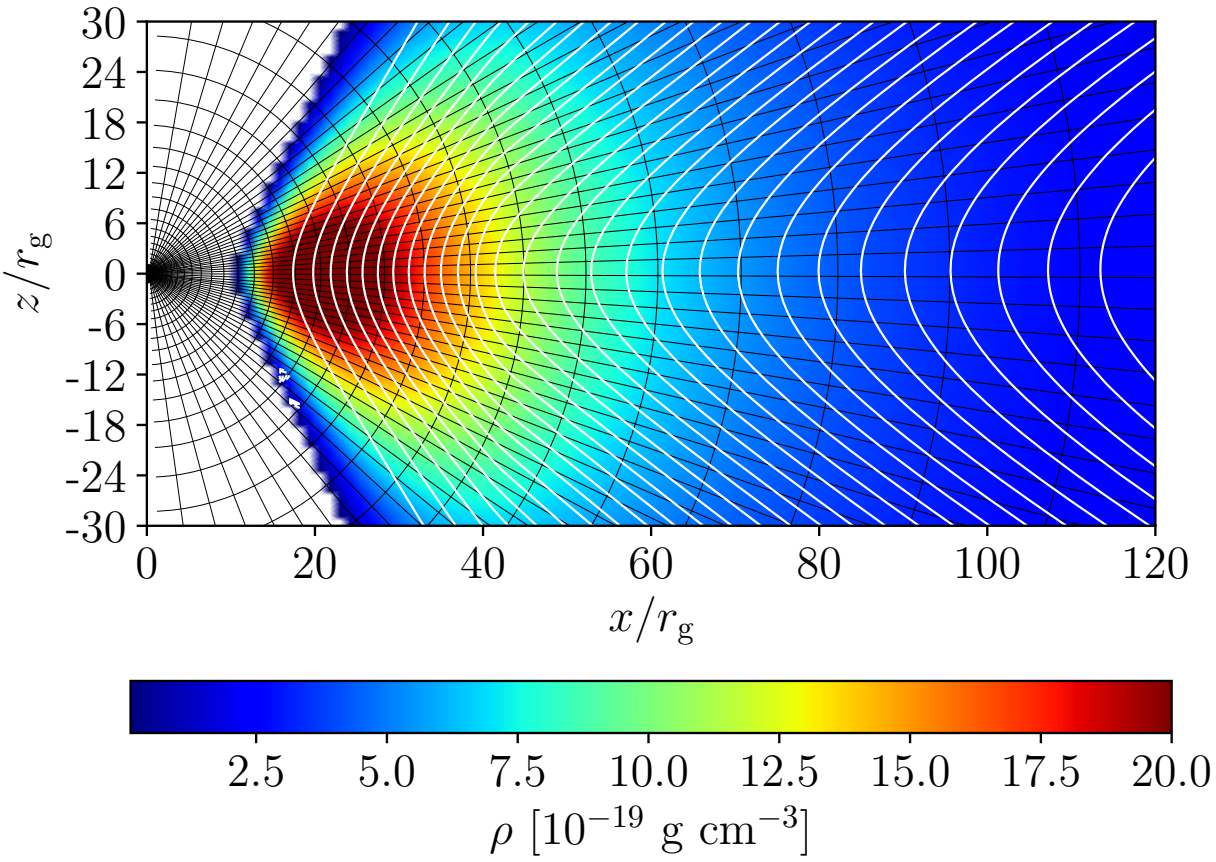


Figure B.2: Coordinate grid and initial torii for the two spin 0.9375 M87 simulations in Chapter 3. The white contours indicate the single loop of magnetic field lines in the initial conditions that leads the accretion disk to develop into a MAD state.

equatorial plane. Making $y_1 > 0$ larger (at fixed h) increases the minimum polar angle at large r, and increasing $y_2 > 0$ increases the minimum polar angle at small r. Adjusting the index $p > 0$ changes how quickly the minimum polar angle at a given radius changes between the value at $r_{\rm min}$ and the value at $r_{\rm max}$. For all the Sgr A* simulations in Chapter 2, $h_0 = 0.7$, $y_2 = 0.02$, $y_1 = 0.002$, $p = 1.3$, and $r_{\rm max} = 5,000$. For the spin $a = 0$ Sgr A* simulations, $r_0 = -2$ and $r_{\rm min} = 1.5$, while for $a = 0.9375$, $r_0 = -1.35$ and $r_{\rm min} = 1$. For the M87 simulations in Chapter 3, the paramaters $h_0 = 0.7$, $y_2 = 0.02$, $y_1 = 0.002$, and $p = 1.3$ are unchanged, but $r_{\rm min} = 1$, $r_{\rm max} = 10^4$, and $r_0 = -1.35$.

The initial plasma torii in all the simulations were set using the Penna et al. (2013) model, which defines a torus that has an angular momentum profile in the equatorial plane that is proportional to the Keplerian value by a factor ξ over a certain radial range $[r_{\rm k,min}, r_{\rm k,max}]$. The equatorial plane angular momentum is constant outside these limits. The initial torus adiabatic index is fixed at $\Gamma_{\rm gas} = 5/3$. For Sgr A* (Chapter 2), the spin zero models R-Lo/H-Lo were initialized with an inner torus edge at $10\,r_{\rm g}$ and an angular momentum profile with values $\xi = 0.708$ times the Keplerian value in the range $[42r_{\rm g}, 1000r_{\rm g}]$. The spin $a = 0.9375$ torus in models R-Hi/H-Hi was nearly identical, except that the strong dependence of the Penna et al. (2013) model on spin means that setting the inner edge at exactly $10r_{\rm g}$ produces a torus that nearly fills the entire grid. To avoid this, the inner edge was set at $11\,r_{\rm g}$, keeping all other values fixed. In the M87 simulations (Chapter 3), the inner edge was set at $10.5\,r_{\rm g}$, $\xi = 0.7135$, and the angular momentum is Keplerian over the smaller range $[42r_{\rm g}, 800r_{\rm g}]$.

To produce SANE accretion disks with little magnetic flux threading the horizon, the initial magnetic field in the Sgr A* simulations was set up in the torus with alternating dipolar field loops; the field polarity alternates roughly every $\sim 30\,r_{\rm g}$. To build up magnetic flux on the black hole and produce magnetically arrested disks for the M87 simulations, the initial torii were threaded with a

single field loop of constant polarity. In all the simulations, the the field strength was normalized such that the minimum value of $\beta_{\rm i}$ in the midplane was 10^{-2}.

In all the simulations of both sources, the initial energy in electrons was taken to be 1% of the total gas energy, with the remainder in the ions. The initial torus was surrounded by a static atmosphere with an r^2 density profile and negligible mass and energy density. The initial radiation energy density was negligible everywhere. The initial torii and simulation grids used in Chapter 2 are displayed in Figure B.1, and the torus used in both M87 simulations in Chapter 3 is displayed in Figure B.2.

Page intentionally left blank

Text in this Appendix was previously published in *ApJL* 875 (2019), L4 (The Event Horizon Telescope Collaboration et al.)

Appendix C

Synthetic data generation with `eht-imaging`

The `eht-imaging` library produces synthetic visibilities from model images by computing their Fourier transform, adding random thermal noise (Equation 5.8), and corrupting the data with systematic station-based effects using the Jones matrix formalism (TMS). Diagonal terms in the Jones matrices scale the measured amplitudes and phases, while off-diagonal terms in the Jones matrices mix the measured polarizations. This ppendix describes the Jones matrix formalism used in `eht-imaging` for adding gain errors and polarimetric leakage to simulated data.

For example, to generate a synthetic data set `obs` from a polarimetric `Image` object `im` observed with an `Array` instance `arr`, including the full affects of atmospheric phase error, gain calibration error, atmospheric opacity, field rotation, and polarimetric leakage, a user could call,

```
obs = im.observe(arr, tint, tadv, tstart, tstop, bw,
                 ttype='nfft', add_th_noise=True,
                 jones=True, inv_jones=True,
                 opacitycal=False, ampcal=False, phasecal=False,
                 dcal=False, frcal=False, rlgaincal=True,
                 tau=0.1, taup=0.1,
                 gainp=0.1, gain_offset=0.1,
                 dterm_offset=0.05)
```

As in Section 6.2.4, `tint` is the scan integration time in seconds, `tadv` is the advance time between scans in seconds, `tstart` and `tstop` are the start and stop time in hours, and `bw` is the observing bandwidth.

`ehtim` takes circular polarizations as its primary basis. On a single baseline ij, the Fourier transform of the image gives the uncorrupted visibility of each polarization (RR'_{ij}, LL'_{ij}, LR'_{ij}, RL'_{ij}), which are assembled into a 2×2 correlation matrix:

$$\mathbf{V}'_{ij} = \begin{pmatrix} RR'_{ij} & RL'_{ij} \\ LR'_{ij} & LL'_{ij} \end{pmatrix}. \tag{C.1}$$

The thermal noise variance on each baseline and polarization is given by Equation 5.8 and assembled into a matrix σ^2_{ij} of the same form. The variance matrix σ^2_{ij} is acted on by the station Jones matrices in the same way as the visibility matrix $\mathbf{V}'$.

Including the effects of systematic and thermal noise, the simulated visibility matrix is

$$\mathbf{V}_{ij} = \mathbf{J}_i \mathbf{V}'_{ij} \mathbf{J}_j^\dagger + \mathcal{N}(\sigma_{ij}), \tag{C.2}$$

where $\mathbf{J}_i$ and $\mathbf{J}_j$ are the station Jones matrices. Jones matrices are enabled in a call to `Image.observe` with the keyword argument `jones=True`. Each Jones matrix has the form:

$$\mathbf{J} = \begin{pmatrix} g_{\mathrm{R}} & e^{\mathrm{i}\varphi}\, d_{\mathrm{R}}\, g_{\mathrm{R}} \\ e^{-\mathrm{i}\varphi}\, d_{\mathrm{L}}\, g_{\mathrm{L}} & g_{\mathrm{L}} \end{pmatrix}, \tag{C.3}$$

where g_{R}, g_{L} are the complex gain terms, d_{R} and d_{L} are the constant complex d-terms, and φ is a term from field rotation. Typically, $|g|$ should be less than unity due to losses in telescope sensitivity.

Table C.1: Field rotation parameters for the EHT stations.

Station	Receiver Mount	$f_{\rm par}$	$f_{\rm el}$	$\varphi_{\rm off}$
ALMA	Cassegrain	1	0	0
APEX	Nasmyth-Right	1	1	0
JCMT	Cassegrain	1	0	0
LMT	Nasmyth-Left	1	−1	0
PV	Nasmyth-Left	1	−1	0
SMA	Nasmyth-Left	1	−1	45°
SMT	Nasmyth-Right	1	1	0
SPT	Cassegrain	1	0	0

The complex gain terms include an absolute amplitude gain offset G and a random atmospheric phase ϕ. For the purposes of EHT synthetic data, $g_{\rm R} = g_{\rm L}$ because the atmosphere is not significantly birefringent at millimeter wavelengths. However, `ehtim` can apply different right- and left-circular polarization gains in constructing the Jones matrices if the keyword `rlgaincal` is set to `False`. The complex gain terms are

$$g_{\rm R} = g_{\rm L} = G(t)\, e^{{\rm i}\phi(t)}. \tag{C.4}$$

In `eht-imaging`, setting `phasecal=False in Image.observe` results in a phase error $\phi(t)$ drawn from a uniform distribution once per scan. The amplitude gain (setting `ampcal=False`) of a site at each time has three components: one from nonzero atmospheric opacity $G_{\rm atten}(t)$, one intrinsic time-stable component G_1, and a time-varying intrinsice component $G_2(t)$:

$$G(t) = G_{\rm atten}(t)\, |1 - |G_1| + G_2(t)|^{1/2}\,. \tag{C.5}$$

Both G_1 and G_2 are sampled from normal distributions with zero mean: $G_1 \sim \mathcal{N}(0, \sigma_{G_1})$ and $G_2(t) \sim \mathcal{N}(0, \sigma_{G_2})$. The standard deviation of the stable part, σ_{G_1} is determined by the `gain_offset` argument to `Image.observe`. The `gain_offset` keyword argument is 0.1 by default on all stations; in

addition to passing in another float, the user can also pass a dictionary associating a unique value of σ_{G_1} for each station. The standard deviation of the time-variable part, σ_{G_2}, is specified by the `gainp` argument which is also 0.1 by default; like `gain_offset`, `gainp` can be set either to a constant for all stations or to a dictionary providing a different uncertainty for each station.

The attenuation term $G_{\rm atten}(t)$ arises from the nonzero atmospheric opacity τ and the changing source elevation angle $\theta_{\rm el}(t)$:

$$G_{\rm atten} = e^{-\tau/(\epsilon+2\sin\theta_{\rm el})}, \tag{C.6}$$

where $\epsilon = 10^{-10}$ prevents the fraction from diverging as the elevation angle $\theta_{\rm el} \to 0$. The opacity τ is formed as the sum of a measured value τ_0 (stored as a column in the `Obsdata.data` table) and a random component τ_1:

$$\tau = \tau_0 + \tau_1, \tag{C.7}$$

where τ_1 is constant in time and normally distributed, $\tau \sim \mathcal{N}(0, \sigma_\tau)$. The standard deviation σ_τ can be specified in a call to `Image.observe` with the `taup` argument; `taup` is 0.1 by default and is the same for all stations.

The complex d terms, $d_{\rm R}$ and $d_{\rm L}$, are stationary in time. In addition to a stable measured part d_0 stored in an `Array.tarr` table, a user can assign a random component d_1 drawn from a complex circular normal distribution. That is, e.g.,

$$d_{\rm L} = d_{\rm L\,,0} + d_{\rm L\,,1}, \tag{C.8}$$

where $d_{\rm L\,,1} \sim \mathcal{N}(0, \sigma_d)$. The standard deviation σ_d is set with the `dterm_offset` keyword argument;

it is 0.05 by default, motivated by the estimates in Johnson et al. (2015) for EHT stations.[1] Like `gain_offset` and `gainp`, `dterm_offset` can either be passed as a float or a dictionary with station name keywords.

The field rotation phase term φ in Equation C.3 has three possible contributions depending upon the receiver mount type:

$$\varphi = f_{\rm el}\theta_{\rm el} + f_{\rm par}\psi_{\rm par} + \varphi_{\rm off}, \tag{C.9}$$

where $\theta_{\rm el}$ is the elevation angle, $\psi_{\rm par}$ is the parallactic angle, and $\varphi_{\rm off}$ is a constant offset. Cassegrain mounts have $f_{\rm par} = 1$ and $f_{\rm el} = 0$. Nasmyth mounts have $f_{\rm par} = 1$ and $f_{\rm el} = \pm 1$, depending on the handedness. For reference, the EHT station field rotation parameters are listed in Table C.1.

The effects of field rotation are deterministic and can be calibrated out in an a priori step. Furthermore, estimates of the opacity and station d-terms may also be applied in a prioi calibration. This a priori step can be simulated in `Image.observe` by setting `inv_jones=True`. This choice constructs the estimated inverse Jones matrices $\mathbf{J}_{{\rm est},j}$ and applies them to the fully corrupted visibility matrix $\mathbf{V}_{ij}$. The estimated Jones matrix contains the terms:

$$\mathbf{J}_{\rm est} = G_{{\rm atten},0} \begin{pmatrix} 1 & e^{i\varphi}\, d_{{\rm R},0} \\ e^{-i\varphi}\, d_{{\rm L},0} & 1 \end{pmatrix}, \tag{C.10}$$

where $G_{{\rm atten},0} = \exp\left[-\tau_0/(\epsilon + \sin\theta_{\rm el})\right]$ is the measured attenuation with no random component, and $d_{{\rm L},0}, d_{{\rm R},0}$ are the measured d-terms, with no random offset.

[1] The mean amplitude of the leakage terms is then 6.3%.

The a priori calibrated visibilities are then given by applying the inverse estimated Jones matrices

$$\mathbf{V}_{\mathrm{meas},ij} = [\mathbf{J}_{\mathrm{est},i}]^{-1} \left(\mathbf{J}_i \mathbf{V}'_{ij} \mathbf{J}_j^{\dagger} + \mathcal{N}(\sigma_{ij}) \right) \left[\mathbf{J}_{\mathrm{est},j}^{\dagger} \right]^{-1}. \tag{C.11}$$

Throughout, the variance matrix σ_{ij}^2 is transformed in the same way as the visibility matrix.

Text in this Appendix was previously published in *ApJ* 857 (2018), 1, 23 (A. Chael, M. Johnson, K. Bouman, L. Blackburn, K. Akiyama, and R. Narayan).

Appendix D

Imaging gradients

This Appendix presents the expressions for the gradients of the various data and regularizer terms (presented in Sections 5.2.2 and 5.2.3) that are used in the `eht-imaging` library.

D.1 Data term gradients

The equations below assume a DTFT matrix F_{ij} (see Equation 6.1); the conjugate transpose matrix $F_{ij}^{\dagger}$ gives the adjoint DTFT matrix (note that since the visibility data is sparsely sampled, $F^{\dagger}F \neq 1$).

The gradient of the complex visibility χ^2 term (Equation 5.16) with respect to an image pixel I_i (already presented in the main text as Equation 6.2) is

$$\frac{\partial}{\partial I_i}\chi^2_{\text{vis}} = -\frac{1}{N_V}\sum_j \text{Re}\left[F_{ij}^{\dagger}\left(\frac{V_j - \hat{V}_j}{\sigma_j^2}\right)\right]. \tag{D.1}$$

The gradient of the visibility amplitude χ^2 (Equation 5.17) is

$$\frac{\partial}{\partial I_i}\chi^2_{\text{amp}} = -\frac{2}{N_V}\sum_j \text{Re}\left[F_{ij}^{\dagger}\frac{\hat{V}_j}{\hat{A}_j}\left(\frac{\left(A_j - \hat{A}_j\right)}{\sigma_j^2}\right)\right]. \tag{D.2}$$

For the bispectrum χ^2 (Equation 5.18), the gradient is

$$\frac{\partial}{\partial I_i}\chi^2_{\text{bispec}} = -\frac{1}{N_{\text{B}}}\sum_j \text{Re}\left[\left(\frac{F^\dagger_{1,ij}}{\hat{V}^*_{1,j}} + \frac{F^\dagger_{2,ij}}{\hat{V}^*_{2,j}} + \frac{F^\dagger_{3,ij}}{\hat{V}^*_{3,j}}\right)\left(\frac{(V_{\text{B},j} - \hat{V}_{\text{B},j})\hat{V}_{\text{B},j}}{\sigma^2_{\text{B},j}}\right)\right], \tag{D.3}$$

where an individual bispectrum measurement $V_{\text{B}\,,j} = V_{1\,j}V_{2\,j}V_{3\,j}$.

The closure phase χ^2 (Equation 5.19) has a gradient

$$\frac{\partial}{\partial I_i}\chi^2_{\text{cl phase}} = -\frac{2}{N_\psi}\sum_j \text{Im}\left[\left(\frac{F^\dagger_{1,ij}}{\hat{V}^*_{1,j}} + \frac{F^\dagger_{2,ij}}{\hat{V}^*_{2,j}} + \frac{F^\dagger_{3,ij}}{\hat{V}^*_{3,j}}\right)\left(\frac{\sin(\psi_j - \hat{\psi}_j)}{\sigma^2_{\psi,j}}\right)\right]. \tag{D.4}$$

And finally, the gradient of the closure amplitude χ^2 term (Equation 5.20) is

$$\frac{\partial}{\partial I_i}\chi^2_{\text{cl amp}} = -\frac{2}{N_{\text{C}}}\sum_j \text{Re}\left[\left(\frac{F^\dagger_{1,ij}}{\hat{V}^*_{1,j}} + \frac{F^\dagger_{2,ij}}{\hat{V}^*_{2,j}} - \frac{F^\dagger_{3,ij}}{\hat{V}^*_{3,j}} - \frac{F^\dagger_{4,ij}}{\hat{V}^*_{4,j}}\right)\left(\frac{\left(A_{\text{C},j} - \hat{A}_{\text{C},j}\right)\hat{A}_{\text{C},j}}{\sigma^2_{\text{C},j}}\right)\right], \tag{D.5}$$

and for log closure amplitudes (Equation 5.21) it is

$$\frac{\partial}{\partial I_i}\chi^2_{\text{log cl amp}} = -\frac{2}{N_{\text{C}}}\sum_j \text{Re}\left[\left(\frac{F^\dagger_{1,ij}}{\hat{V}^*_{1,j}} + \frac{F^\dagger_{2,ij}}{\hat{V}^*_{2,j}} - \frac{F^\dagger_{3,ij}}{\hat{V}^*_{3,j}} - \frac{F^\dagger_{4,ij}}{\hat{V}^*_{4,j}}\right)\frac{\hat{A}^2_{\text{C},j}}{\sigma^2_{\text{C},j}}\log\left(\frac{A_{\text{C},j}}{\hat{A}_{\text{C},j}}\right)\right]. \tag{D.6}$$

D.2 Regularizer term gradients

The gradient of the entropy regularizer term S_{MEM} (Equation 5.22) with respect to a pixel I_k is

$$\frac{\partial S_{\text{MEM}}}{\partial I_k} = -\frac{1}{\zeta}\left(\log\frac{I_k}{P_k} + 1\right). \tag{D.7}$$

The gradient of the ℓ_1 norm regularizer S_{ℓ_1} (Equation 5.23) is

$$\frac{\partial S_{\ell 1}}{\partial I_k} = -\frac{1}{\zeta}\,\text{sign}[I_k]. \tag{D.8}$$

This gradient is not continuous; it may be preferable to use a continuously differentiable approximation to the absolute value operator (Akiyama et al. 2017a,b; Paper IV).

The total variation regularizer S_{TV} (Equation 5.24) with respect to a pixel $I_{k,l}$ (indexed now by its x and y position) also has a gradient that is not continuously differentiable everywhere.

$$\frac{\partial S_{TV}}{\partial I_{k,l}} = -\frac{1}{\zeta}\left[\frac{2I_{k,l} - I_{k+1,l} - I_{k,l+1}}{\sqrt{(I_{k+1,l} - I_{k,l})^2 + (I_{k,l+1} - I_{k,l})^2}} + \frac{I_{k,l} - I_{k-1,l}}{\sqrt{(I_{k,l} - I_{k-1,l})^2 + (I_{k-1,l+1} - I_{k-1,l})^2}} + \frac{I_{k,l} - I_{k,l-1}}{\sqrt{(I_{k+1,l-1} - I_{k,l-1})^2 + (I_{k,l} - I_{k,l-1})^2}}\right]. \tag{D.9}$$

Again, singularities in the gradient can be accounted for by modifying the form of the ℓ_2 norm to include a small bias ϵ when the argument is zero (see Appendix A of Paper IV).

In contrast, the gradient of the Total Squared Variation regularizer S_{TSV} (Equation 5.25) with respect to a pixel $I_{k,l}$ is continuously differentiable:

$$\frac{\partial S_{TSV}}{\partial I_{k,l}} = -\frac{2}{\zeta}\left[(2I_{k,l} - I_{k+1,l} - I_{k,l+1}) + (I_{k,l} - I_{k-1,l}) + (I_{k,l} - I_{k,l-1})\right]. \tag{D.10}$$

The total flux regularizer $S_{\text{tot flux}}$ (Equation 5.26) has a gradient (returning to a single index k for each pixel):

$$\frac{\partial S_{\text{tot flux}}}{\partial I_k} = -\frac{2}{\zeta}\left(I_k - f\right). \tag{D.11}$$

Finally, the gradient of the centroid regularizer S_{centroid} (Equation 5.27) is

$$\frac{\partial S_{\text{centroid}}}{\partial I_k} = -\frac{2}{\zeta}\left[\left(I_k x_k - f\delta_x\right) x_k + \left(I_k y_k - f\delta_y\right) y_k\right]. \tag{D.12}$$

Page intentionally left blank

Appendix E

A sample `eht-imaging` script

This Appendix presents a sample imaging script for `eht-imaging`, very similar to the one used to generate the `eht-imaging` fiducial images of M87 presented in Chapter 7. The imaging script relies only on `ehtim` funcions and concepts introduced in Chapter 6. It consists of several stages including: (1) data preparation, (2) generation of an initial image, (3) initial self-calibration, (4) alternating imaging and self-calibration stages in a loop, and finally (5) saving the final output and diagnostic plots.

```
################################################
# consensus_script_simple.py
# A simplified version of the Paper IV M87 imaging script
# March 19, 2019
################################################

import ehtim as eh
import numpy as np
```

The first part of the script sets parameters used throughout the imaging process. These include the data file to load, the number of pixels and field of view of the image to be produced, the FHWM of the initial Gaussian, the weights α_D and β_R on the objective function χ^2 and regularizer terms (Equation 5.14), the convergence criterion, and the maximum number of iterations in each imaging step. Unique to EHT data, the initial parameters also include the FWHM and flux density of a

Gaussian for initial self-calibration of LMT baselines (since the LMT amplitude calibration was particularly poor in 2017), and an a priori systematic noise tolerance to include on each baseline based on each telescope's estimated performance.

```
################################################
# Imaging Parameters
################################################

infile = './obs_m87_scanavg.uvfits' # input data file
outfile = './m87_out'               # output file name

# image parameters
zbl = 0.6        # Total flux in Jy
npix = 64        # number of pixels
fov = 128        # field of view (uas)
prior_fwhm = 40  # FWHM of the initial/prior image in uas

# initial data weights
amp_w = 0.2      # weight on amplitudes
cphase_w = 1     # weight on closure phases
camp_w = 1       # weight on log closure amplitudes

# regularizer weights
simple = 100    # entropy regularizer weight
tv = 0          # TV regularizer weight
tv2 = 1         # TSV regularizer weight
l1 = 10         # l1 regularizer weight
flux = 1.e2     # Weight on the total flux regularizer

# other imager parameters
zero_uv_max = 1.e8 # Baselines shorter than this are effectively zero
syserr = 0.02      # Non-closing noise tolerance
stop = 1.e-4       # Convergence criterion
major = 3          # Number of convergence cycles on imaging & blurring
maxit = 100        # Maximum number of iterations for imaging
ttype = 'nfft'     # Fourier transform type
transform = 'log'  # enforce positivity ('log') or not (None)
```

```
updates = False    # display updates as the imager progresses


# which sites to selfcal
self_cal_sites = ['SM','JC','AA','AP', 'LM','SP','AZ']

# frac gain tolerance below and above 1 for gains
gain_tol = [0.02,0.2]

LZgauss_flux = 0.6 # Flux density for initial self-cal Gaussian
LZgauss_size = 60  # FWHM (uas) for initial self-cal Gaussian

# Systematic noise tolerance for amplitude a-priori calibration errors
systematic_noise = {'AA': 0.012,'AP': 0.013,'AZ': 0.008,
                    'JC': 0.017,'LM': 0.18, 'PV': 0.012,
                    'SM': 0.018,'SP': 0.009}
```

The next step loads the data in the .uvfits file 'obs_m87_scanavg.uvfits' into an `Obsdata` object and prepares the data for imaging. Many of the steps in this stage are unique to EHT reconstructions. For instance, because of the large gap between the short intra-site baselines (e.g., ALMA-APEX) and VLBI baselines, the visibility amplitudes on the short baselines include contributions from large-scale emission that are resolved out by even the shortest VLBI baselines (LMT-SMT). Because the imager cannot produce an image with this missing flux density, the script rescales these short baselines ($< 10^8\lambda$) to the user-specified total compact flux density (`zbl`).

This step also adds a systematic noise tolerance `syserr` to the reported thermal noise level σ on each baseline; this additional noise tolerance reflects non-closing errors due to polarimetric leakage and other factors not accounted for in the a priori calibration. Finally, the LMT's a priori calibration in 2017 was poor (Paper IV), with many scan dropouts from poor pointing. To account for the LMT's large gain errors, the short LMT baselines to the SMT are initially self-calibrated to a compact Gaussian (with FWHM `LZgauss_size` and flux density `LZgauss_flux`) before imaging.

```
################################################
# Prepare the data
################################################

# load the uvfits file
obs = eh.obsdata.load_uvfits(infile)

# Find the resolution of the observation
res = obs.res()

# Estimate the total flux density from AA-AP
zbl_tot  = np.median(obs.unpack_bl('AA','AP','amp')['amp'])

# Rescale short baselines to excise contributions from extended flux.
if zbl != zbl_tot:
  obs = obs.rescale_zbl(zbl, zero_uv_max)

# Reorder stations based on snr
obs.reorder_tarr_snr()

# Add non-closing systematic noise to the observation for imaging
obs = obs.add_fractional_noise(syserr)

# Make a static copy of the observation
obs_static = obs.copy()

# Self calibrate the problematic LMT to a Gaussian
if LZgauss_flux > 0.0 and LZgauss_size > 0:

  # flag long baselines
  obs_LMT = obs_static.flag_uvdist(uv_max=2e9)

  # make Gaussian image
  gausspriorLMT_size = LZgauss_size * eh.RADPERUAS
  gausspriorLMT_dims = (gausspriorLMT_size, gausspriorLMT_size,  0, 0, 0)
  gausspriorLMT = eh.image.make_square(obs, npix, fov)
  gausspriorLMT = gausspriorLMT.add_gauss(LZgauss_flux,  gausspriorLMT_dims)
```

```
# derive LMT self-calibration solution from only short baselines
caltab = eh.selfcal(obs_LMT, gausspriorLMT,
                    sites=['LM'],
                    ttype=ttype,
                    caltable=True,
                    gain_tol=1.0)

# apply selfcal to full observation
obs = caltab.applycal(obs_static, interp='nearest', extrapolate=True)
```

The next section of the script prepares the initial image, which is also used as a prior image for the MEM regularizer. The initial image for this M87 imaging script is a 40 μas Gaussian with 0.6 Jy of total flux density. The script adds this Gaussian to an empty image with the built in `Image` method `Image.add_gauss`. Furthermore, the script also adds a weak (scaled by 10^{-3}) Gaussian component offset by the FWHM (`gaussprior_dim`). This additional offset Gaussian breaks the perfect symmetry of the initial image and prevents initial singularities in the total variation regularizer (Equation D.10).

```
################################################
# Prepare the initial image
################################################
# Make a Gaussian initial / prior image
# to avoid gradient singularities, add a slightly offset component

gaussprior_size = prior_fwhm*eh.RADPERUAS
gaussprior_dim   = (gaussprior_size, gaussprior_size, 0, 0, 0)
gaussprior_dim_off = (gaussprior_size, gaussprior_size, 0,
                      gaussprior_size, gaussprior_size)

gaussprior = eh.image.make_square(obs, npix, fov*eh.RADPERUAS)
gaussprior = gaussprior.add_gauss(zbl, gaussprior_dim)
gaussprior = gaussprior.add_gauss(zbl*1e-3, gaussprior_dim_off)
```

Next, the script sets up the `Imager` object with the specified initial data weights, regularizer weights, `Obsdata` object, and initial image. This instance of `Imager` is called `imgr`; the script uses this instance for the rest of the imaging process, updating its internal attributes as needed when the

data weights change or the `Obsdata` data object is self-calibrated. This section of the script also sets up a helper function called `converge`; this function runs the `Imager` several times in succession with `imgr.make_image_I`, replacing the initial image for the next imager run with the blurred output from the last run. This procedure acts to smooth out spurious high-frequency structure not constrained by the data and helps the imager avoid local minima in minimizing the objective function.

```
################################################
# Prepare the imager
################################################

# Define the imager
data_term1 = {'amp':amp_w, 'cphase':cphase_w, 'logcamp':camp_w }

data_term2 = {'vis':10*amp_w, 'cphase':10*cphase_w, 'logcamp':10*camp_w }

reg_term = {'simple': simple,
            'tv'  : tv,
            'tv2'  : tv2,
            'l1'  : l1,
            'flux' : flux }

imgr = eh.imager.Imager(obs, gaussprior,
                        prior_im = gaussprior,
                        flux = zbl,
                        data_term = data_term1,
                        reg_term = reg_term,
                        stop = stop,
                        maxit = maxit,
                        ttype = ttype,
                        systematic_noise=systematic_noise,
                        cp_uv_min=zero_uv_max)


# Define a helper function to repeat imaging
# with blurring to assure good convergence
def converge(imgr, major=major, blur_frac=1.0):
  imgr.make_image_I(show_updates=updates)
  for repeat in range(major):
```

```
    init = imgr.out_last().blur_circ(blur_frac*res)
    imgr.init_next = init
    imgr.make_image_I(show_updates=updates)
  return imgr
```

At this stage, the script begins the actual imaging and self-calibration procedure. The script first images using χ^2 terms on the visibility amplitudes, closure phases, and closure amplitudes. Then, the script self-calibrates the station phase terms to the final image from this first round.

```
################################################
# Imaging / Self-Calibration
################################################

# First round of imaging
print("Imaging with visibility amplitudes and closure quantities...")
imgr = converge(imgr)
out1 = imgr.out_last()
out1_blur = out1.blur_circ(res)

# First round of self-calibration (phase-only)
print("Self-Calibrating phases...")
obs_sc = eh.selfcal(obs_static, out1, ttype=ttype, method='phase')
```

In the second round, the script replaces the original data `obs_static` in `imgr` with the self-calibrated `Obsdata` object `obs_sc`. As a consequence, the amplitude χ^2 term is replaced with a the complex visibility term χ^2_{vis}, and the weights on all the data terms are increased by a factor of 10 relative to the fixed regularizer weights. The script then runs `converge` again to produce an image, and it follows this round of imaging with another round of self calibration on the phases. The script then calibrates the LMT amplitude alone, and the data in `imgr` are replaced with the latest self-calibrated data.

```
# Second round of imaging, increasing the data term weights 10x
print("Imaging with visibilities and closure quantities...")
imgr.init_next = out1_blur
imgr.obs_next = obs_sc
imgr.dat_term_next = data_term2
imgr = converge(imgr)
out2 = imgr.out_last()
out2_blur = out2.blur_circ(res)

# Second round of self-calibration (phase for all sites; amp for LMT)
print("Self-Calibrating phases and LMT amplitude...")
obs_sc = eh.selfcal(obs_static, out2, ttype=ttype, method='phase')
caltab = eh.selfcal(obs_sc, out2, sites=['LM'], ttype=ttype,
                    method='both',
                    gain_tol=gain_tol, caltable=True)
obs_sc = caltab.applycal(obs_sc, interp='nearest', extrapolate=True)
```

The script then concludes with two final rounds of imaging and self-calibration. In the self-calibration step, the script now solves for the amplitude and the phase offsets for all stations.

```
# Image and selfcal until end
for repeat_selfcal in range(2):

  # Image
  print("Imaging with visibility amplitudes and closure quantities...")
  imgr.init_next = out2_blur
  imgr.obs_next = obs_sc
  imgr.systematic_noise_next = 0.01 #reset systematic noise
  imgr = converge(imgr)

  # Self-calibrate
  print("Self-Calibrating phases and amplitudes...")
  caltab = eh.selfcal(obs_static, imgr.out_last(), ttype=ttype,
                      method='both',
                      gain_tol=gain_tol, caltable=True)
  obs_sc = caltab.applycal(obs_static, interp='nearest', extrapolate=True)
```

Finally, the script saves out the final image as a .fits file and produces a final self-calibrated data set that it saves to the .uvfits format. After generating these final outputs, the script produces and saves a .pdf file image-data consistency sheet (described in Section 6.6.2) that summarizes the final image, the final self-calibration gain solution, and the consistency of the self-calibrated data to the image.

```
################################################
# Save the results and produce a summary sheet
################################################

# Final image
im_out = imgr.out_last().copy()
im_out.save_fits(outfile + '.fits')

# Final self-cal data
obs_sc_out = eh.selfcal(obs_sc, im_out, ttype=ttype, method='both')
obs_sc_out.save_uvfits(outfile + '.uvfits')

# Image-data summary sheet
eh.imgsum(im_out, obs_sc_out, obs_static, outfile+'_imgsum.pdf', cp_uv_min=zero_uv_max)
```

Page intentionally left blank

References

ALMA Partnership et al., 2015, ApJL, 808, L3

Abramowicz M. A., Czerny B., Lasota J. P., Szuszkiewicz E., 1988, ApJ, 332, 646

Abramowicz M. A., Chen X., Kato S., Lasota J.-P., Regev O., 1995, ApJL, 438, L37

Abramowski A., et al., 2012, ApJ, 746, 151

Agol E., 2000, ApJL, 538, L121

Akiyama K., et al., 2015, ApJ, 807, 150

Akiyama E., Hasegawa Y., Hayashi M., Iguchi S., 2016, ApJ, 818, 158

Akiyama K., et al., 2017a, AJ, 153, 159

Akiyama K., et al., 2017b, ApJ, 838, 1

An T., et al., 2005, ApJL, 634, L49

Asada K., Nakamura M., 2012, ApJL, 745, L28

Baade W., Minkowski R., 1954, ApJ, 119, 215

Baganoff F. K., et al., 2003, ApJ, 591, 891

Balbus S., Hawley J., 1998, R. Mod. Phys., 70, 1

Balick B., Brown R. L., 1974, ApJ, 194, 265

Ball D., Özel F., Psaltis D., Chan C.-K., 2016, ApJ, 826, 77

Ball D., Sironi L., Özel F., 2018, ApJ, 862, 80

Bardeen J. M., Press W. H., Teukolsky S. A., 1972, ApJ, 178, 347

Baron F., Monnier J. D., Kloppenborg B., 2010, in Optical and Infrared Interferometry II. p. 77342I, doi:10.1117/12.857364

Barrière N. M., et al., 2014, ApJ, 786, 46

Bisnovatyi-Kogan G. S., Ruzmaikin A. A., 1976, Ap&SS, 42, 401

Blackburn L., et al., 2019, in prep.

Blandford R. D., Begelman M. C., 1999, MNRAS, 303, L1

Blandford R. D., Königl A., 1979, ApJ, 232, 34

Blandford R. D., Znajek R. L., 1977, MNRAS, 179, 433

Blandford R., Meier D., Readhead A., 2019, ARA&A, in press

Boldyrev S., Loureiro N. F., 2017, ApJ, 844, 125

Bouman K. L., Johnson M. D., Zoran D., Fish V. L., Doeleman S. S., Freeman W. T., 2016, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). , doi:10.1109/CVPR.2016.105

Bouman K. L., Johnson M. D., Dalca A. V., Chael A. A., Roelofs F., Doeleman S. S., Freeman W. T., 2018, IEEE Transactions on Computational Imaging, 4

Bower G. C., Wright M. C. H., Falcke H., Backer D. C., 2003, ApJ, 588, 331

Bower G. C., et al., 2015, ApJ, 802, 69

Bridle A. H., Perley R. A., 1984, ARA&A, 22, 319

Briggs D. S., 1995, in American Astronomical Society Meeting Abstracts. p. 1444

Broderick A., Loeb A., 2006, MNRAS, 367, 905

Broderick A., Loeb A., 2009, ApJ, 697, 1164

Broderick A. E., Tchekhovskoy A., 2015, ApJ, 809, 97

Buchler J. R., Yueh W. R., 1976, ApJ, 210, 440

Buscher D. F., 1994, in Robertson J. G., Tango W. J., eds, IAU Symposium Vol. 158, Very High Angular Resolution Imaging. p. 91

Byrd R. H., Lu P., Nocedal J., 1995, SIAM Journal on Scientific and Statistical Computing, 16, 1190

Carbone V., Veltri P., Mangeney A., 1990, Physics of Fluids A, 2, 1487

Carroll S. M., 2004, Spacetime and Geometry. An Introduction to General Relativity. Addison-Wesley, San Francisco

Chael A., Johnson M. D., Narayan R., Doeleman S. S., Wardle J. F. C., Bouman K. L., 2016, ApJ, 829, 11

Chael A., Narayan R., Sądowski A., 2017, MNRAS, 470, 2367

Chael A., Rowan M., Narayan R., Johnson M., Sironi L., 2018a, MNRAS, 478, 5209

Chael A., Johnson M. D., Bouman K. L., Blackburn L. L., Akiyama K., Narayan R., 2018b, ApJ, 857, 23

Chael A., et al., 2019a, ehtim: Imaging, analysis, and simulation software for radio interferometry, Astrophysics Source Code Library (ascl:1904.004), doi:http://doi.org/10.5281/zenodo.2614009

Chael A., Narayan R., Johnson M. D., 2019b, MNRAS, 486, 2873

Chan C.-K., Psaltis D., Özel F., Narayan R., Sądowski A., 2015a, ApJ, 799, 1

Chan C.-K., Psaltis D., Özel F., Medeiros L., Marrone D., Sądowski A., Narayan R., 2015b, ApJ, 812, 103

Chandra M., Foucart F., Gammie C. F., 2017, ApJ, 837, 92

Chandrasekhar S., 1939, An Introduction to the Study of Stellar Structure. The University of Chicago Press, Chicago

Chandrasekhar S., 1983, The Mathematical Theory of Black Holes. Oxford University Press, New York

Charbonnier P., Blanc-Feraud L., Aubert G., Barlaud M., 1997, IEEE Transactions on Image Processing, 6, 298

Chatzopoulos S., Fritz T. K., Gerhard O., Gillessen S., Wegg C., Genzel R., Pfuhl O., 2015, MNRAS, 447, 948

van Cittert P. H., 1934, Physica, 1, 201

Clark B. G., 1980, A&A, 89, 377

Comisso L., Asenjo F. A., 2018, Phys. Rev. D, 97, 043007

Cornwell T. J., 2008, IEEE Journal of Selected Topics in Signal Processing, 2, 793

Cornwell T. J., Evans K. F., 1985, A&A, 143, 77

Cornwell T. J., Fomalont E. B., 1999, in Taylor G. B., Carilli C. L., Perley R. A., eds, Astronomical Society of the Pacific Conference Series Vol. 180, Synthesis Imaging in Radio Astronomy II. p. 187

Cornwell T. J., Wilkinson P. N., 1981, MNRAS, 196, 1067

Cotera A., Morris M., Ghez A. M., Becklin E. E., Tanner A. M., Werner M. W., Stolovy S. R., 1999, in Falcke H., Cotera A., Duschl W. J., Melia F., Rieke M. J., eds, Astronomical Society of the Pacific Conference Series Vol. 186, The Central Parsecs of the Galaxy. p. 240

Coughlan C. P., Gabuzda D. C., 2016, MNRAS, 463, 1980

Curtis H. D., 1918, Publications of Lick Observatory, 13, 9

Davelaar J., Mościbrodzka M., Bronzwaer T., Falcke H., 2018, A&A, 612, A34

Dexter J., 2016, MNRAS, 462, 115

Dexter J., Agol E., Fragile P. C., McKinney J. C., 2010, ApJ, 717, 1092

Dexter J., McKinney J. C., Agol E., 2012, MNRAS, 421, 1517

Di Matteo T., Allen S. W., Fabian A. C., Wilson A. S., Young A. J., 2003, ApJ, 582, 133

Dodds-Eden K., et al., 2009, ApJ, 698, 676

Dodds-Eden K., et al., 2011, ApJ, 728, 37

Doeleman S., et al., 2008, Nature, 455, 78

Doeleman S., et al., 2012, Science, 338, 355

Dolence J. C., Gammie C. F., Mościbrodzka M., Leung P. K., 2009, ApJS, 184, 387

Eastwood J. P., Phan T. D., Drake J. F., Shay M. A., Borg A. L., Lavraud B., Taylor M. G. G. T., 2013, Phys. Rev. Lett., 110, 225001

Eckart A., et al., 2006, A&A, 450, 535

Eckart A., García-Marín M., Vogel S. N., Teuben P., Morris M. R., Baganoff F., Dexter J., et al., 2012, A&A, 537, A52

Einstein A., 1916, Annalen der Physik, 354, 769

Esin A. A., McClintock J. E., Narayan R., 1997, ApJ, 489, 865

Falcke H., Biermann P. L., 1995, A&A, 293, 665

Falcke H., Goss W. M., Matsuo H., Teuben P., Zhao J.-H., Zylka R., 1998, ApJ, 499, 731

Falcke H., Melia F., Agol E., 2000, ApJL, 528, L13

Farris B. D., Li T. K., Liu Y. T., Shapiro S. L., 2008, Phys. Rev. D, 78, 024023

Finkelstein D., 1958, Phys. Rev., 110, 965

Fish V. L., et al., 2014, ApJ, 795, 134

Fish V. L., Johnson M. D., Doeleman S. S., Broderick A. E., Psaltis D., Lu R.-S., Akiyama K., et al., 2016, ApJ, 820, 90

Fouka M., Ouichaoui S., 2013, Research in Astronomy and Astrophysics, 13, 680

Frieden B. R., 1972, Journal of the Optical Society of America (1917-1983), 62, 511

GRAVITY Collaboration et al., 2018a, A&A, 615, L15

GRAVITY Collaboration et al., 2018b, A&A, 618, L10

Gammie C. F., McKinney J. C., Tóth G., 2003, ApJ, 589, 444

de Gasperin F., et al., 2012, A&A, 547, A56

Gebhardt K., Adams J., Richstone D., Lauer T. R., Faber S. M., Gültekin K., Murphy J., Tremaine S., 2011, ApJ, 729, 119

Genzel R., Eckart A., 1999, in Falcke H., Cotera A., Duschl W. J., Melia F., Rieke M. J., eds, Astronomical Society of the Pacific Conference Series Vol. 186, The Central Parsecs of the Galaxy. p. 3

Genzel R., Schödel R., Ott T., Eckart A., Alexander T., Lacombe F., Rouan D., Aschenbach B., 2003, Nature, 425, 934

Genzel R., Eisenhauer F., Gillessen S., 2010, Reviews of Modern Physics, 82, 3121

Gillessen S., et al., 2006, ApJL, 640, L163

Gillessen S., Eisenhauer F., Trippe S., Alexander T., Genzel R., Martins F., Ott T., 2009, ApJ, 692, 1075

Ginzburg V. L., Syrovatskii S. I., 1964, The Origin of Cosmic Rays. Macmillan, New York

Giroletti M., et al., 2012, A&A, 538, L10

Gold R., McKinney J. C., Johnson M. D., Doeleman S. S., 2017, ApJ, 837, 180

Greene J. E., Ho L. C., 2007, ApJ, 667, 131

Gull S. F., Daniell G. J., 1978, Nature, 272, 686

Guo X., Sironi L., Narayan R., 2014, ApJ, 794, 153

Hada K., Doi A., Kino M., Nagai H., Hagiwara Y., Kawaguchi N., 2011, Nature, 477, 185

Hada K., et al., 2016, ApJ, 817, 131

Hada K., et al., 2017, PASJ, 69, 71

Hamaker J. P., Bregman J. D., Sault R. J., 1996, A&AS, 117, 137

Hawking S. W., 1974, Nature, 248, 30

Hawking S. W., Penrose R., 1970, Proceedings of the Royal Society of London Series A, 314, 529

Heinz S., Begelman M. C., 1997, ApJ, 490, 653

Herrnstein R. M., Zhao J.-H., Bower G. C., Goss W. M., 2004, AJ, 127, 3399

Hilbert D., 1917, Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen - Mathematisch-physikalische Klasse. Weidmannsche Buchhandlung, Berlin, pp 53–76

Ho L. C., 2008, ARA&A, 46, 475

Högbom J. A., 1974, A&AS, 15, 417

Holdaway M. A., 1990, PhD thesis, Brandeis Univ., Waltham, MA.

Holdaway M. A., Wardle J. F. C., 1990, in Gmitro A. F., Idell P. S., Lahaie I. J., eds, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 1351, Digital Image Synthesis and Inverse Optics. pp 714–724, doi:10.1117/12.23679

Honma M., Akiyama K., Uemura M., Ikeda S., 2014, PASJ, 66, 95

Hornstein S. D., Matthews K., Ghez A. M., Lu J. R., Morris M., Becklin E. E., Rafelski M., Baganoff F. K., 2007, ApJ, 667, 900

Howes G., 2010, MNRAS, 409, L104

Howes G., 2011, ApJ, 738, 40

Howes G., Dorland W., Cowley S. C., Hammett G. W., Quataert E., Schekochihin A. A., Tatsuno T., 2008a, Physical Review Letters, 100, 065004

Howes G., Cowley S. C., Dorland W., Hammett G. W., Quataert E., Schekochihin A. A., 2008b, Journal of Geophysical Research (Space Physics), 113, A05103

Hunter J. D., 2007, Computing in Science & Engineering, 9, 90

Ichimaru S., 1977, ApJ, 214, 840

Igumenshchev I. V., Narayan R., Abramowicz M. A., 2003, ApJ, 592, 1042

Issaoun S., et al., 2019, ApJ, 871, 30

Jennison R. C., 1958, MNRAS, 118, 276

Johnson M. D., 2016, ApJ, 833, 74

Johnson M. D., et al., 2015, Science, 350, 1242

Johnson M. D., et al., 2017, ApJ, 850, 172

Johnson M. D., et al., 2018, ApJ, 865, 104

Jones E., Oliphant T., Peterson P., et al., 2001, SciPy: Open source scientific tools for Python, http://www.scipy.org/

Jorstad S., Marscher A., 2016, Galaxies, 4, 47

Junor W., Biretta J. A., 1995, AJ, 109, 500

Junor W., Biretta J. A., Livio M., 1999, Nature, 401, 891

Kawazura Y., Barnes M., Schekochihin A. A., 2019, Proceedings of the National Academy of Science, 116, 771

Keiner J., Kunis S., Potts D., 2009, ACM Trans. Math. Softw., 36, 19:1

Kerr R. P., 1963, Physical Review Letters, 11, 237

Kim J.-Y., et al., 2018, A&A, 616, A188

King A., 2003, ApJL, 596, L27

King A., Pounds K., 2015, ARA&A, 53, 115

Kino M., Takahara F., Hada K., Doi A., 2014, ApJ, 786, 5

Kino M., Takahara F., Hada K., Akiyama K., Nagai H., Sohn B. W., 2015, ApJ, 803, 30

Komissarov S. S., 1999, MNRAS, 303, 343

Komissarov S. S., Barkov M. V., Vlahakis N., Königl A., 2007, MNRAS, 380, 51

Kormendy J., Ho L. C., 2013, ARA&A, 51, 511

Kovalev Y. Y., Lister M. L., Homan D. C., Kellermann K. I., 2007, ApJL, 668, L27

Kuo C. Y., et al., 2014, ApJL, 783, L33

Kuramochi K., Akiyama K., Ikeda S., Tazaki F., Fish V. L., Pu H.-Y., Asada K., Honma M., 2018, ApJ, 858, 56

Kusunose M., Takahara F., 2011, ApJ, 726, 54

Lee S.-S., Lobanov A. P., Krichbaum T. P., Witzel A., Zensus A., Bremer M., Greve A., Grewing M., 2008, AJ, 136, 159

Leung P. K., Gammie C. F., Noble S. C., 2011, ApJ, 737, 21

Levermore C. D., 1984, J. Quant. Spectrosc. Radiative Transfer, 31, 149

Li Y.-P., Yuan F., Wang Q. D., 2017, MNRAS, 468, 2552

LIGO Scientific Collaboration, the Virgo Collaboration, et al., 2016, Phys. Rev. Lett., 116, 061102

LIGO Scientific Collaboration, the Virgo Collaboration, et al., 2018, arXiv e-prints,

Lindquist R. W., 1966, Annals of Physics, 37, 487

Liska M., Hesp C., Tchekhovskoy A., Ingram A., van der Klis M., Markoff S., 2018, MNRAS, 474, L81

Lister M. L., Aller M. F., Aller H. D., Hodge M. A., Homan D. C., Kovalev Y. Y., Pushkarev A. B., Savolainen T., 2018, ApJS, 234, 12

Liu H. B., et al., 2016a, A&A, 593, A44

Liu H. B., et al., 2016b, A&A, 593, A107

Loureiro N. F., Boldyrev S., 2017, ApJ, 850, 182

Lu R.-S., Broderick A. E., Baron F., Monnier J. D., Fish V. L., Doeleman S. S., Pankratius V., 2014, ApJ, 788, 120

Lu R.-S., Krichbaum T. P., Roy A. L., Fish V. L., Doeleman S. S., Johnson M. D., Akiyama K., et al., 2018, ApJ, 859, 60

Luminet J.-P., 1979, A&A, 75, 228

Ly C., Walker R. C., Junor W., 2007, ApJ, 660, 200

Lynden-Bell D., 1969, Nature, 223, 690

Mahadevan R., 1998, Nature, 394, 651

Mahadevan R., Quataert E., 1997, ApJ, 490, 605

Mallet A., Schekochihin A. A., Chandran B. D. G., 2017, MNRAS, 468, 4862

Manolakou K., Horns D., Kirk J. G., 2007, A&A, 474, 689

Mao S. A., Dexter J., Quataert E., 2016, MNRAS

Marrone D. P., Moran J. M., Zhao J.-H., Rao R., 2007, ApJL, 654, L57

Marrone D. P., et al., 2008, ApJ, 682, 373

McKinney J. C., 2006, MNRAS, 368, 1561

McKinney J. C., Blandford R. D., 2009, MNRAS, 394, L126

McKinney J. C., Tchekhovskoy A., Blandford R. D., 2012, MNRAS, 423, 3083

McKinney J. C., Tchekhovskoy A., Sądowski A., Narayan R., 2014, MNRAS, 441, 3177

Mei S., et al., 2007, ApJ, 655, 144

Mertens F., Lobanov A. P., Walker R. C., Hardee P. E., 2016, A&A, 595, A54

Mihalas D., Mihalas B. W., 1984, Foundations of Radiation Hydrodynamics. Oxford University Press, New York

Moderski R., Sikora M., Coppi P. S., Aharonian F., 2005, MNRAS, 363, 954

Mościbrodzka M., Falcke H., 2013, A&A, 559, L3

Mościbrodzka M., Gammie C. F., Dolence J. C., Shiokawa H., Leung P. K., 2009, ApJ, 706, 497

Mościbrodzka M., Gammie C. F., Dolence J. C., Shiokawa H., 2011, ApJ, 735, 9

Mościbrodzka M., Falcke H., Shiokawa H., Gammie C. F., 2014, A&A, 570, A7

Mościbrodzka M., Falcke H., Shiokawa H., 2016a, A&A, 586, A38

Mościbrodzka M., Falcke H., Shiokawa H., 2016b, A&A, 586, A38

Mościbrodzka M., Dexter J., Davelaar J., Falcke H., 2017, MNRAS, 468, 2214

Narayan R., Nityananda R., 1986, ARA&A, 24, 127

Narayan R., Yi I., 1994, ApJL, 428, L13

Narayan R., Yi I., 1995a, ApJ, 444, 231

Narayan R., Yi I., 1995b, ApJ, 452, 710

Narayan R., Yi I., Mahadevan R., 1995, Nature, 374, 623

Narayan R., Mahadevan R., Grindlay J. E., Popham R. G., Gammie C., 1998, ApJ, 492, 554

Narayan R., Igumenshchev I. V., Abramowicz M. A., 2003, PASJ, 55, L69

Narayan R., Sądowski A., Penna R. F., Kulkarni A. K., 2012, MNRAS, 426, 3241

Narayan R., Zhu Y., Psaltis D., Sądowski A., 2016, MNRAS, 457, 608

Neilsen J., et al., 2013, ApJ, 774, 42

Nityananda R., Narayan R., 1983, A&A, 118, 194

Noble S. C., Krolik J. H., Schnittman J. D., Hawley J. F., 2011, ApJ, 743, 115

Novikov I. D., Thorne K. S., 1973, in Dewitt C., Dewitt B. S., eds, Black Holes (Les Astres Occlus). pp 343–450

Numata R., Loureiro N. F., 2015, Journal of Plasma Physics, 81, 305810201

Oliphant T. E., 2007, Computing in Science & Engineering, 9, 10

Oppenheimer J. R., Volkoff G. M., 1939, Physical Review, 55, 374

Owen F. N., Eilek J. A., Kassim N. E., 2000, ApJ, 543, 611

Özel F., Psaltis D., Narayan R., 2000, ApJ, 541, 234

Palmer H. P., Rowson B., Anderson B., Donaldson W., Miley G. K., 1967, Nature, 213, 789

Palumbo D., et al., 2019, in prep.

Pandya A., Zhang Z., Chandra M., Gammie C. F., 2016, ApJ, 822, 34

Paper II, see The Event Horizon Telescope Collaboration et al. (2019b)

Paper III, see The Event Horizon Telescope Collaboration et al. (2019c)

Paper IV, see The Event Horizon Telescope Collaboration et al. (2019d)

Paper V, see The Event Horizon Telescope Collaboration et al. (2019e)

Paper VI, see The Event Horizon Telescope Collaboration et al. (2019f)

Pearson T. J., Readhead A. C. S., 1984, ARA&A, 22, 97

Penna R. F., McKinney J. C., Narayan R., Tchekhovskoy A., Shafee R., McClintock J. E., 2010, MNRAS, 408, 752

Penna R. F., Kulkarni A., Narayan R., 2013, A&A, 559, A116

Penrose R., 1965, Phys. Rev. Lett., 14, 57

Perlman E. S., Sparks W. B., Radomski J., Packham C., Fisher R. S., Piña R., Biretta J. A., 2001, ApJL, 561, L51

Phan T. D., et al., 2013, Geophysical Research Letters, 40, 4475

Phan T. D. Drake J. F., Shay M. A., Gosling J. T., Paschmann G., Eastwood J. P., Oieroset M., Fujimoto M., Angelopoulos M., 2014, Geophysical Research Letters, 41, 7002

Ponsonby J. E. B., 1973, MNRAS, 163, 369

Ponti G., et al., 2017, MNRAS, 468, 2447

Porquet D., et al., 2008, A&A, 488, 549

Porth O., Olivares H., Mizuno Y., Younsi Z., Rezzolla L., Mościbrodzka M., Falcke H., Kramer M., 2017, Computational Astrophysics and Cosmology, 4, 1

Prieto M. A., Fernández-Ontiveros J. A., Markoff S., Espada D., González-Martín O., 2016, MNRAS, 457, 3801

Psaltis D., Narayan R., Fish V. L., Broderick A. E., Loeb A., Doeleman S. S., 2015, ApJ, 798, 15

Quataert E., 1998, ApJ, 500, 978

Quataert E., Narayan R., 1999, ApJ, 520, 298

Readhead A. C. S., Wilkinson P. N., 1978, ApJ, 223, 25

Readhead A. C. S., Walker R. C., Pearson T. J., Cohen M. H., 1980, Nature, 285, 137

Rees M. J., 1984, ARA&A, 22, 471

Rees M. J., Begelman M. C., Blandford R. D., Phinney E. S., 1982, Nature, 295, 17

Reid M. J., Schmitt J. H. M. M., Owen F. N., Booth R. S., Wilkinson P. N., Shaffer D. B., Johnston K. J., Hardee P. E., 1982, ApJ, 263, 615

Remillard R. A., McClintock J. E., 2006, ARA&A, 44, 49

Ressler S. M., Tchekhovskoy A., Quataert E., Chandra M., Gammie C. F., 2015, MNRAS, 454, 1848

Ressler S. M., Tchekhovskoy A., Quataert E., Gammie C. F., 2017, MNRAS, 467, 3604

Reynolds C. S., Fabian A. C., Celotti A., Rees M. J., 1996, MNRAS, 283, 873

Roberts D. H., Wardle J. F. C., Brown L. F., 1994, ApJ, 427, 718

Roedig C., Zanotti O., Alic D., 2012, MNRAS, 426, 1613

Rogers A. E. E., et al., 1974, ApJ, 193, 293

Rogers A. E. E., Doeleman S. S., Moran J. M., 1995, AJ, 109, 1391

Rowan M., Sironi L., Narayan R., 2017, ApJ, 850, 29

Rowan M. E., Sironi L., Narayan R., 2019, ApJ, 873, 2

Rudin L. I., Osher S., Fatemi E., 1992, Physica D, 60, 256

Ryan B. R., Dolence J. C., Gammie C. F., 2015, ApJ, 807, 31

Ryan B. R., Ressler S. M., Dolence J. C., Tchekhovskoy A., Gammie C., Quataert E., 2017, ApJL, 844, L24

Ryan B. R., Ressler S. M., Dolence J. C., Gammie C., Quataert E., 2018, ApJ, 864, 126

Rybicki G. B., Lightman A. P., 1979, Radiative processes in astrophysics. Wiley-Interscience, New York, doi:10.1002/9783527618170

Sądowski A., Narayan R., 2015, MNRAS, 454, 2372

Sądowski A., Narayan R., Tchekhovskoy A., Zhu Y., 2013a, MNRAS, 429, 3533

Sądowski A., Narayan R., Penna R., Zhu Y., 2013b, MNRAS, 436, 3856

Sądowski A., Narayan R., McKinney J. C., Tchekhovskoy A., 2014, MNRAS, 439, 503

Sądowski A., Narayan R., Tchekhovskoy A., Abarca D., Zhu Y., McKinney J. C., 2015, MNRAS, 447, 49

Sądowski A., Wielgus M., Narayan R., Abarca D., McKinney J. C., Chael A., 2017, MNRAS, 466, 705

Salpeter E. E., 1964, ApJ, 140, 796

Schmidt M., 1963, Nature, 197, 1040

Schödel R., Eckart A., Mužić K., Meyer L., Viehmann T., Bower G. C., 2007, A&A, 462, L1

Schödel R., Morris M. R., Muzic K., Alberdi A., Meyer L., Eckart A., Gezari D. Y., 2011, A&A, 532, A83

Schwab F. R., 1980, in Rhodes W. T., ed., Proc. SPIEVol. 231, 1980 International Optical Computing Conference I. pp 18–25, doi:10.1117/12.958828

Schwarzschild K., 1916, Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften (Berlin), 1916, Seite 189-196,

Shafee R., McKinney J. C., Narayan R., Tchekhovskoy A., Gammie C. F., McClintock J. E., 2008, ApJL, 687, L25

Shakura N. I., Sunyaev R. A., 1973, A&A, 24, 337

Shay M. A., Haggerty C. C., Matthaeus W. H., Parashar T. N., Wan M., Wu P., 2018, Physics of Plasmas, 25, 012304

Shcherbakov R. V., Penna R. F., McKinney J. C., 2012, ApJ, 755, 133

Shepherd M. C., 1997, in Hunt G., Payne H., eds, Astronomical Society of the Pacific Conference Series Vol. 125, Astronomical Data Analysis Software and Systems VI. p. 77

Shiokawa H., 2013, PhD thesis, University of Illinois at Urbana-Champaign

Sironi L., Spitkovsky A., 2011, ApJ, 726, 75

Sironi L., Spitkovsky A., 2014, ApJL, 783, L21

Stawarz Ł., Aharonian F., Kataoka J., Ostrowski M., Siemiginowska A., Sikora M., 2006, MNRAS, 370, 981

Stepney S., Guilbert P. W., 1983, MNRAS, 204, 1269

TMS, see Thompson et al. (2017)

Takahashi R., 2004, ApJ, 611, 996

Tchekhovskoy A., Narayan R., McKinney J. C., 2010, ApJ, 711, 50

Tchekhovskoy A., Narayan R., McKinney J. C., 2011, MNRAS, 418, L79

The Event Horizon Telescope Collaboration et al., 2019a, ApJL, 875, L1

The Event Horizon Telescope Collaboration et al., 2019b, ApJL, 875, L2

The Event Horizon Telescope Collaboration et al., 2019c, ApJL, 875, L3

The Event Horizon Telescope Collaboration et al., 2019d, ApJL, 875, L4

The Event Horizon Telescope Collaboration et al., 2019e, ApJL, 875, L5

The Event Horizon Telescope Collaboration et al., 2019f, ApJL, 875, L6

Thiébaut É., 2013, in Mary D., Theys C., Aime C., eds, EAS Publications Series Vol. 59, EAS Publications Series. pp 157–187, doi:10.1051/eas/1359009

Thiébaut É., Young J., 2017, Journal of the Optical Society of America A, 34, 904

Thompson A. R., Moran J. M., Swenson Jr. G. W., 2017, Interferometry and Synthesis in Radio Astronomy, 3rd Edition. Springer, doi:10.1007/978-3-319-44431-4

Twiss R. Q., Carter A. W. L., Little A. G., 1960, The Observatory, 80, 153

de Villiers J.-P., Hawley J. F., Krolik J. H., 2003, ApJ, 599, 1238

Wakker B. P., Schwarz U. J., 1988, A&A, 200, 312

Walker R. C., Hardee P. E., Davies F., Ly C., Junor W., Mertens F., Lobanov A., 2016, Galaxies, 4, 46

Walker R. C., Hardee P. E., Davies F. B., Ly C., Junor W., 2018, ApJ, 855, 128

Walsh J. L., Barth A. J., Ho L. C., Sarzi M., 2013, ApJ, 770, 86

Walt S. v. d., Colbert S. C., Varoquaux G., 2011, Computing in Science & Engineering, 13, 22

Webb G. M., 1985, ApJ, 296, 319

Webb G. M., 1989, ApJ, 340, 1112

Webster B. L., Murdin P., 1972, Nature, 235, 37

Werner G. R., Uzdensky D. A., Begelman M. C., Cerutti B., Nalewajko K., 2018, MNRAS, 473, 4840

White C. J., Stone J. M., Gammie C. F., 2016, ApJS, 225, 22

Whysong D., Antonucci R., 2004, ApJ, 602, 116

Wilkinson P. N., Readhead A. C. S., Purcell G. H., Anderson B., 1977, Nature, 269, 764

Witzel G., et al., 2012, ApJS, 203, 18

Yamada M., Yoo J., Jara-Almonte J., Ji H., Kulsrud R. M., Myers C. E., 2014, Nature communications, 5, 4774

Yuan F., Narayan R., 2014, ARA&A, 52, 529

Yuan F., Markoff S., Falcke H., 2002, A&A, 383, 854

Yuan F., Quataert E., Narayan R., 2003, ApJ, 598, 301

Yuan F., Quataert E., Narayan R., 2004, ApJ, 606, 894

Yusef-Zadeh F., et al., 2006, ApJ, 644, 198

Yusef-Zadeh F., Bushouse H., Wardle M., Heinke C., Roberts D. A., Dowell C. D., Brunthaler A., et al., 2009, ApJ, 706, 348

Zamaninasab M., Clausen-Brown E., Savolainen T., Tchekhovskoy A., 2014, Nature, 510, 126

Zdziarski A. A., Sikora M., Pjanka P., Tchekhovskoy A., 2015, MNRAS, 451, 927

Zernike F., 1938, Physica, 5, 785

Zhang S., Baganoff F. K., Ponti G., Neilsen J., Tomsick J. A., Dexter J., et al., 2017, ApJ, 843, 96

Zhu Y., Narayan R., Sądowski A., Psaltis D., 2015, MNRAS, 451, 1661