Black hole parameter estimation with synthetic very long baseline interferometry data from the ground and from space

Freek Roelofs\textsuperscript{1,2}, Christian M. Fromm\textsuperscript{3,4,5}, Yosuke Mizuno\textsuperscript{6,4}, Jordy Davelaar\textsuperscript{7,8,1}, Michael Janssen\textsuperscript{4,1}, Ziri Younsi\textsuperscript{9,4}, Luciano Rezzolla\textsuperscript{4,10,11}, and Heino Falcke\textsuperscript{1}

\textsuperscript{1} Department of Astrophysics, Institute for Mathematics, Astrophysics and Particle Physics (IMAPP), Radboud University, PO Box 9010, 6500 GL Nijmegen, The Netherlands
\textsuperscript{2} Center for Astrophysics | Harvard & Smithsonian, 60 Garden Street, Cambridge, MA 02138, USA
\textsuperscript{3} Black Hole Initiative at Harvard University, 20 Garden Street, Cambridge, MA 02138, USA
\textsuperscript{4} Institut für Theoretische Physik, Goethe-Universität Frankfurt, Max-von-Laue-Straße 1, 60438 Frankfurt am Main, Germany
\textsuperscript{5} Max-Plank-Institut für Radioastronomie, Auf dem Hügel 69, 53121 Bonn, Germany
\textsuperscript{6} Tsung-Dao Lee Institute and School of Physics and Astronomy, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, PR China
\textsuperscript{7} Department of Astronomy and Columbia Astrophysics Laboratory, Columbia University, 550 W 120th St., New York, NY 10027, USA
\textsuperscript{8} Center for Computational Astrophysics, Flatiron Institute, 162 Fifth Avenue, New York, NY 10010, USA
\textsuperscript{9} Mullard Space Science Laboratory, University College London, Holmbury St. Mary, Dorking, Surrey RH5 6NT, UK
\textsuperscript{10} School of Mathematics, Trinity College, Dublin 2, Ireland
\textsuperscript{11} Frankfurt Institute for Advanced Studies, Ruth-Moufang-Straße 1, 60438 Frankfurt, Germany

Received 22 October 2020 / Accepted 19 March 2021

ABSTRACT

Context. The Event Horizon Telescope (EHT) has imaged the shadow of the supermassive black hole in M\textsubscript{87}. A library of general relativistic magnetohydrodynamics (GRMHD) models was fit to the observational data, providing constraints on black hole parameters.

Aims. We investigate how much better future experiments can realistically constrain these parameters and test theories of gravity.

Methods. We generated realistic synthetic 230\,GHz data from representative input models taken from a GRMHD image library for M\textsubscript{87}, using the 2017, 2021, and an expanded EHT array. The synthetic data were run through an automated data reduction pipeline used by the EHT. Additionally, we simulated observations at 230, 557, and 690\,GHz with the Event Horizon Imager (EHI) Space VLBI concept. Using one of the EHT parameter estimation pipelines, we fit the GRMHD library images to the synthetic data and investigated how the black hole parameter estimations are affected by different arrays and repeated observations.

Results. Repeated observations play an important role in constraining black hole and accretion parameters as the varying source structure is averaged out. A modest expansion of the EHT already leads to stronger parameter constraints in our simulations. High-frequency observations from space with the EHI rule out all but ~15\% of the GRMHD models in our library, strongly constraining the magnetic flux and black hole spin. The 1\,\sigma constraints on the black hole mass improve by a factor of five with repeated high-frequency space array observations as compared to observations with the current ground array. If the black hole spin, magnetization, and electron temperature distribution can be independently constrained, the shadow size for a given black hole mass can be tested to ~0.5\% with the EHI space array, which allows tests of deviations from general relativity. With such a measurement, high-precision tests of the Kerr metric become within reach from observations of the Galactic Center black hole Sagittarius A*.

Key words. galaxies: nuclei – black hole physics – telescopes – atmospheric effects – techniques: high angular resolution – techniques: interferometric

1. Introduction

The Event Horizon Telescope (EHT) is a millimeter very long baseline interferometry (VLBI) array imaging supermassive black holes on event horizon scales (Event Horizon Telescope Collaboration 2019b). In April 2019, the Event Horizon Telescope Collaboration (EHTC) published its first results from 230\,GHz observations of the supermassive black hole in M\textsubscript{87} conducted in 2017 (Event Horizon Telescope Collaboration 2019a,b,c,d,e,f). The source was imaged as an asymmetric ring with a diameter of 42 ± 3\,mas that is brighter in the south than in the north. This ring structure is interpreted as the black hole “shadow” (Falcke et al. 2000), which is formed by gravitational lensing of synchrotron photons emitted by the near-horizon accretion flow plasma or relativistic jet. The observed shadow is largely dominated by properties of the black hole itself, but also affected by the astrophysics of the emitting region. Hence, simulations are used to disentangle these effects. The structure is asymmetric due to Doppler boosting of emission from plasma moving toward us on the southern side of the ring, which is in agreement with the orientation of the jet seen on larger scales for a clockwise rotation of the accretion flow. Through the combination of fitting a library of general relativistic magnetohydrodynamics (GRMHD) simulations and a crescent model to the interferometric data, and fitting a ring to the reconstructed images, a mass measurement of $M = 6.5_{-0.2}^{+0.7}\,\text{sys}\times10^6\,M_\odot$ could be
made, which is consistent with earlier measurements from stellar dynamics in M87 (Gebhardt et al. 2011).

Constraining black hole and accretion parameters is important for several reasons. Accurate measurements of the mass and spin of the black hole would together determine the Kerr metric, describing the spacetime near the black hole in general relativity. In the Kerr metric, the size and shape of the black hole shadow are completely determined by the mass and spin combined with the distance to the source. If an independent mass measurement is made, the accuracy of deriving the black hole mass from VLBI imaging data can be used as a proxy for testing general relativity. Especially high-resolution VLBI observations of the Galactic Center black hole Sagittarius A* (Sgr A*) could provide a unique opportunity to test the Kerr metric (e.g., Psaltis et al. 2015), since its mass is known accurately (at the sub-percent level) and independently from monitoring stellar orbits. The mass of Sgr A* was measured to be $(4.148 \pm 0.014) \times 10^6 M_\odot$ by Do et al. (2019) and $(3.964 \pm 0.047_{\text{stat}} \pm 0.026_{\text{syst}}) \times 10^6 M_\odot$ by GRAVITY Collaboration (2019). With an expected shadow size of 51 µas, Sgr A* is the other important target for black hole imaging by the EHT.

Apart from measuring and testing the Kerr metric, high-resolution VLBI observations also provide the opportunity to establish the near-horizon plasma properties and behavior. Potentially measurable quantities include the plasma magnetization and the electron temperature distribution. The black hole spin is important here as well, as it determines the plasma orbits. The black hole spin and plasma properties play an important role in the launching and collimation of relativistic jets as seen in M87 and potentially in Sgr A*, which is currently not clearly understood. For instance, in the Blandford-Znajek mechanism (Blandford & Znajek 1977), the jet launching energy is extracted from the black hole as its spin twists magnetic field lines; whereas, in the Blandford-Payne mechanism (Blandford & Payne 1982), the accretion disk is the dominant source of jet launching energy. While the black hole spin, plasma magnetization, and electron temperature distribution could not be determined from the analysis by Event Horizon Telescope Collaboration (2019a,b,c,d,e,f), an enhanced EHT array may provide the necessary resolution and fidelity to start constraining these parameters. Analysis of the 2017 polarization data strongly favors magnetically arrested disks (Event Horizon Telescope Collaboration 2021a,b, see also Sect. 2.1).

Apart from extensions of the EHT on the ground, several studies have been done on the employment of space-based antennas for high-resolution observations. The baseline lengths attainable from Earth are limited by its size. Also, a ground-based imaging array at frequencies higher than about 345 GHz is extremely challenging due to strong attenuation and turbulence introduced by water vapor in the troposphere. A space-based array could overcome these limitations and potentially increase the array resolution by an order of magnitude (Roelofs et al. 2019). However, image comparisons are challenging to quantify except for simple metrics evaluating pixel-by-pixel differences or cross-correlations, which do not always match a quality assessment by eye where a person looks at the reconstruction of certain model features, such as a photon ring or extended jet. Furthermore, because of the sparsely sampled in-plane, image reconstruction generally requires additional assumptions in the form of, for example, regularization by image smoothness or sparsity (Event Horizon Telescope Collaboration 2019d). In this work, we aim to overcome both of these limitations by evaluating the current and future EHT and EHI array performance based on model parameters that are estimated from synthetic visibility data directly.

There are two main motivations for testing array performance under varying observational circumstances. For the near future, studies such as these are required for setting a set of triggering requirements for the EHT. During an EHT campaign, which typically lasts two weeks, a few observing nights can be triggered based on the technical readiness of the telescopes and weather conditions at the different sites. These triggering decisions could be optimized by quantifying the effect of losing certain stations or varying weather conditions on the science output.

The second motivation for this project and main focus of this paper is for the longer term: By adding new stations to the synthetic observations at different sites, their effect on the science output can be measured directly. Such a procedure can help to optimize the design of the future EHT and other VLBI arrays, and establish how accurately parameters can be measured and ultimately how accurately general relativity and models for black hole accretion and jet launching can be tested.

In this paper, we investigate the potential of current and future EHT and EHI arrays to constrain black hole and accretion parameters. To this end, we employed the GRMHD library model fitting framework of Event Horizon Telescope Collaboration (2019e,f) to fit the black hole mass and spin, electron temperature distribution, plasma magnetization, and sky orientation for a suite of realistic EHT and EHI synthetic data sets. In Sect. 2, we present our model image generation, synthetic data generation, and model fitting framework. In Sect. 3, we show how the recovered parameter constraints are affected by the array and observing strategy used. We summarize and provide suggestions for next steps in Sect. 4.

2. Simulation and parameter estimation methods
2.1. Model image generation
2.1.1. The EHT GRMHD model image library

The data and images from the EHT 2017 campaign were compared to a library of ray-traced GRMHD simulations (Event Horizon Telescope Collaboration 2019e,f). In a GRMHD simulation, the plasma dynamics near the event horizon are simulated from a set of initial conditions by evolving the plasma according to the laws of magnetohydrodynamics embedded in, for example, the Kerr metric as described by general relativity (e.g., Gammie et al. 2003; Mościbrodzka et al. 2009; Porth et al. 2017, 2019). The resulting plasma quantities, such as the density, magnetic field, and temperature, are then used to calculate synchrotron emissivities and absorptivities, and photons are...
ray-traced along the geodesics through the simulation domain to a distant observer to form an image at a specific frequency.

The images in the EHT M87 GRMHD library depend on several parameters. At the GRMHD stage, there is a distinction between standard and normal evolution (SANE, Narayan et al. 2012; Sadowski et al. 2013) models, which have a relatively low magnetic flux crossing a hemisphere of the event horizon, and magnetically arrested disk (MAD, Igumenshchev et al. 2003; Narayan et al. 2003) models, which are characterized by a high magnetic flux. Intermediate magnetic fluxes (INSANE) are possible as well, but these were not included in the analysis by Event Horizon Telescope Collaboration (2019e,f). The dimensionless black hole spin parameter $a_*$ can be set to values between $-1$ and $1$, where negative and positive values represent a retrograde and prograde spin, respectively, with respect to the rotation of the outer large-scale accretion flow.

At the ray-tracing stage, a choice needs to be made for the ratio of the ion temperature $T_i$, which follows from the GRMHD simulation, and the electron temperature $T_e$, which characterizes the emission. In Event Horizon Telescope Collaboration (2019e), the model from Mościbrodzka et al. (2016) was adopted, where

$$\frac{T_i}{T_e} = \frac{R_{\text{high}}}{1 + \beta_p^2} + \frac{1}{1 + \beta_p^2}. \quad (1)$$

Here, $\beta_p$ is the gas-to-magnetic pressure ratio, which is large in the disk and small in the strongly magnetized jet. It follows from the GRMHD simulation, so that the free parameter $R_{\text{high}}$ fixes the electron temperature. An increased value of $R_{\text{high}}$ represents a larger ion-to-electron temperature ratio in the disk, so that emission in that region is suppressed and the resulting ray-traced image becomes more jet-dominated. A thermal electron temperature parametrization is justified here, as the inclusion of nonthermal particles is mostly relevant for high-energy emission. The appearance of the black hole shadow depends only weakly on the inclusion of nonthermal particles (Davelaar et al. 2019). Other ray-tracing parameters that need to be set are the inclination angle of the line of sight with respect to the black hole spin axis, the plasma mass unit, which sets the average total flux density of the model, the black hole mass, and the emission frequency.

2.1.2. GRMHD images used in this work

Our ray-traced images, which were used as input for the synthetic data generation and for fitting the synthetic data against, were generated from the BIAC (Porth et al. 2017) EHT GRMHD library using the general-relativistic ray-tracing radiative transfer code BioSS (Younsi et al. 2012, 2016, and in prep.). We used MAD and SANE models with spins $-0.94$, $-0.5$, $0$, $0.5$, $0.94$ (MAD), and $0.97$ (SANE). Once the accretion rate had reached a steady state at $t > 6000 \text{GM}/c^3$ for SANE models and at $t > 12000 \text{GM}/c^3$ for MAD models, we computed the emission. Following Event Horizon Telescope Collaboration (2019e,c), $R_{\text{high}}$ was set to $1, 10, 20, 40, 80, \text{or} 160$. During the radiative transfer calculations, we excluded regions in the GRMHD simulations with magnetization (ratio of magnetic and kinetic energy density) $\sigma > 1$. These regions, mainly found in the funnel, are typically subject to numerical uncertainties which could lead to spurious features in the final images. We use a black hole mass $6.2 \times 10^9 M_\odot$ and a distance of 16.9 Mpc. For the flux density normalization, we selected the last GRMHD snapshot of each model and iterated the accretion rate until a flux density of 0.8 Jy at 230 GHz was obtained. The inclination angle was fixed at $163^\circ$, as motivated by the large-scale orientation of the M87 jet and the fact that the 230 GHz data favor an image that is brighter in the south\(^1\) (Walker et al. 2018; Event Horizon Telescope Collaboration 2019d,e). If the inclination angle is not known a priori from other measurements, which is the case for Sgr A*, it should be added as an additional GRMHD parameter.

For each combination of magnetic flux, spin, and $R_{\text{high}}$, 100 frames were generated at a cadence of $10 \text{GM}/c^3$, at frequencies of 230, 557, and 690 GHz. The image field of view was set to 160 $\mu$m, with a pixel size of 1 $\mu$m. At the higher frequencies, the emission is optically thin and the average total flux density decreases from 0.8 Jy at 230 GHz to 0.42 and 0.33 Jy at 557 and 690 GHz, respectively.

2.2. Synthetic data generation

2.2.1. EHT simulations

The EHT synthetic data generation for this work was performed with SYMBA (Roelofs et al. 2020). This pipeline mimics the full VLBI observation process, including realistic observation and calibration effects. Raw frequency-resolved synthetic data are generated following real observation schedules with MeqSilhouette (Blecher et al. 2017; Natarajan et al., in prep.). Added data corruptions include an atmospheric model which introduces delays, phase turbulence, amplitude attenuation, and sky noise. Furthermore, the effects of antenna pointing offsets due to atmospheric seeing, wind wobbling the dish, and inaccurate pointing solutions are simulated. Fixed antenna gain offsets and polarimetric leakage are added as well. The raw synthetic data is then passed through the VLBI data calibration pipeline rPICARD (Janssen et al. 2019), which is used to calibrate real EHT data as well (Event Horizon Telescope Collaboration 2019e,c). The calibrated data sets can then be used for further analysis, such as image reconstruction or parameter estimation.

The synthetic data generation setup, which includes the locations and properties of the antennas and the local weather conditions during the observations, was set equal to the setup used for the EHT2017 and EHT2020 arrays in Roelofs et al. (2020), where the latter was renamed to the EHT2021 array as the EHT2020 observations were cancelled. The EHT2017 coverage is identical to that of 11 April, where scans that were scheduled but not observed were flagged. For the simulation of atmospheric corruptions, SYMBA takes the ground pressure, ground temperature, precipitable water vapor, and coherence time at each site as input. The input antenna and weather parameters were based on site measurements taken during the EHT2017 campaign or estimated from the EHT2017 data. For new stations that did not participate in 2017, weather parameters were estimated using climatological modeling (Paine 2019), assuming decent observing conditions in April. See Roelofs et al. (2020) for more details and the values used for the different observation parameters.

We also simulated an array configuration with six additional stations added to the EHT2021 array. These are the Africa Millimeter Telescope, which is planned to be built on the Gamsberg in Namibia (Backes et al. 2016), the Haystack Observatory in Massachusetts, and potential new sites on the summit of the

\(^1\) Due to frame dragging, the plasma in the innermost accretion flow is forced to rotate with the black hole spin, even if the orbits are retrograde further out. Given that the approaching jet of M87 extends toward the northwest, the Doppler-boosted bright region in the south thus implies that the black hole spin axis is pointed away from us.
Greenland ice sheet, in La Palma in Spain, in Río Negrá in Argentina, and on Pikes Peak in Colorado. These sites would be suitable for M 87 observations based on the atmosphere. Longer baselines have thus not been included in the simulations, even though the satellites can be further apart in these orbits. For synthetic data generated for this work, we adopted the three-satellite setup with 4 m antenna diameters and system noise parameters from Roelofs et al. (2019). The model visibilities were calculated with the eht-imaging library (Chael et al. 2016, 2018). The \( uv \)-coverage of the EHI at 230, 557, and 690 GHz is shown in Fig. 2.

2.3. Parameter estimation

2.3.1. The EHT GRMHD library parameter estimation framework

In Event Horizon Telescope Collaboration (2019e,f), the GRMHD library was scored against the EHT 2017 data by rotating and scaling (in size and total flux density) each model image frame to provide the best possible fit to the EHT data. The size scaling is of particular importance here since it sets the angular size of one gravitational radius \( GM/D^2 \). Combined with distance measurements of M 87, the size scaling then gives a best-fit black hole mass for each model frame. The scoring was performed and vetted with two independent pipelines: the MCMC parameter estimation framework Themis (Broderick et al. 2020a) and the genetic algorithm GENA (Fromm et al. 2019), which is used in this work.

Due to variability from plasma orbits and turbulence, different GRMHD frames of the same movie may appear significantly different on the scales probed by the EHT. Therefore, synthetic EHT data generated from a frame of that model do not necessarily give a formally acceptable fit (\( \chi^2 > 1 \)) to the other movie frames. To account for this model variance, Event Horizon Telescope Collaboration (2019e,f) developed the method of average image scoring (AIS). In this procedure, synthetic EHT data generated from the average image of a GRMHD movie are scored against the individual frames of the same model. The resulting \( \chi^2 \) distribution is then compared to the \( \chi^2 \) obtained from fitting the average model image to the observed data. A \( p \)-value is then computed as

\[
p_{\text{AIS}}(M|\mathcal{D}) = \frac{N_{\chi^2_{\text{obs}} < \chi^2_{\text{med}}}}{N}
\]

(2)

Here, \( M \) is a GRMHD model, \( D \) is the observed data, and \( N \) is the number of frames in the model; \( N_{\chi^2_{\text{obs}} < \chi^2_{\text{med}}} \) is the number of frames in the model for which the difference between the reduced \( \chi^2 \) of the frame scored against synthetic data from the average model and the median \( \chi^2 \) of all frames is larger than the difference between the reduced \( \chi^2 \) of the average model scored against the observed data and the median \( \chi^2 \) of the average model. Equation (2) is effectively computing a two-sided \( p \)-value (see for details Event Horizon Telescope Collaboration 2019f). The model is rejected when \( p_{\text{AIS}} < 0.01 \).

Using this method, MAD models with \( a_s = -0.94 \) formed the only class of models that could be ruled out based on the EHT2017 data alone. With the addition of other criteria, such as...
the X-ray flux and jet power as predicted by the models compared to other observational data, the number of fitting models was reduced significantly. The remaining models, combined with a distance measurement of \( D = 16.8^{\pm 0.8} \text{Mpc} \) and mass estimates from a crescent model fit to the data and from a ring fit to the reconstructed image, gave the measured mass of \( M = 6.5 \pm 0.2 \text{[stat]} + 0.7 \text{[sys]} \times 10^6 \text{M}_\odot \).

2.3.2. GENA pipeline description

In the following, we provide a short description of the mathematical and numerical methods used within GENA, which is the genetic algorithm pipeline we used for the fitting of the GRMHD model library to our input synthetic data.

**Constrained nonlinear optimization.** The extraction of physical parameters from observational data via numerical, semi-analytical, or geometrical models can be considered as a constrained non-linear optimization problem. Its mathematical formulation is given by the following:

\[
\min f(x) \\
\text{subject to } g_j(x) \leq 0, \quad j = 1, \ldots, n, \\
x_{L,i} \leq x_i \leq x_{R,i}, \quad i = 1, \ldots, m,
\]

where \( x \) is an \( m \)-dimensional vector including the model parameters, \( f(x) \) is the cost function (also objective or minimization function), \( g_j(x) \) are the constraints, and \( x_{L,i} \) and \( x_{R,i} \) are lower and upper boundaries for the model parameters.

The cost function and the constraints should be constructed in such a way that key properties of the data and prior knowledge of the source, such as the mass of the black hole, are used to guide and speed up the convergence of the optimization process.

**Optimization algorithm.** The above stated optimization problem can be solved by several kinds of algorithms: gradient-based, gradient-free, or MCMC algorithms (see Themis, Broderick et al. 2020a). Given the high computational effort in scoring the GRRT images, which includes Fourier transforms and gain optimizations, as well as the fast convergence of gradient-free in contrast to gradient-based algorithms (a gradient-based algorithm requires a lot of computational resources for mapping out the gradient with sufficient resolution), we decided to apply a gradient-free search algorithm in the form of a population-based evolutionary algorithm. The basic idea and steps of an evolutionary algorithm (EA) are described in the next paragraph.

**Evolutionary algorithm (EA).** EAs are motivated by biological evolution and follow the principle of survival of the fittest. The main steps of EAs are selection, crossover (mating), and mutation of the individuals for several generations. An individual can be seen as a set of parameters, in our case the orientation of the image in the sky, \( \phi \), the flux scaling, \( f \), and the mass of the black hole, \( M \). We note that \( x \) can thus be written as \( x = [\phi, f, M]^T \). Each entry, for example, \( \phi \), is typically labeled as a gene. During the initial step, \( N \) random individuals are created in such a way that they cover the parameter space (e.g., using the Latin hypercube sampling), and their fitness is computed using \( f(x) \).

For the creation of the next generation, offsprings are created from the current population, including random mutations to ensure diversity in the population so that the algorithm does not get stuck in local minima. The details of this step depend on the implementation of the EA. In this work, we use the Non Sorting Genetic Algorithm II (NSGA2; Deb et al. 2002) and the differential evolution algorithm following Storm & Price (1997). In this algorithm, one member of the population, \( x_1 \), is selected together with three additional randomly selected population members labeled \( x_2, x_3 \), and \( x_4 \). From these three random members, a new member is created, which is referred to as the mutated member \( x_{\text{mut}} = x_1 + F(x_2 - x_3) \), where \( F \) is a constant factor in the interval \([0,2] \). A new trial population member \( x_{\text{trial}} \) is then created from crossover between \( x_1 \) and \( x_{\text{mut}} \). During the crossover, a random number in the interval \([0,1] \) is drawn for each component \( (\phi, f \) and \( M) \). If this number is larger than the crossover probability, the component from \( x_{\text{mut}} \) is copied over to \( x_{\text{trial}} \). If not, the component from \( x_1 \) is taken as a new component for \( x_{\text{trial}} \). After the crossover step, \( x_{\text{trial}} \) thus includes components from the initial member \( x_1 \) and from the mutated member \( x_{\text{mut}} \). In the last step, \( x_{\text{trial}} \) is evaluated with respect to the cost function \( f(x) \). If \( f(x_{\text{trial}}) < f(x_1) \), \( x_1 \) is replaced by \( x_{\text{trial}} \). If not, \( x_1 \) is kept in the next generation. This procedure is repeated for each member of the population so that finally a new generation is created.

New generations are created up to a maximum number of iterations or until a convergence criterion is reached, that is when the standard deviation \( \sigma_f \) of \( f(x) \) across the population is smaller than a specified fraction of the mean \( f(x) \). The final fitted parameters are then the components of the \( x \) for which \( f(x) \) is smallest. In this work, we use a population size of \( 20 \), a crossover probability of \( 0.6 \), we varied \( F \) randomly between different generations in the interval \([0.7,1] \) (a process called dithering, which speeds up the convergence), and the process was stopped after 100 generations or until \( \sigma_f < 0.02f(x) \).

**Application to VLBI observations.** Here, we apply the concept of numerical optimization via EAs to the fitting of radio astronomical observations obtained via VLBI. Interferometers sample the brightness distribution \( I(l,m) \) of an astronomical object in Fourier-space via its projected baseline between antenna \( i \) and antenna \( j \). The observed visibilities \( V_{ij} \) can be written as

\[
V_{ij}(l,m) = \int \int \mathcal{D} \mathcal{D} l_1 l_2 m_1 m_2 e^{2\pi i (\alpha u + \nu v)}.
\]

where \( l \) and \( m \) are angular coordinates on the sky, and \( u \) and \( v \) are the projected baseline components measured in wavelengths, which act like spatial frequencies. The complex visibility can be converted into the visibility amplitude \( |V_{ij}| \) and the visibility phase \( \Phi_{ij} \). The latter is severely affected by atmospheric fluctuations. The number of complex visibility data points of an interferometer is given by \( N(N-1)/2 \), where \( N \) is the total number of antennas forming the interferometer, assuming that they are all able to observe the source at the same time.

A secondary observable quantity is the so-called closure phase \( \Phi_{ij} \), computed as the sum of the visibility phases on a closed triangle of baselines (Jennison 1958; Rogers et al. 1974). The advantage of closure phases is that station-dependent phase variations cancel out. Closure phases are thus of great importance to measure the structure of the observed source. The total number of independent closure phases computed from an interferometer consisting of \( N \) individual antennas is given by \( (N-1)(N-2)/2 \) (Thompson et al. 2017). If prior information about the station gain phases is available, the dependence on closure phases comes with a loss of information as fewer data are available. If no prior information about the gain phases is available, as is usually the case for our observations, the information
content of the visibility phases and closure phases is the same (Blackburn et al. 2020).

Throughout this paper, we use the visibility amplitude $|V_{ij}|$ and the closure phase $\Phi_{ijk}$ to fit models to VLBI observations. The visibility amplitude may still be corrupted by, for example, gain offsets due to antenna mispointings that cannot be calibrated a priori. However, these gains are usually limited to approximately a few tens of percent, while the visibility phases are randomized due to the atmospheric corruptions. We note that there are additional quantities which could be computed from the complex visibilities, such as the closure amplitudes (e.g., Thompson et al. 2017), which are also free of station-based gain errors. For the minimization function, we use the least squares computed from the visibility amplitudes, $\chi^2_{VA}$, and the closure phase, $\chi^2_{CP}$, (we follow the convention given in Chael et al. 2018, for the calculation of the $\chi^2$). Finally, the cost function for the optimization process can be written as:

$$f(x) = \chi^2_{VA} + w\chi^2_{CP},$$  

where $w$ is a weighting factor. The information on the structure of the source is mostly encoded in the closure phases. Moreover, these quantities are free of gain corruptions and therefore they are more reliable than visibility amplitudes. Using a weighting factor $w < 1$ enforces the optimization to prioritize the reduction of the $\chi^2_{CP}$. This setup typically increases the convergence of the optimization procedure.

**Station-dependent gain factors.** In addition to the free parameters of our model ($\phi$, $f$, and $M$), the antenna gains $g_i$ can be considered as free parameters (see Thompson et al. 2017). Within GENA, we used a numerical minimization as implemented in eht-imaging to obtain the antenna gains following Chael et al. (2018) and Themis (Broderick et al. 2020a) in a self-calibration procedure minimizing

$$\sum_{\text{time} \times \text{set}} \sum_{i,j} (V_{ij,\text{obs}} - g_i g_j^* V_{ij,\text{model}})^2.$$  

During the minimization of Eq. (6), we assumed that the gains were constant during a scan, which speed up the calculations. To avoid compensation of large $\chi^2$ on the visibility amplitude and/or closure phase by strong gain variations, we included a flat prior on the individual antenna gains $g_i$, disfavoring station gain solutions larger than 0.2 (see below).

### 2.3.3. Changes made to GENA

For this work, several changes were made to the GENA pipeline as compared to the version used in Event Horizon Telescope Collaboration (2019e,f). First, self-calibration, which solves for the station gains due to, for example, unknown pointing offsets, has been included in the process of optimizing the rotation and scaling for each frame. Previously, self-calibration was only performed for the best fitting rotation and scaling parameters at the end of the optimization process to determine the eventual goodness-of-fit. Because the station gains can be significant, they strongly increase the $\chi^2_{VA}$ of non self-calibrated data. In the optimization process, the weight $w$ of the $\chi^2_{VA}$ therefore needed to be set low ($\sim 0.01$) with respect to the weight of the $\chi^2_{CP}$ for the optimization to converge.

However, the visibility amplitudes contain valuable information on the source morphology that can be included in the optimization process. We found that including self-calibration in the optimization process and setting $w$ to 0.1 resulted in a better goodness-of-fit and narrower (by $\sim 10\%$) mass distributions in the end. A $w$ larger than 0.1 gave a small but significant increase in the best-fit $\chi^2_{CP}$. Since the closure phase remains the most reliable data product because of its robustness against station gains, we adopted $w = 0.1$, which is still significantly larger than the $w = 0.01$ that needed to be set previously when self calibration was not included in the optimization process. In order to prevent the self calibration from overfitting the visibility amplitudes, station gain offset solutions larger than 0.2 were disfavored. Due to low signal-to-noise ratio $(S/N)$ decoherence and randomized pointing offsets, the gain offsets may exceed this value in some cases despite the fact that the input station gain offsets did not exceed 0.1. Including self calibration in the optimization process increases the runtime of GENA by a factor of $\sim 10$ to 20. In order to mitigate this effect, we performed self calibration only when the closure phase fit was reasonable, that is $\chi^2_{CP} < 10$.

The second change in the GENA pipeline is the treatment of stations with intra-site baselines. For the current EHT, these are ALMA and APEX in Chile and JCMT and SMA in Hawaii. If the total flux density of the source is known, gains of these stations can be solved for by fixing the visibility amplitudes on the intra-site baselines and comparing the amplitudes on the baselines to a third station (network calibration, Blackburn et al. 2019), and no further self calibration is required. In the EHT2017 model fitting, these stations were included in the self calibration nevertheless because the total compact flux density was uncertain due to a lack of short baselines, and the contribution of the large-scale (jet) structure to the intra-baseline amplitudes was not strongly constrained. Since there is no large-scale jet component in the GRMHD library from which we took our input models, and the input total flux density is known, our input synthetic data were network-calibrated using the implementation in eht-imaging (Chael et al. 2016, 2018; Blackburn et al. 2019), and the gains for ALMA, APEX, JCMT, and SMA were fixed throughout the fitting procedure.

### 3. Parameter estimation results

In this section, we show our scoring results for a range of observing conditions. We start by investigating the best fit parameter distributions as a function of stations included in the past, current, and potential future EHT array. We also include results from simulating Space VLBI observations of the Event Horizon Imager (EHI), and investigate the effect of repeated observations of the same source with a fixed array.

As the input model for our synthetic observations, we used the 20th frame of the SANE model with $a_\ast = 0.5$ and $\theta_{\text{high}} = 80$, shown in Fig. 3. Fits to the EHT2017 data were not considered when choosing this particular model, but it was chosen because its appearance is fairly representative for most library frames in that it has no strong jet footprint or emission far from the horizon. No specific preference was involved when choosing the 20th frame as the input model for our simulated ground truth data. The typical correlation time between GRRT images created from BHAC data is around 50M. Thus, images separated by five frames can be regarded as uncorrelated and independent realizations of the variability. The model was observed with SYMBa and the EHI simulator using the setups described in Sect. 2.2.

#### 3.1. General trends

Figure 3 shows the visibility amplitude and closure phase fits for four different library frames that were fit to the EHT2017
synthetic data. The best fit parameters (mass, amplitude, and sky orientation) for the input model frame (top left) were recovered approximately at the percent level. As illustrated by the fit to a different frame from the same model (top right), the best fit parameters and goodness-of-fit can indeed vary considerably due to the variability within the model (model variance).

Larger variations in mass estimates may occur where the image is not dominated by the flux on or close to the photon ring. The bottom row of Fig. 3 shows two examples. On the left is a model with strong emission from the jet footprint, which appears as a small ring in front of the shadow. To optimally fit the EHT data, the model was scaled up in size, resulting in a large mass estimate of $10.38 \times 10^9 M_\odot$. On the right is a model dominated by emission further away from the photon ring, which was scaled down in size to fit the data, resulting in a small mass estimate of $3.07 \times 10^9 M_\odot$. 

Fig. 3. Four example fits to visibility amplitudes (upper panels) and closure phases (middle panels) of the EHT2017 synthetic data generated from frame 20 of the SANE model with $a_*=0.5$ and $R_{high}=80$. Lower panels: model frames that were scaled and rotated by the best-fit values indicated in the upper panels, without (left) and with (right) blurring by a 20 µas Gaussian beam.
Figure 4 shows the best fit masses for all SANE models for synthetic data generated with the EHT2017 (left, blue distributions) and EHT2021 (right, orange distributions) arrays. All model frames have been included in the distributions, that is to say no selection was made based on the goodness-of-fit. The plotted curves were generated with a kernel density estimator. The annotation below each distribution indicates whether the model was accepted of rejected based on the average image scoring procedure. See Appendix A for an analysis and discussion on the validity of the mass scaling in the context of this work.

The estimated mass is generally lower than the true value for models with low $R_{\text{high}}$ and/or retrograde spin. The models with $R_{\text{high}} = 1$ are disk-dominated, resulting in emission further away from the horizon as the observer is looking down the jet and viewing the disk face-on (see also Fig. 3, bottom right panel). A retrograde spin pushes the orbits further out, resulting in lower mass estimates. The models with high prograde spin ($\alpha_*=0.97$) and $R_{\text{high}} \geq 10$ show a strong jet footprint in front of the shadow (see also Fig. 3, bottom left panel), and they result in a large mass estimate. Similar trends were observed in the comparison of the real EHT2017 data to the GRMHD image library (Event Horizon Telescope Collaboration 2019e).

The inclusion of the GLT, KP, and PDB in the EHT2021 array generally narrows the individual mass distributions. Some models with mass estimates that are substantially lower (most $R_{\text{high}} = 1$ models) or higher (the $R_{\text{high}} = 40$, $\alpha_*=0.97$ model) were accepted when observed with the EHT2017 array but rejected when observed with the EHT2021 array, indicating an increased ability to constrain the black hole mass when more stations are added to the array.

A few models with retrograde spin are rejected with the EHT2017 array but accepted with the EHT2021 array (the $\alpha_* = -0.94$ models with $R_{\text{high}} = 10$–40, and the $\alpha_* = -0.5$ model with $R_{\text{high}} = 160$), which is counter-intuitive. This effect can be understood from the average image scoring procedure and investigating the $\chi^2$ distributions obtained for these cases. For example, the median $\chi^2_{\text{CP}}$ from fitting the model frames to synthetic data from the average frame of the same model increases from 5.5 to 15 for the SANE, $\alpha_* = -0.94$, $R_{\text{high}} = 20$ model between EHT2017 and EHT2021, and the standard deviation increases from to 3.9 to 8.3. The overall fit quality thus decreases and the spread across the frames increases as more baselines are added, which is indeed plausible if there are large variations between the different model frames. The additional and longer baselines, which allow one to see more detail in the source structure, make it more difficult for GEVA to fit the varying substructure to data obtained from the average model, where the detailed structure has been washed out. In other words, the additional baselines of the EHT2021 array make the model variance more apparent. In the average image scoring procedure, the acceptance of a model is determined by the goodness-of-fit of the average model image to the input synthetic data, which is generated from a frame of a different model, as compared to the goodness-of-fit of the model frames to synthetic data generated from the average image of the same model. The $\chi^2_{\text{CP}}$ for the former decreases from 32 to 29 between EHT2017 and EHT2021 for the SANE, $\alpha_* = -0.94$, $R_{\text{high}} = 20$ model, where the input synthetic data were generated from our fixed SANE, $\alpha_* = -0.94$, $R_{\text{high}} = 80$, frame 20 input model. So, the average of the model can be fit slightly better to the input synthetic data when more baselines are added, although the overall fit quality remains poor; whereas the $\chi^2$ variations become larger within the model due to its strong variability, and some $\chi^2$-values now exceed that for the average model image fit to the input synthetic data. The rejection criterion was thus reached for EHT2017, but not for EHT2021. This phenomenon is thus a consequence of not setting an absolute $\chi^2$ criterion for the acceptance of a model, but comparing the fit to the model variance. It is not expected to occur for models which do not
In this section, we show how the recovered mass and sky orientation distributions and model selection by average image scoring are affected by the observation setup.

### 3.2. Recovered parameters for different observing conditions

We show strong variability and fit the input synthetic data well, as the additional baselines will then allow the fits to improve. It should not occur for the true input model as the input synthetic data are then from a sample of the model variance. Once rejected for a particular array, it should therefore in principle be safe to consider the model rejected for other arrays as well, provided that the model variance is well sampled by the model frames.

Apart from the mass, the sky orientation $\phi$ is optimized for each frame in the fitting procedure as well. Figure 5 shows the recovered distribution in $\phi$ for all SANE models. Just like the mass, these distributions show a general trend with the black hole spin. The sky orientation is near the true value for models that, similar to the input model from which synthetic data were generated, have a prograde spin. However, the preferred orientation flips when the spin changes direction. The plasma is forced to rotate with the black hole spin due to frame dragging, and it gets boosted in the opposite direction when the spin changes sign, hence changing the image asymmetry. The width of the distributions is generally smaller for models with a high prograde spin, which appear most asymmetric. For the more jet-dominated models ($R_{\text{high}} \geq 10$), many retrograde models show a particularly wide distribution, most of which are rejected by the average image scoring procedure. These models indeed appear symmetric, with most emission outside the black hole shadow. For jet-dominated models ($R_{\text{high}} = 40 – 160$) with strongly retrograde spins ($a_\ast = -0.94$), Doppler boosting becomes significant and the images appear asymmetric, which can be made consistent with the input model synthetic data by rotating the image by $\sim 180^\circ$.

For the MAD models (not plotted), the results are different in that the optimal mass is less dependent on the spin and usually lower than the true mass as the emission is mostly outside the black hole shadow. The trends in $\phi$ are similar to those of the SANE models.

#### 3.2.1. Multi-epoch observations and fitting

The synthetic data were either generated with a single observation of a single frame (designation “1 frame”), or with ten observations from ten randomly selected frames from the same model (designation “10 frames”). The latter case represents a situation where repeated observations of the same source take place with a fixed array, which could become an operational mode for the future EHT. The visibility with equal $uv$-coordinates were coherently averaged after the network calibration step for the ten observations generated with independent data corruption and calibration realizations. For the EHI simulations, the source was set to remain static over the $uv$-spiral completion time of 29 days because of the limited number of movie frames available. In reality, more realizations of the variability would be sampled within an EHI observation. Since the Fourier transform is linear, the average of ten observations of different frames corresponds to an observation of the average frame (modulo observational data corruptions).

The effect of averaging on the visibility data is illustrated in Fig. 6, which shows comparisons between the fast Fourier transform (FFT) of the average image of ten frames, an EHT2021 observation of this average image, and an average of ten independent EHT2021 observations of the ten different frames. The scatter in the visibility amplitudes and closure phases reduces as the visibilities are averaged. The amplitude gain offsets between the single and averaged observation are comparable. The closure phases of a single observation have no offset from the true value, but a small offset from the FFT of the average image caused by averaging the data with gain offsets included occurs for the averaged ten-epoch observation of ten different frames. However, this offset mostly remains within the noise. On the ALMA-LMT-SMT triangle, which has the sensitive ALMA-LMT baseline and large gain variations on the SMT-LMT baseline, a systematic offset beyond the noise can be observed, but it is limited to $\sim 3^{\circ}$. This offset is not enough to cause significant biases in the fits as all recovered mass and sky orientation distributions have the true value within 1$\sigma$ (see next subsections). The variation between model frames and other observational uncertainties (e.g., the limited baseline length) are dominant over this effect.
For the average image scoring, the models were also averaged down to ten frames which are the averages of ten randomly selected model frames. In order to capture the variability between ten-frame averages, each 100-frame GRMHD movie was randomly divided into ten ten-frame averages ten times, so that a total of 100 ten-frame averages were created for each model. The data generated from the average model images should be generated in the same way as the data generated from the ground truth input model, so that the $\chi^2$ can be compared. The average image of each model was therefore observed ten times with independent data corruption and calibration realizations, and these observations were averaged as well. The fitting and average image scoring procedure was then executed exactly as described in the preceding sections.

As the models were averaged down to ten frames, the variable small-scale turbulent structure was averaged out and the images became more dominated by the lensed photon ring and overall size and shape of the emitting region (Fig. 7). The model variance was therefore reduced. Combined with the decreased scatter in the averaged data, we show below that this effect results in significantly narrower best-fit mass and sky orientation distributions and an increased ability to rule out models that do not correspond to the observed input model.

### 3.2.2. Recovered mass and orientation for the input model

First, we investigate the recovered mass and sky orientation distributions by fitting synthetic data to the frames of the input model (SANE, $a_*=0.5$, $R_{\text{high}}=80$) as shown in Fig. 8. As a measure for the width of the distributions, Table 1 shows the standard deviations $\sigma_M$ and $\sigma_\phi$ for the best fit mass and orientation values, respectively, where $\sigma_\phi$ is the circular standard deviation accounting for wraps around $360^\circ$.

The recovered mass distribution clearly narrows and peaks closer to the true value as the array improves. The significant improvement as multiple observations were combined especially stands out. Repeated observations with the 2021 array even recover the mass better than a single observation with the EHI Space VLBI concept ($\sigma_M$ of $0.092 \times 10^9 M_\odot$ and $0.217 \times 10^9 M_\odot$, respectively), because the model variance is significantly reduced and the lensed photon ring becomes more prominent (Fig. 7), allowing for a sharper determination of the mass scaling. A potential explanation for the increase in mass estimates as the array improves is that GRMHD models generally show more emission outside the photon ring than inside, causing a tendency for frames with more large-scale emission.
than the input model to be scaled down, leading to an underestimation of the mass. However, it should be noted that the peak offsets are within $\sim 1\sigma$, and this could depend on the input model.

Adding in the six EHT2021+ stations marginally improved the mass and position angle measurements here. The precision obtained with EHT2021 coverage in combination with using multiple epochs already seems to be approaching a plateau within the maximum baseline length. The widths of the distributions are thus likely not dominated by uncertainties due to gaps in the $uv$-plane, but by the variation in the different model frames and by the maximum baseline length (resolution), which is set to the Earth’s size and observing frequency. We only see a large improvement beyond EHT2021 here when we go to long space baselines. Repeated observations with the EHI further sharpen the distribution. The 230 GHz distribution peaks at a slightly lower mass here, but it is still about $1\sigma$ from the true value. The mass determination is not as sharp for 690 GHz observations: The ratio between accepted MAD and SANE models for each array configuration in Table 2. The magnetic flux is challenging to recover with ground-based observations; The ratio between accepted MAD and SANE models is close to 1:1. A clear preference for SANE models becomes apparent with EHI observations. For the EHT2021+ and high-frequency EHI observations, the model acceptance for the repeated EHT2021 observations and 230 GHz EHI observations was respectively used as a filter, resulting in the additional rejection of a few models.

Figure 9 shows the distribution of accepted models between MAD and SANE models for each array configuration in Table 2. The magnetic flux is challenging to recover with ground-based observations: The ratio between accepted MAD and SANE models is close to 1:1. A clear preference for SANE models becomes apparent with EHI observations. For the 690 GHz EHI observations, all accepted models are SANE models, which indeed corresponds to the ground truth input model.

The black hole spin distribution (Fig. 10) is close to flat for the single EHT2017 and EHT2021 observations, although retrograde spins are slightly disfavored compared to prograde spins. With repeated EHT2021+ observations, the $a_*=0.94$ models are disfavored more strongly. With EHI observations, the true input spin of 0.5 is strongly favored. In this case, the $a_*=0.5$ and $a_*=0.94$ models are completely ruled out with 230 GHz observations. Only $a_*=0$ and $a_*=0.5$ models remain with high-frequency EHI observations.

The most difficult parameter to measure is $R_{\text{high}}$ (Fig. 11). Its distribution remains relatively flat, with little changes between the observing configurations. Only EHI observations rule out the
disk-dominated (\(R_{\text{high}} = 1\)) models. Low \(R_{\text{high}}\)-values seem to be favored over the true value of 80, although it should be noted that only five to eight of the GRMHD models are accepted for the EHT observations (see Table 2). The result that \(R_{\text{high}}\) is more difficult to constrain than, for example, the black hole spin can be understood from the fact that, when the magnetic flux and black hole spin are fixed, it typically does not have a large influence on the model images. This can also be seen in Figs. 4 and 5: While varying the spin can change the fitted mass by a factor of 2 or more due to, for example, the appearance of a jet footprint and the changing plasma orbits for high spins (Fig. 3), the differences between the panels with different \(R_{\text{high}}\) is small. An exception is the \(R_{\text{high}} = 1\) panel. These models are disk-dominated, resulting in more emission outside the black hole shadow and a lower mass estimate (Fig. 3). The \(R_{\text{high}} = 10–160\) models all show more of the jet emission, resulting in a similar morphology. Tighter constraints on \(R_{\text{high}}\) may therefore require multi-wavelength observations and spectral fitting.

**3.2.4. Recovered mass and orientation for the library models**

Figure 12 shows the mass and sky orientation distributions for the accepted models after average image scoring (Table 2).

---

**Table 1. 1σ width of the mass and orientation distributions in Fig. 8.**

|           | EHT2017 1 frame | EHT2021 1 frame | EHT2021 10 frames | EHT2021+ 10 frames | EHL 230 GHz 1 frame | EHL 230 GHz 10 frames | EHL 557 GHz 10 frames | EHL 690 GHz 10 frames |
|-----------|-----------------|-----------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| \(\sigma_M\) (10^9 M_\odot) | 0.534           | 0.320           | 0.092             | 0.087             | 0.217             | 0.063             | 0.039             | 0.049             |
| \(\sigma_\phi\) (%) | 8.61            | 5.15            | 1.48              | 1.41              | 3.51              | 1.03              | 0.63              | 0.79              |
| \(\sigma_\phi\) (deg) | 15.7            | 12.9            | 3.7               | 3.5               | 12.2             | 3.0               | 1.2               | 1.0               |
| \(\sigma_\phi\) (%) | 4.35            | 3.58            | 1.02              | 0.96              | 3.40             | 0.84              | 0.33              | 0.28              |

**Notes.** The percentages are relative to the true input mass of 6.2 \(\times\) 10^9 M_\odot and the full 360° circle for \(\sigma_M\) and \(\sigma_\phi\), respectively.

---

![Fig. 8. Best-fit mass (top) and sky orientation (bottom) distributions from fitting the frames or averages of ten frames of the input model (SANE, \(a_*=0.5, R_{\text{high}}=80\)) to synthetic data generated either from frame 20 of this model or a random selection of ten frames from this model, using a ground (left) or space-based (right) array. The blue and orange line in the top left panel correspond to the distributions in the red ellipse in Fig. 4, and the blue and orange line in the bottom left panel correspond to the distributions in the red ellipse in Fig. 5. The true input mass and sky orientations are indicated with a red dashed line.](image-url)
Table 2. GRMHD library model acceptance (ACC) and rejection (REJ) by the average image scoring procedure for different arrays.

| Magn. | $a_*$ | $R_{\text{high}}$ | EHT2017 1 frame | EHT2021 1 frame | EHT2021 10 frames | EHT2021+ 10 frames | EHL 230 GHz 10 frames | EHL 557 GHz 10 frames | EHL 690 GHz 10 frames |
|-------|-------|-------------------|-----------------|-----------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| MAD   | −0.94 | 1                 | ACC             | ACC             | REJ               | REJ               | REJ               | REJ               | REJ               |
| MAD   | −0.94 | 10                | ACC             | ACC             | REJ               | REJ               | REJ               | REJ               | REJ               |
| MAD   | −0.94 | 20                | ACC             | ACC             | ACC               | ACC               | REJ               | REJ               | REJ               |
| MAD   | −0.94 | 40                | ACC             | ACC             | ACC               | ACC               | REJ               | REJ               | REJ               |
| MAD   | −0.94 | 80                | ACC             | ACC             | ACC               | ACC               | REJ               | REJ               | REJ               |
| MAD   | −0.5  | 1                 | ACC             | ACC             | REJ               | REJ               | REJ               | REJ               | REJ               |
| MAD   | −0.5  | 10                | REJ             | ACC             | ACC               | ACC               | REJ               | REJ               | REJ               |
| MAD   | −0.5  | 20                | REJ             | ACC             | ACC               | ACC               | REJ               | REJ               | REJ               |
| MAD   | −0.5  | 40                | REJ             | ACC             | ACC               | ACC               | REJ               | REJ               | REJ               |
| MAD   | −0.5  | 80                | ACC             | ACC             | ACC               | ACC               | REJ               | REJ               | REJ               |
| SANE  | −0.94 | 1                 | ACC             | ACC             | REJ               | REJ               | REJ               | REJ               | REJ               |
| SANE  | −0.94 | 10                | ACC             | ACC             | ACC               | ACC               | REJ               | REJ               | REJ               |
| SANE  | −0.94 | 20                | ACC             | ACC             | ACC               | ACC               | REJ               | REJ               | REJ               |
| SANE  | −0.94 | 40                | ACC             | ACC             | ACC               | ACC               | REJ               | REJ               | REJ               |
| SANE  | −0.94 | 80                | ACC             | ACC             | ACC               | ACC               | REJ               | REJ               | REJ               |
| SANE  | −0.05 | 1                 | ACC             | ACC             | ACC               | ACC               | ACC               | ACC               | ACC               |
| SANE  | −0.05 | 10                | REJ             | ACC             | ACC               | ACC               | ACC               | ACC               | ACC               |
| SANE  | −0.05 | 20                | REJ             | ACC             | ACC               | ACC               | ACC               | ACC               | ACC               |
| SANE  | −0.05 | 40                | REJ             | ACC             | ACC               | ACC               | ACC               | ACC               | ACC               |
| SANE  | −0.05 | 80                | ACC             | ACC             | ACC               | ACC               | ACC               | ACC               | ACC               |
| SANE  | −0.5  | 80                | ACC             | ACC             | ACC               | ACC               | ACC               | ACC               | ACC               |
| SANE  | −0.5  | 160               | ACC             | ACC             | ACC               | ACC               | ACC               | ACC               | ACC               |
| SANE  | −0.97 | 1                 | ACC             | ACC             | ACC               | ACC               | ACC               | ACC               | ACC               |
| SANE  | −0.97 | 10                | REJ             | ACC             | ACC               | ACC               | ACC               | ACC               | ACC               |
| SANE  | −0.97 | 20                | ACC             | ACC             | ACC               | ACC               | ACC               | ACC               | ACC               |
| SANE  | −0.97 | 40                | ACC             | ACC             | ACC               | ACC               | ACC               | ACC               | ACC               |
| SANE  | −0.97 | 80                | ACC             | ACC             | ACC               | ACC               | ACC               | ACC               | ACC               |
| SANE  | −0.97 | 160               | ACC             | ACC             | ACC               | ACC               | ACC               | ACC               | ACC               |

% ACC 85 90 47 42 15 13 12

Notes. The model used for the input synthetic data generation is indicated in boldface.

Table 3 shows the standard deviations $\sigma_{M}$ and $\sigma_{\phi}$ for the best fit mass and orientation values, respectively. The $\sigma$ values from Tables 1 and 3 are displayed in Fig. 13 as well. The distributions from fitting to the full GRMHD library are significantly broader than those from the fit to the input model only (Fig. 4), indicating that the uncertainties in recovered mass and orientation are dominated by the array’s ability to rule out GRMHD models with different parameters.

The EHT2017 and EHT2021 single observation mass distributions are similar, where the EHT2021 distribution peaks slightly higher and closer to the true value due to the rejection of some SANE models with offset best-fit masses (Fig. 4).
and Sect. 3.1). The $\sigma_M$ of $1.50 \times 10^9 M_\odot$ for the single observation with the EHT2017 array is substantially larger than the error of $0.7 \times 10^9 M_\odot$ measured from the real 2017 data (Event Horizon Telescope Collaboration 2019a,f); this is because that measurement was made from data on four different days using a combination of three different methods, and most models were rejected based on constraints other than those from the EHT data. Repeated observations with the EHT2021 array sharpen the distribution more significantly as, for example, the high mass estimates for the SANE models with $\alpha_s = 0.94$ (Fig. 4) are rejected. The six EHT2021+ stations reduce the number of accepted models by 11% compared to EHT2021 (Table 2), but the mass distributions are similar ($\sigma_M$ is even slightly larger for EHT2021+). The models that were additionally rejected with the ten-epoch EHT2021+ observation were thus not at the tail of the mass distribution for the ten-epoch EHT2021 observation. If the large-scale jet emission is taken into account in the fitting, EHT2021+ observations are expected to be able to rule out a larger part of the GRMHD parameter space (see also Sect. 4). The single observation with the EHI lacks S/N for a sharp peak, but repeated observations show a sharp bimodal distribution as only a few models are still accepted (Table 2). High-frequency EHI observations bring the distributions closer to the true value with a smaller width. The repeated 557 GHz observations give the lowest $\sigma_M$ of $0.27 \times 10^9 M_\odot$, corresponding to an uncertainty of 4.4% with respect to the true input mass.

The sky orientation distribution shows similar trends. Especially the high-frequency EHI observations substantially improve the position angle constraints from a $\sigma_\phi$ of $13.5^\circ$ with repeated EHI observations at 230 GHz and a $\sigma_\phi$ of $5.0^\circ$ and $5.1^\circ$ at 557 and 690 GHz, respectively.

3.2.5. MAD input model

The above analysis was done using a single model (SANE, $\alpha_s = 0.5$, $R_{\text{high}} = 80$) as input. The parameter estimates obtained for the different arrays may depend on the input model. In order to get a rough idea of this input model dependence of the parameter estimates, here, we repeat the analysis performed above using a MAD, $\alpha_s = 0.5$, $R_{\text{high}} = 80$ input model. MAD models are generally less variable than SANE models. As the parameter estimation procedure is computationally expensive, we only focus on the single and ten-epoch EHT2021 and ten-epoch 557 GHz EHI observations here.

With this input model, the number of accepted models after the single and ten-epoch EHT2021 and ten-epoch 557 GHz EHI observations are 87%, 63%, and 18%, respectively. Hence, slightly more models are accepted after repeated ground-based and space-based observations, but the overall trend is similar to using a SANE input model (Table 2). Figure 14 shows the distribution over magnetic flux, spin, and $R_{\text{high}}$ after average image scoring. The true input magnetic flux (MAD, left panel) is favored more strongly for the ground-based array, but the distribution becomes flatter again for the space-based array as some MAD models with a different spin and $R_{\text{high}}$-values are rejected. This is also visible in the middle and right panels. As for the SANE input model (Fig. 10), the spin distribution remains relatively flat with ground-based arrays, but the true input spin of 0.5 is favored with the space-based array. Finally, the $R_{\text{high}}$
EHT2021 observation also gives a smaller $\sigma$ when fitting to the input model only, the single EHT2021 observation of the MAD model already ruled out there is less of an effect of repeated observations here. The single EHT2021 observation of the MAD model already ruled out both the high-spin and disk-dominated SANE models, which caused the mass distribution to be wide when using the SANE input model. When fitting to the input model only, the single EHT2021 observation also gives a smaller $\sigma_M$ when using the MAD model as input, which could be due to weaker variability in MAD compared to SANE. The repeated EHT2021 observations give similar $\sigma_M$ for the MAD and SANE models. The high-frequency EHI observations constrain the mass to 0.4% when fitting to the MAD input model, compared to 0.6% when fitting to the SANE input model.

### 4. Summary and outlook

In this paper, we use the methods developed in Event Horizon Telescope Collaboration (2019e,f) to investigate the potential black hole and accretion parameter estimation capabilities of the current and future Event Horizon Telescope and the Event
We find that the percentage of models rejected from the library from two single input models ($M_{\text{AD}}$, $R_{\text{high}} = 80$) is significantly reduced beyond those ruled out by high-frequency VLBI observations, lightcurve analysis, or polarization measurements and modeling, the library of acceptable models could be significantly reduced beyond those ruled out by high-frequency VLBI.

When the array is extended or a Space VLBI array is employed, the parameter constraints improve as well. Multi-epoch observations with the EHT2021 array improve the mass constraint by 38% compared to a single EHT2017 observation. The Event Horizon Imager, with potential maximum baseline lengths of 19 to 57 G1 depending on frequency, significantly reduces the GRMHD model parameter space: It strongly favors models with the correct input magnetic flux, rules out retrograde and strongly prograde spin models when using our SANE input model, and rules out disk-dominated ($R_{\text{high}} = 1$) models when using both input models. We note that $R_{\text{high}}$ is difficult to constrain further: Once the emission is jet-dominated, the results are only weakly dependent on the microphysics of particle heating. The 1σ black hole mass constraint is 4 to 5% for EHI observations at 557 GHz. This constraint is still limited by the array’s ability to rule out different GRMHD models: When only the input model is fit to the data, it is 0.4 to 0.6%. The mass and sky orientation constraints from fitting to the input model only are about an order of magnitude better than those from fitting to the full GRMHD library for all synthetic observations (Fig. 13). With, for example, multiwavelength observations, lightcurve analysis, or polarization measurements and modeling, the library of acceptable models could be significantly reduced beyond those ruled out by high-frequency VLBI.

Fig. 13. $\sigma_{M}$ (left panel) and $\sigma_{\phi}$ (right panel) values and percentages from fitting synthetic data from different observations to the input model (SANE, $a_*=0.5$, $R_{\text{high}}=80$) and the full GRMHD library. The values are from Tables 1 and 3. The percentages are relative to the true input mass of $6.2 \times 10^9 M_\odot$ and the full 360° circle for $\sigma_{M}$ and $\sigma_{\phi}$, respectively.

Fig. 14. Left to right: same as Figs. 9–11, respectively, but using a MAD, $a_*=0.5$, $R_{\text{high}}=80$ input model.

Fig. 15. Same as Fig. 12, left panel, but using a MAD, $a_*=0.5$, $R_{\text{high}}=80$ input model.

Horizon Imager Space VLBI concept. We use synthetic data from two single input models (MAD+SANE, $a_*=0.5$, $R_{\text{high}}=80$) fit to a GRMHD image library at 230, 557, and 690 GHz. We find that the percentage of models rejected from the library and the ability to measure the black hole mass and black hole spin axis sky orientation increases significantly when observations from multiple epochs are combined. By averaging multiple model frames and epochs, the variance within the models is reduced, leaving images of the quiescent source structure, which has a more prominent photon ring than the individual frames and can therefore be fit more precisely to the observed data.
observations only, leading to better constraints. Multiwavelength constraints were used to rule out 32 of the 60 library models in Event Horizon Telescope Collaboration (2019e) in addition to the nine that were ruled out based on the EHT data only. With the 2017 M87 polarization data, all SANE models in the EHT GRMHD library could be ruled out (Event Horizon Telescope Collaboration 2021a,b).

Assuming that the mass estimation results are, at least in order of magnitude, applicable to observations of Sgr A* as well, they provide an estimate of the precision with which general relativity can be tested. This precision is limited by the ~0.5% mass measurement uncertainty from fitting the input model to the ten-epoch 557 GHz synthetic EHI data. This uncertainty is comparable to the current uncertainty in the Sgr A* mass measurements by Do et al. (2019) and GRAVITY Collaboration (2019). If, in the coming decades, we indeed manage to independently constrain the GRMHD plasma parameters and potentially get a spin measurement from further Gravity observations of stellar orbits and infrared flares in the Galactic Center (e.g., Eisenhauer et al. 2011) or from timing measurements of a pulsar in the Galactic Center (Liu et al. 2012; Goddi et al. 2017), this means a potential null hypothesis test of the Kerr metric (Psaltis et al. 2015; Johannsen et al. 2016) at sub-percent level with high-frequency Space VLBI. Additionally, such a measurement could potentially strongly constrain non-Kerr alternatives, such as boson stars (Olivares et al. 2020), dilaton black holes (Mizuno et al. 2020), and simulations at 345 GHz could be considered as well. For the EHI, the system noise is a limiting factor, and it could be varied to formulate the system requirements more robustly. Combined arrays with space and ground stations (Palumbo et al. 2019; Fish et al. 2020; Andrianov et al. 2021) could be investigated as well.

Analyses such as these are not only useful for investigating the science potential of future VLBI arrays, but they can also be used to optimize observing strategies with current arrays. A future study on the effects of weather, antenna pointing offsets, station dropouts, and scheduling on the parameter estimations could help formulate a set of trigger requirements for the array.

The GRMHD model library is limited in that it has only a few discrete parameter values for the magnetization, black hole spin, and electron temperature distributions. Apart from increasing the number of discrete values, one could think about ways to interpolate between these using, for example, machine learning techniques (van der Gucht et al. 2020; Yao-Yu Lin et al. 2020) or fit to semi-analytic models (e.g., Broderick et al. 2016). The variability within the GRMHD models was found to be an important limitation for constraining black hole parameters, as attested by, for example, the small difference in recovered parameters between the EHT2021 and EHT2021+ arrays. The analysis pipeline may be extended to include a characterization of the source variability as part of the model selection process (e.g., Kim et al. 2016; Roelofs et al. 2017; Medeiros et al. 2017, 2018; Johnson et al. 2017; Bouman et al. 2018; Wielgus et al. 2020), which could improve the constraining power beyond the averaging method introduced here. EHT expansions are expected to make the large-scale jet visible in reconstructed images of the black hole shadow due to an increased dynamic range (Doelman et al. 2019; Roelofs et al. 2020; Raymond et al. 2021). This connection between event-horizon scales and the extended jet has not been taken into account in the parameter estimation framework used here, as the GRMHD library images have a small field of view (160 µas). With the development of GRMHD simulations that have the ability to connect large (e.g., Fromm et al. 2017, 2018, 2019; Liska et al. 2018; Chatterjee et al. 2019) and small scales at different wavelengths and of an extended fitting framework, the constraining power is expected to improve especially for EHT extensions and space arrays. For a mass measurement, feature extraction techniques such as a ring fit (Event Horizon Telescope Collaboration 2019d,f) may be used, potentially in combination with fitting the more extended (variability) techniques (van der Gucht et al. 2020; Yao-Yu Lin et al. 2020) or interpolation between these using, for example, machine learning techniques for Sgr A* and polarization could be considered as well. These possible avenues for further simulation and fitting framework development mean that the parameter constraints presented in this paper should not be interpreted as set limits on the constraining power of the considered arrays. Rather, they show what is achievable with the current state of the art.

While the analysis can indeed be extended, this work already provides some directions for steps to take in the future. Repeated observations with a modestly extended array (EHT2021) already improve black hole and accretion parameter estimates significantly. Repeated observations seem more important than large array expansions, at least within the current fitting framework. Order of magnitude improvements become possible with a small Space VLBI array, and even stronger constraints may be attainable with a larger and more sensitive Space VLBI array, involving more and larger dishes and longer baselines. Multiwavelength and possibly also polarization data can help to constrain the model parameter space, allowing precise mass measurements and hence tests of general relativity. In the long term, we have the machinery and tools to make high-precision tests of the physics and astrophysics of black holes a reality.

Acknowledgements. This work is supported by the ERC Synergy Grant “Black-HoleCam: Imaging the Event Horizon of Black Holes” (Grant 610058). C.M.F. is supported by the Black Hole Initiative at Harvard University, which is supported by a grant from the John Templeton Foundation. JD is supported by NASA grant NNX17AL82G and a Joint Columbia/Flatiron Postdoctoral Fellowship. Research at the Flatiron Institute is supported by the Simons Foundation. Z.Y. is supported by a Leverhulme Trust Early Career Fellowship and a UKRI Stephen Hawking Fellowship. The GRMHD and ray-tracing simulations were performed on GOETHE at the CSC-Frankfurt and Iboga at ITP Frankfurt. The GERA pipeline was developed primarily by C.M.F. We thank Alexander Raymond for providing a list of potential EHT2021+ stations that are suitable for observations of M87. We thank Kotaro Moriyama, Dominic Pesce, and Rachel Doelman for useful comments and discussion on this work. This work has made use of NASA's astrophysics data system (ADS), the Numpy (van der Walt et al. 2011), Scipy (Jones et al. 2001), and Astropy (Astropy Collaboration 2013, 2018) libraries, and the KERN software bundle (Molenar & Simnov 2018).

References
Andrianov, A. S., Baryshev, A. M., Falcke, H., et al. 2021, MNRAS, 500, 4866
Astropy Collaboration (Robitaille, T. P., et al. 2013, A&A, 558, A33
Astropy Collaboration (Price-Whelan, A. M., et al. 2018, AJ, 156, 123
Backes, M., Müller, C., & Conway, J. E. 2016, Proceedings of the 4th Annual Conference on High Energy Astrophysics in Southern Africa (HEASA 2016), 25–26 August, 2016. South African Astronomical Observatory (SAAO), Cape Town, South Africa, 29
Blackburn, L., Chan, C.-K., Crew, G. B., et al. 2019, ApJ, 882, 23
Appendix A: Black hole mass scaling

During the optimization process, we allowed the black hole mass $M$ to vary and we rescaled the pixel size of the images accordingly. The obtained range of black hole masses within our work varies between about $2 \times 10^9 M_\odot$ and $12 \times 10^9 M_\odot$ (see Figs. 4 and 12). In order to test the applicability and the uncertainties introduced by the black hole mass scaling, we performed the radiative transfer calculations for our reference model (SANE, $a_*=0.5$, $R_{\text{high}}=80$, $M=6.2 \times 10^9 M_\odot$) using several black hole masses ($2, 3, \ldots, 12 \times 10^9 M_\odot$) and compared them to the rescaled reference model. The comparison was computed in the image plane via the mean square error (MSE) and in the Fourier plane using the visibility amplitude (VA).

During the radiative transfer calculations, we fixed the field of view to 160 $\mu$as, used a resolution of 1 pixel per $\mu$as, and iterated the mass accretion rate until a total flux of 0.8 Jy was obtained. In Fig. A.1, we show the GRRT images for several different black hole masses using $R_{\text{high}}=80$. The GRRT images using a black hole mass $>4 \times 10^9 M_\odot$ show very similar structures: a prominent photon ring and the foot point of the jet as an inner smaller ring. While we decreased the black hole mass keeping the field of view fixed to 160 $\mu$as, more jet structures moved into the field of view (see for example the first panel for $M=2 \times 10^9 M_\odot$). More importantly, for smaller black hole masses, we needed to increase the mass accretion rate to obtain a total flux density of 0.8 Jy. Connected to the increase of the accretion rate, the opacity of the plasma orbiting the black hole increases. The density in the disk is typically higher than in the jet, and thus the disk turns optically thick and the emitting regions are shifted into the jet.

In the next step, we rescaled our standard SANE model and compared it to the SANE model computed with different black hole masses. The result of this analysis can be seen in Fig. A.2. For $M<4 \times 10^9 M_\odot$, the largest difference comes from the jet and secondly from the optically thick regions in the disk. For larger black hole masses, the differences are minor and mainly occur from slightly better resolved arcs in the highest mass cases. The MSE between the images computed from the different black hole masses and the rescaled one reflects the abovementioned behavior and increases faster for the lower masses than for the higher masses. In Fig. A.3 we present the evolution of the MSE (top) and the largest differences in the visibility amplitude (VA) using the EHT2017 coverage (bottom) for a SANE $a_*=0.5$ (left) and a MAD $a_*=0.5$ model (right) with different values for $R_{\text{high}}$. Both models show the same behavior: Up-scaling the images to larger black hole masses than used in the radiative transfer introduces smaller errors than down-scaling the images to smaller black hole masses\(^2\). The comparison across the accretion model (SANE or MAD) shows that rescaling MAD models introduces smaller errors than rescaling SANE models. A possible explanation for this behavior could be the less variable and more compact emission structure seen in the MAD models.

Inspecting the images (Figs. A.1 and A.2), the structure starts to change significantly around $M=4 \times 10^9 M_\odot$. From the bottom panels of Fig. A.3, the VA error is then limited to about 0.1 Jy. Up-scaling the black hole mass does not introduce VA errors larger than 0.1 Jy (see bottom panels in Fig. A.3). Based on this analysis, images using a black hole mass of $6.2 \times 10^9 M_\odot$ during the radiative transfer should not be down-scaled to black hole masses smaller than $4 \times 10^9 M_\odot$, while up-scaling them leads to less serious issues. Since for this work the GRMHD models were ray-traced at $6.2 \times 10^9 M_\odot$ and the input synthetic data were also generated from a $6.2 \times 10^9 M_\odot$ model, the mass scaling did not lead to serious errors in this work. Some library models had a best-fit mass below $4 \times 10^9 M_\odot$ (Fig. 4), but most of these were rejected by the average image scoring procedure, so that the final mass distributions (Fig. 12) hardly extend below $4 \times 10^9 M_\odot$. When fitting to real data, it is recommended to perform a similar analysis as presented in this appendix and redo the ray-tracing if a significant fraction of the fitted masses is in the problematic regions.

---

\(^2\) Assuming a fixed field of view and constant flux of 0.8 Jy.

---

Fig. A.1. Reference model (SANE, $a_*=0.5$, $R_{\text{high}}=80$, $i=163^\circ$) using a different black hole mass during the radiative transfer calculations, with masses of $2 \times 10^9 M_\odot$ to $12 \times 10^9 M_\odot$ increasing from left to right. The black hole mass used is indicated at the top of each panel.
Fig. A.2. Pixel differences between the rescaled reference model (SANE, $a = 0.5$, $i = 163^\circ$, using $M = 6.2 \times 10^9 M_\odot$) and images using a different black hole mass during the radiative transfer calculations from left to right $2 \times 10^9 M_\odot$ to $12 \times 10^9 M_\odot$.

Fig. A.3. Result of the mass scaling analysis for SANE (left) and MAD (right) model with $a = 0.5$ for different $R_{\text{high}}$ values. Top panels: MSE and the bottom ones show the maximal difference in the visibility amplitude.