Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **494,** 4771–4793 (2020)
Advance Access publication 2020 April 7

doi:10.1093/mnras/staa921

Systematic versus statistical uncertainties in masses and magnifications of the Hubble Frontier Fields

Catie A. Raney,[1]★ Charles R. Keeton,[1] Sean Brennan[1] and Hsin Fan[2]

[1]*Department of Physics and Astronomy, Rutgers University, 136 Frelinghuysen Road, Piscataway, NJ 08854, USA*
[2]*Department of Physics and Astronomy, Stony Brook University, Stony Brook, NY 11794, USA*

Accepted 2020 March 29. Received 2020 March 26; in original form 2020 January 10

ABSTRACT

The Hubble Frontier Fields data, along with multiple data sets obtained by other telescopes, have provided some of the most extensive constraints on cluster lenses to date. Multiple lens modelling teams analyzed the fields and made public a number of deliverables. By comparing these results, we can then undertake a unique and vital test of the state of cluster lens modelling. Specifically, we see how well the different teams can reproduce similar magnifications and mass profiles. We find that the circularly averaged mass profiles of the fields are remarkably constrained (scatter < 5 per cent) at distances of 1 arcmin from the cluster core, yet magnifications can vary significantly. Averaged across the six fields, we find a bias of -6 per cent (-17 per cent) and a scatter of $\sim$40 per cent ($\sim$65 per cent) at a modest magnification of 3 (10). Statistical errors reported by individual teams are often significantly smaller than the differences among all the teams, indicating the importance of continued systematics studies in cluster lensing.

Key words: gravitational lensing: strong – galaxies: high-redshift, clusters: general, individual: (Abell 2744, MACS J0416.1+2403, MACS J1149.5+2223, MACS J0717.5+3745, Abell S1063, Abell 370).

1 INTRODUCTION

Galaxy clusters are the largest gravitationally bound objects in our Universe, with masses of 10^{14}–$10^{15}\,\mathrm{M}_{\odot}$. They are dominated by dark matter but are also made up of both hot gas in the intracluster medium (ICM) and hundreds to thousands of galaxies. These structures are built up by mergers of groups and other clusters of galaxies, which can give them complicated mass distributions. However, they can be very informative to study. For example, how common these extreme systems are and how mass is distributed within them can give constraints on dark matter properties. An example of the latter is the well-known Bullet cluster (Clowe et al. 2006), and a similar analysis has been applied to many systems since (e.g. Bradač et al. 2008; Merten et al. 2011; Harvey et al. 2015).

Gravitational lensing can be a useful tool in studying the mass of these galaxy clusters (see review by Hoekstra et al. 2013). Lensing occurs when light from a background source is bent by intervening mass. Since galaxy clusters are both very massive and large on the sky, they offer a wide area over which this lensing can be detected. In the weak lensing regime, the image of the background galaxy is only very slightly stretched tangentially around the cluster. While this stretch usually cannot be seen by eye, it can be detected through statistical studies of thousands of galaxies. This allows for the mass distribution of the cluster to be constrained out to large radii but with low resolution (see e.g. Umetsu et al. 2014; Bartelmann & Maturi 2017; Murata et al. 2019).

Strong lensing occurs closer to the core of the cluster, where the density is highest. In this case, the light from a background galaxy is more strongly affected, and two or more images of the galaxy are produced. These multiple images can be used to constrain the mass of the cluster within the strong lensing region, i.e. where the multiple images are found. This offers higher resolution than weak lensing but is limited in radius (Cibirka et al. 2018; Jauzac, Harvey & Massey 2018; Andrade et al. 2019, etc.)

In the case of strong lensing, galaxy clusters can also be used as cosmic telescopes (see review by Kneib & Natarajan 2011). The multiple images produced often have a magnification that makes the images of the source appear brighter than they would without the lensing effect. Further, they can be stretched out into long arcs; this allows the study of the galaxy at a higher resolution than it would have otherwise, down to sub-kiloparsec scale (e.g. Livermore et al. 2012; Johnson et al. 2017; Dunham et al. 2019). This has been particularly useful in the study of intermediate- and high-redshift galaxies ($z > 6$), which are intrinsically small and very faint (e.g. Zheng et al. 2012; Coe et al. 2013; Salmon et al. 2018).

The goal of the Hubble Frontier Fields (HFF; Lotz et al. 2017) program was to use galaxy clusters in this way to study galaxies

★ E-mail: raney@physics.rutgers.edu

Published by Oxford University Press on behalf of the Royal Astronomical Society

from the first billion years of cosmic history. The program included an extensive observing campaign to produce very deep, multiband images of six known lensing clusters. In addition, a number of other campaigns utilized different ground-based telescopes, which provided both spectroscopic and photometric data in different bands and over a wider area. Combining these produced a wealth of information on galaxies, both in the cluster and along the line of sight, as well as on candidate lensed images. The program proved successful, with a number of images found at high redshift, allowing the luminosity functions at $z \sim 6$ and greater to be better estimated (McLeod, McLure & Dunlop 2016; Bouwens et al. 2017; Oesch et al. 2018).

An important part of the program was that multiple teams were invited and/or funded to model the fields. In order to determine the intrinsic properties of a lensed galaxy, e.g. its size and luminosity, one must use a lens model to determine how much it is being magnified. To do that, a model of the mass in the field must be constructed. Of course, with such complicated systems, there are many possible sources of error in the models. Some of these errors have been studied (e.g. Host 2012; Johnson & Sharon 2016; Acebron et al. 2017; Chirivì et al. 2018; Raney, Keeton & Brennan 2019) but not all of them. If many teams model the fields, some of these errors will be marginalized over, or at least explored, when combining results.

The Hubble Frontier Fields data set then is extremely useful, not just in creating detailed models of the fields in question but also in comparing results from multiple teams. Priewe et al. (2017) examined magnifications within the core of two HFF clusters, Abell 2744 and MACS J0416, finding high dispersion (30 per cent at low magnifications) between the version 3 models analyzed. Remolina González, Sharon & Mahler (2018) also considered models of the field MACS J0416, though they studied scatter in rms of images and how well old models could predict the positions of new images. Meneghetti et al. (2017) generated two mock clusters, aiming to produce mass distributions that were similar to clusters of the HFF sample, in both mass and complexity. They then asked teams to model the two fields and compared the results with a variety of metrics. In the case of mock clusters, the true mass distribution is known, as are the magnifications of the images, which makes comparing the models easier than with real clusters where it is not known. However, mock clusters might not capture the full complexity of a real mass distribution.

In this work, we aim to expand on previous studies by comparing the publicly available results[1] in the latest round (version 4) of modelling all six HFF clusters. In particular, we examine mass profiles and magnifications. By surveying how well the models of various teams agree, we can both test the current state of the field and use the results as a way to inform future cluster lensing work. This is especially useful since it is not given that cluster lensing studies in the future will have the amount of modelling effort that the HFF project did: only one or two teams might model a field and thus would likely not be able to capture the full errors in magnification.

We begin this paper with an overview of the HFF modelling process in Section 2. From there, we look at mass profiles in Section 3, as well as give a brief introduction to each field. In Section 4, we examine the magnification maps submitted. We discuss results from both mass and magnification comparisons in Section 5. We conclude our findings and offer broader implications of the work in Section 6.

2 HFF MODELLING OVERVIEW

2.1 Data and process

In this work, we compare models created in the latest (version 4) round of modelling. The process started with teams numerically ranking candidate lensed images based on spectroscopic data, matching colours and morphologies, and whether or not a team would use an image as a constraint on its model. The images were then given a medal ranking. GOLD images were those for which the majority of teams were confident that the image was part of a lensed family and it had a spectroscopic redshift; SILVER images also required high confidence but did not have secure spectroscopic redshifts. More tenuous images were given the BRONZE ranking, while some images received no ranking if, for example, they were added late in the process and thus not all teams ranked them. Tables of images we used to constrain our models, as well as the catalogues used for cluster member and line-of-sight (LOS) galaxy selection, can be found in Raney et al. (2019).

In creating the models, teams were left to choose their own inputs and modelling methodology. Techniques for lens modelling fall within two categories: parametric and free-form (sometimes called non-parametric). Parametric models consist of small-scale haloes for galaxies and large-scale haloes for dark matter and ICM/hot gas. Mass is usually assigned to galaxies using scaling relations tied to some reference galaxy, e.g. the brightest cluster galaxy (BCG) or an $L*$ galaxy at the cluster's redshift. This allows the model to have only a few free parameters for all of the cluster members since positions (and sometimes ellipticity/position angle) are informed by the light distribution. Large-scale haloes, on the other hand, are usually allowed to vary freely. Both kinds of haloes are parametrized by given density profiles. Free-form models, on the other hand, do not put such constraints on the mass of the haloes. This freedom is both useful in that it can capture oddities in the mass distribution but can also be a disadvantage if there are less constraints than free parameters. Hybrid techniques are those that have free-form large-scale haloes but use given density profiles for small-scale haloes.

2.2 Modelling deliverables

Each team submitted a number of deliverables for its fiducial model, as well as a number of realizations of the model. These realizations, which we will refer to in this work as 'range maps', varied from 40 to over 200 and were meant to sample the uncertainty in a model. It is important to note that lensing quantities depend on the distances between the observer, lens, and source. For the range maps, all teams submitted shear ($\hat{\gamma}$) and convergence ($\hat{\kappa}$), or surface mass density, maps that correspond to a source at infinite distance. From there, the quantities can be found at any source redshift using

$$\kappa = \frac{D_{\rm ls}}{D_{\rm s}}\hat{\kappa} \;\; ; \; \gamma = \frac{D_{\rm ls}}{D_{\rm s}}\hat{\gamma}, \tag{1}$$

where $D_{\rm s}$ and $D_{\rm ls}$ represent angular–diameter distances from the observer to the source and from the lens to the source, respectively.

Further, while the fiducial model submitted had to include magnification maps for $z = 1$, 2, 4, and 9, one can find the magnification at any redshift by using

$$\mu = \frac{1}{(1-\kappa)^2 - \gamma^2}. \tag{2}$$

We note that these equations are true only for a 2D model, i.e. a single lens plane. With a 3D model, there are multiple lens planes

[1] https://archive.stsci.edu/prepds/frontier/lensmodels/

and thus the shear and convergence are not so easily scaled (see e.g. Schneider, Ehlers & Falco 1992).

One can also use these $\hat{\kappa}$ maps to find the mass predicted by a model. The convergence is defined as the surface mass density divided by a critical surface density:

$$\hat{\kappa} = \frac{\Sigma}{\hat{\Sigma}_{\rm crit}}, \text{ where } \hat{\Sigma}_{\rm crit} = \frac{c^2 D_l}{4\pi G}, \quad (3)$$

and $D_{\rm l}$ is the angular–diameter distance from the observer to the lens. By summing the convergence, for example in circular apertures as we do in this work, the mass can be computed.

2.3 Participating teams

In this work, we consider models from five teams using parametric methods, two using free-form methods, and one using a hybrid technique. Three teams (Caminha, CATS, and Sharon) share the same modelling code (Lenstool), while all other teams use separate codes. For an in-depth overview of the techniques for each team, we point the reader to Meneghetti et al. (2017) or Priewe et al. (2017).

The teams using parametric methods are Caminha (Caminha et al. 2017), Clusters As TelescopeS (CATS) (Jullo et al. 2007; Jauzac et al. 2012, 2014; Richard et al. 2014), Glafic (Oguri 2010; Ishigaki et al. 2015; Kawamata et al. 2016, 2018), Keeton (Raney et al. 2019), and Sharon (Jullo et al. 2007; Johnson et al. 2014).

Two teams use free-form methods: Bradač/Strait (shortened to Bradač in plots for space; Bradač et al. 2005, 2009; Strait et al. 2018) and Williams (Liesenborgs et al. 2007; Mohammed et al. 2014; Grillo et al. 2015). One team, Diego, uses a free-form method but assumes that light traces mass for the galaxies, i.e. each galaxy is initally assigned mass based on its surface brightness and later optimized (Diego et al. 2005a, b, 2007, 2015; Vega-Ferrero, Diego & Bernstein 2019).

All teams use only strong lensing constraints except Bradč/Strait, who also employ weak lensing. The number of haloes (large-scale and galactic) varies among the teams, as do the density profiles of the haloes for the parametric models. The number of images used as constraints can differ as well and, in some fields, by large amounts (e.g. ∼100 images).

3 MASS COMPARISON

3.1 Overview

One of the ways that we can compare the results from all teams is by looking at mass profiles. The goal of a lens model is to find the underlying mass distribution and source configuration that can produce the lensed images seen in the data. This is not an easy task, especially in cluster lensing due to the inherent complexity of galaxy clusters. Further, the clusters that are most likely to be chosen as cosmic telescopes are those that are both large on the sky and very massive. These two factors combine to give a larger area on the sky where background galaxies can be strongly lensed. However, this can cause a selection bias for clusters that are undergoing a merger, which can increase both the density and physical size of a cluster. A configuration that is also favourable to lensing many images is multiple large-scale haloes along the line of sight, which can boost lensing strength (Wong et al. 2012; Bayliss et al. 2014).

It can be difficult for lens models to differentiate between mass profiles in a cluster using just image positions as constraints. For example, a recent study of the Hubble Frontier Field MACS J0717.5+3745 found that the data fit models with cored and non-cored dark matter haloes equally well, even with 132 constraints (Limousin et al. 2016). This is also seen in mock data: a model with isothermal haloes can fit position data just as well as a model with NFW haloes even though the density profiles are obviously different, as are the resulting image magnifications (Shu et al. 2008).

A common metric used to compare mass distributions found by lens modelling is the 1D mass profile. This was used in Meneghetti et al. (2017) to compare the results from multiple teams modelling two mock clusters, as a way of determining how accurate and precise the models were. It was found that, though the multiple teams used different density profiles for the haloes and different modelling techniques, they were able to recover 1D mass profiles to within 15 per cent of the true value.

In this work, we do not know the true mass distribution of the cluster, but it is still useful to compare the mass profiles obtained by the different modelling teams and see the extent to which they agree or disagree. We construct our 1D profiles by computing the mass in circular apertures centred on the BCG. In the following subsections, we give a brief introduction to each lensing field, including a sky map. This map includes two solid circles at 5 and 100 arcsec from the BCG, which correspond to the x-axis limits of the 1D mass profiles, shown in the right-hand panels. The profiles are split between parametric (top) and free-form (bottom) techniques for clarity. We note that for each model, we plot the 1D profiles for all of the submitted range maps, such that the thickness of the line illustrates the uncertainty in the model. We also note that the lines very often overlap. The median across all models and realizations is plotted in black on both panels for reference.

Since the modelling teams were allowed to choose the size of their maps, the mass profiles do not all go out to the same radii. Further, parametric models use certain density profiles for their haloes; thus, mass continues to grow at large radii. Free-form techniques, on the other hand, have different priors and regularizations. This can produce flatter profiles at larger radii where there are no lensed image constraints and, as we will see in Section 4, lower magnifications. We also note that, though we have created both 2D and 3D models of each field, we include only the 2D models in the mass analysis. A multiplane model has mass at different redshifts and thus is not a fair comparison to single-plane models.

We indicate the locations of lensing constraints in two ways. The dashed circle in the sky map indicates the spatial extent of the lensed images we used in our models (see Raney et al. 2019), which are primarily the GOLD sample. In many cases, the images are not centred around the BCG because its position does not coincide with the centre of the mass distribution due to merging systems. We also mark the image positions as vertical lines in the top mass profile panel. This helps to show the distribution of these images and informs where the models might be most tightly constrained. We stress that the sample shown is unique to our team and fairly conservative since it is primarily restricted to images with spectroscopic information. Other teams may have used different images and thus their models will be constrained differently.

3.2 Abell 2744

This field, part of the Abell galaxy cluster catalogue (Abell, Corwin & Olowin 1989), was the first HFF cluster to be observed by the *Hubble Space Telescope* (*HST*) and has a redshift of $z = 0.308$. It is a system undergoing a merger, as evidenced by a number of

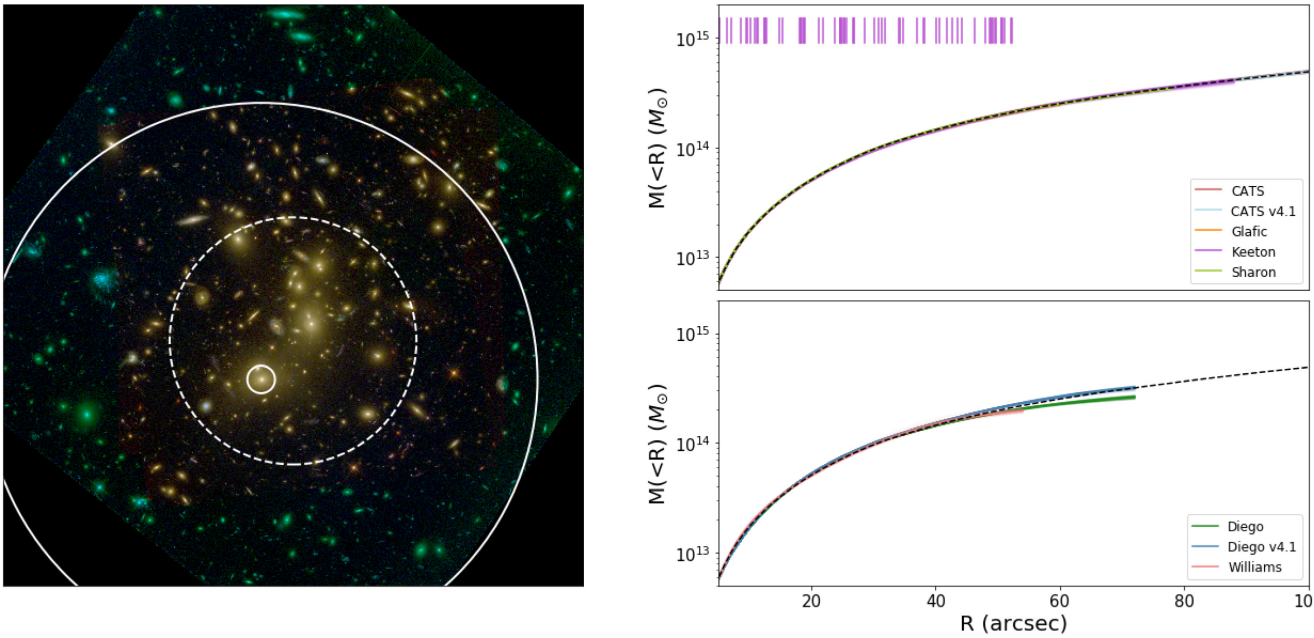

Figure 1. *Left: HST* multiband colour image (produced using Trilogy, Coe et al. 2012) of the cluster Abell 2744. Solid circles are centred on the BCG and have 5 and 100 arcsec radii, corresponding to the x-axis limits in the panels on the right. The dashed circle encloses the majority of spectroscopically confirmed images: specifically, those images we used in our own modelling (see appendix B in Raney et al. 2019). The panel is 3.5 arcmin on a side and is oriented such that North is up and East is left. We note that the size of the panel does not correspond to the size of the submitted maps of the teams. *Right:* Mass profiles centred on the BCG and circularly averaged, computed from the publicly available $\hat{\kappa}$ maps for a source at infinity. All submitted maps are plotted, including the realizations such that the thickness of the line describes the error. The median profile across all teams is also plotted (black, dashed). Models employing parametric techniques are shown on top while free-form/hybrid models are in the bottom panel. We note that some submitted maps covered a smaller area than others, causing the profiles to truncate at different radii. We also include lines indicating the distance of images from the BCG for the constraints used in our model.

factors. The first is that the cluster is physically very large. In Fig. 1, we show the *HST* colour image of the field, but we note that the cluster extends to the north-west, past the field of view (FOV). In the figure, we see two galaxies with similar brightness, which could both be classed as BCGs; ∼2 arcmin away, there are three more galaxies with the same brightness down to photometric errors (Mann & Ebeling 2012). However, since there are more cluster member galaxies around the southern two BCGs, this is considered the main part of the cluster.

Optical and X-ray studies suggest that the system has undergone two mergers in the recent past, one of which was line of sight (Kempner & David 2004; Merten et al. 2011; Owers et al. 2011). This would explain both the large number of BCGs and the offsets found between peaks in the X-ray data and the positions of the cluster members. While the mass outside the *HST* FOV is affecting the lensing on some scale, it is not well constrained due to the lack of lensed images in that region, far from the southern core. Most modelling teams did find that the models preferred to place a large-scale halo to the north-west of the main cluster, as we will see in the magnification maps in the next section. The image constraints in the main part of the cluster are fairly numerous: around 70 images have spectroscopic redshifts. This is in large part due to a recent spectroscopic survey (Mahler et al. 2018) using MUSE.

In the latest round of modelling, six teams created models of the field. We show the 1D mass profiles for each model in the right-hand panels of Fig. 1. In the top panel, we show models that were made using a parametric method, while those shown in the bottom panel were made with free-form methods. The width of the line represents the scatter in the model using the submitted range maps. Some teams cut off before the edge of the plot due to smaller area of their submitted maps.

It is immediately obvious that all the models agree fairly well. The two Diego models, which here differ in their constraint selection (GOLD+SILVER+BRONZE versus GOLD), are fairly different at larger radii: the v4.1 profile agrees with the parametric and median curves, while the v4 profile has a shallower slope. We will see in Section 4 that the magnification maps of these two models are also quite different. Nonetheless, the scatter among all models is surprisingly low with 1σ scatter of < 5 per cent out to an arcminute from the BCG. In fact, the scatter becomes < 1 per cent at 14 arcsec from the BCG, the lowest value out of all the fields in the HFF sample.

3.3 MACS J0416.1-2403

The second of the Hubble Frontier Fields to be observed by *HST* is this cluster at $z = 0.396$ from the Massive Cluster Survey (MACS; Ebeling, Edge & Henry 2001). Similar to Abell 2744, there is evidence that it is undergoing a merger, though one that is not quite as dramatic. From the sky map in Fig. 2, one can see that there are two BCGs with similar brightness. Further, the X-ray map is distinctly doubly peaked (Mann & Ebeling 2012). The merger is likely one that is along the line of sight. Due to this orientation, the lensing area is elongated in such a way to produce a large number of triple images in a ladder configuration.

Indeed, this field has the most images in the GOLD sample out of all the six fields: ∼95. This allows for models that can be well

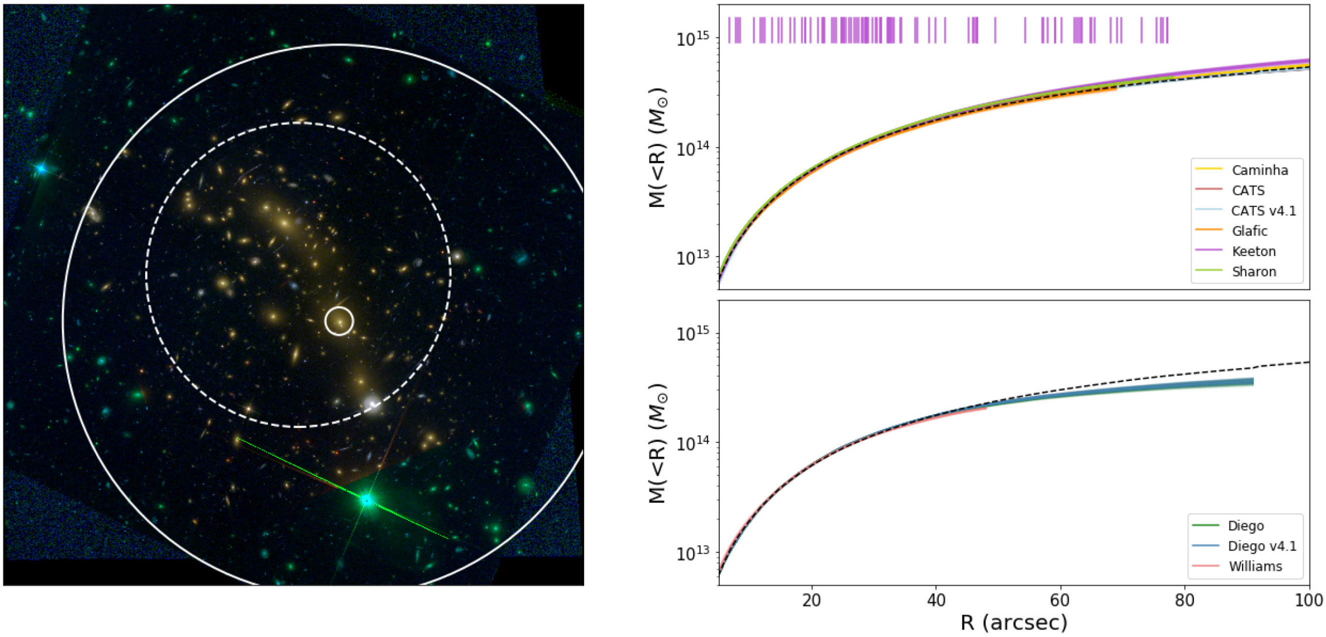

Figure 2. Similar to Fig. 1 for MACS J0416.

constrained, which is indeed what we see in the right-hand panels of Fig. 2. The mass profiles are very similar, specifically at radii between 10 and 40 arcsec, where the scatter is around 2.5 per cent. It is not surprising that this is also the range in which the bulk of the images are found. It is interesting to note that there is more scatter at larger radii in this field than in Abell 2744, where the mass distribution is known to extend beyond the modelled area.

3.4 MACS J0717.5+3745

This cluster, found as part of the MACS survey (Ebeling et al. 2001), is superlative among the HFF sample in many respects. It has the highest redshift at $z = 0.545$, slightly higher than MACS J1149 at $z = 0.543$. It is the most massive cluster in the sample and also likely the most complicated; it was considered the most disturbed system at $z > 0.5$ due to the complex nature of its X-ray data (Ebeling et al. 2007). Part of the complexity comes from a filament (Ebeling, Barrett & Donovan 2004; Jauzac et al. 2012), which could be causing the odd elongated nature of the lensing critical curves that we will see in the next section.

We see in Fig. 3 that the field is not a typical cluster with a BCG in the centre of smaller cluster member galaxies. Indeed, the galaxy classed as the BCG (within the smallest circle in the figure) is at the centre of neither the cluster members nor the area covered by lensed images (shown by the dashed circle). The proposed filament can be seen in the figure as the swath of cluster galaxies extending to the upper right. We note that the bright galaxy to the bottom left is likely a foreground galaxy based on a photometric redshift of $z = 0.155 \pm 0.03$ from CLASH (Postman et al. 2012; Molino et al. 2017) and Subaru/Suprimecame imaging (Medezinski et al. 2013). Yet another source of complication comes in the form of a possible LOS structure in the field for which Williams, Sebesta & Liesenborgs (2018) found evidence.

Unfortunately, this complex cluster also has the least number of spectroscopically confirmed images with which its mass can be constrained: less than 30. That is not to say the field lacks candidate images; the CATS team, for example, used 132 images in their v4 and v4.1 models. These two models are different in that they have either cored or non-cored haloes, respectively. Even with the large number of constraints they used, they found that both models were able to fit the data equally well (Limousin et al. 2016).

This is evident in the mass profiles shown in the right-hand panels of Fig. 3, where the CATS v4 and v4.1 models (red, blue) do indeed disagree at low radii. Interestingly, there does not appear to be a lot of intrinsic scatter in each model. This is not true for the other two parametric teams; the Sharon team's model has a fairly large spread around the core of the cluster, as does our model. All of these models converge at higher radii, though, which is unsurprising: the constraints also extend to a large radius.

For the free-form teams, the results are slightly different. There is some scatter at smaller radii but not as much as among the parametric models. Further, there is more scatter at larger radii. The two Williams models are different but do straddle the median curve. The two Diego models, on the other hand, agree with each other very well but lie the farthest from the median profile.

3.5 MACS J1149.4+2223

This cluster is also at a fairly high redshift ($z = 0.543$) but is less complex than MACS J0717. For example, the BCG is notably brighter than any other galaxy in the field and lies nicely at the epicentre of the lensed images, as seen in Fig. 4. It does have a somewhat elongated mass distribution, so it is likely undergoing a merger, but one that is in later stages than some of the other fields.

The cluster has been the focus of many studies due in large part to a triply imaged spiral galaxy. Two of its images sit close to the BCG, the closer of which is fairly distorted. The second image has a spiral arm further lensed into an Einstein-cross configuration by a cluster member galaxy but otherwise shows only a small amount of distortion. The third image, ~20 arcsec from the BCG, is also mostly intact. These three images can thus be used to give constraints on the mass distribution of the BCG and

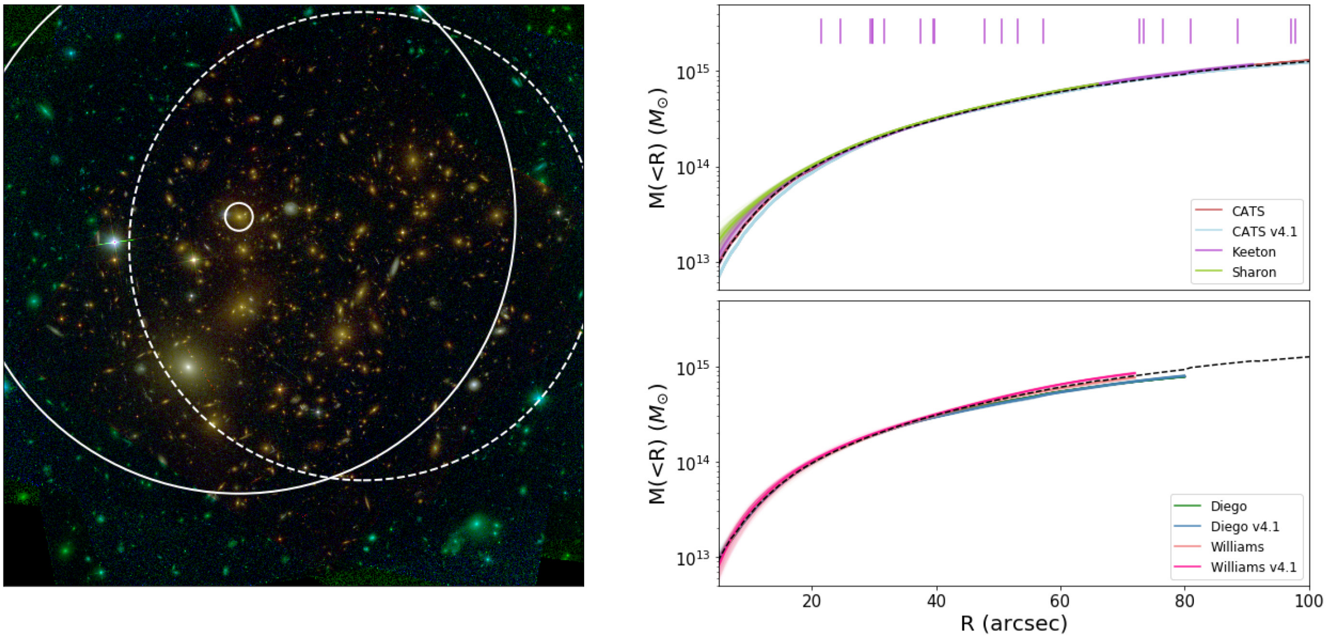

Figure 3. Similar to Fig. 1 for MACS J0717.

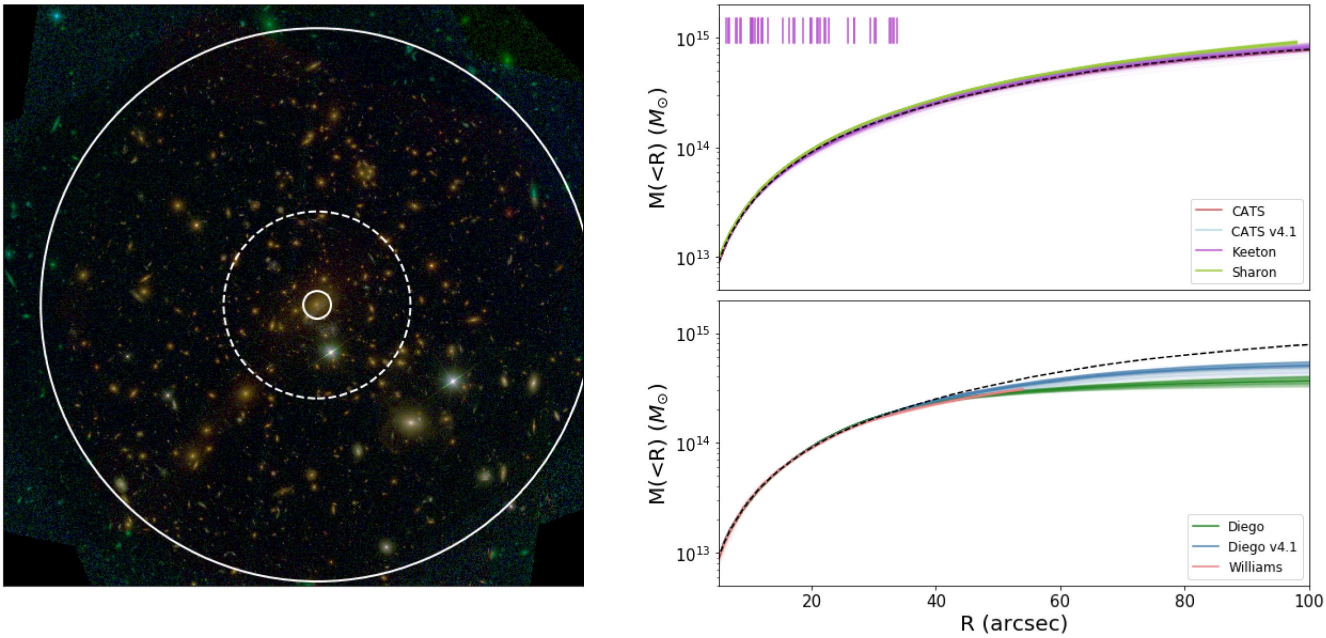

Figure 4. Similar to Fig. 1 for MACS J1149.

cluster core (e.g. Zitrin & Broadhurst 2009; Rau, Vegetti & White 2014).

This spiral galaxy was also the host of SN Refsdal, a Type II supernova that was found in the arm of the galaxy that was further lensed by a cluster member (Rodney et al. 2016; Treu et al. 2016). The four images of the SN in the cross-configuration were named S1–S4. The SN was also set to appear in the image of the galaxy closest to the cluster core but not for a time after S1–S4. Thus, this other image (SX) could be used as a test of the predictive abilities of lens models. The models were able to predict the position of SX quite well but its time delay, i.e. when it would appear, proved harder to pin down (Kelly et al. 2016). Still, the ability to make somewhat accurate predictions is a good sign that the lens modelling is headed in the right direction.

It is important to note that, while this field was the subject of many studies, there are still relatively few lensed images with spectroscopic redshifts; only 22 images from nine sources were ranked GOLD. Star-forming knots within the spiral arms of the Refsdal host galaxy (e.g. see Kawamata et al. 2016) can be used as further constraints on the model. We do note, however, that two of the images of this galaxy are <10 arcsec from the BCG, and thus the majority of the constraints are on the inner region of the cluster.

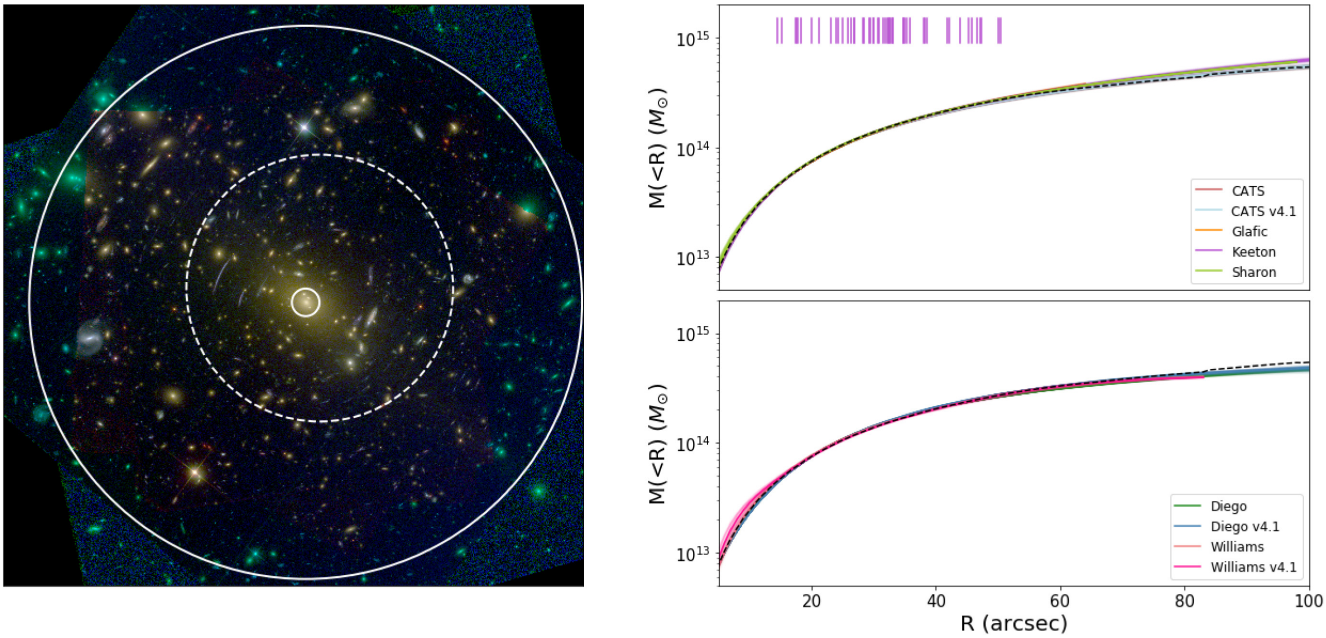

Figure 5. Similar to Fig. 1 for Abell S1063.

Given that there are many constraints close to the BCG and so few farther out, it is then unsurprising that the mass profiles shown in the right-hand panels of Fig. 4 are tight at small radii but become broader as radius increases. In fact, at an arcminute from the BCG, MACS J1149 has the highest scatter across the six fields: 20 per cent. This is driven by the differences between the two Diego models, which are not only very different from each other but are also quite far from the median and the parametric models.

3.6 Abell S1063

This cluster ($z = 0.348$) is the most well behaved in the HFF sample. For example, there is one clear BCG that lies at the centre of the cluster galaxies, as shown in Fig 5. It can also be seen that the GOLD sample, consisting of almost 50 images from 19 sources, are mostly clustered around the BCG, though there are quite a few candidate images to the north-east. There is evidence that the system is undergoing a merger based on dynamical studies (Gómez et al. 2012), which could explain this. Nonetheless, it is not as dramatic of a merger, or perhaps is in a later stage than other clusters in the sample.

The mass distribution of the cluster core is well constrained among the parametric models, though the free-form/hybrid models show scatter some at low radii. Like the Diego models, the two Williams models differ in their constraints: in this case, v4.1 is only the GOLD sample, while v4 uses GOLD+SILVER+BRONZE. It is interesting that the largest differences are seen near the core of the cluster in the free-form models while parametric models show more (though still a small amount of) scatter at larger radii. Nonetheless, the mass of the cluster is very well constrained with 1σ scatter of less than 5 per cent past 10 arcsec.

3.7 Abell 370

This cluster ($z = 0.375$) was the first in which a strongly lensed galaxy was discovered, stretched into a giant arc (Soucail et al. 1987; Lynds & Petrosian 1989). In the 30 yr since it was found, it has been the subject of many studies on both weak and strong lensing (e.g. Abdelsalam, Saha & Williams 1998; Bézecourt et al. 1999; Medezinski et al. 2010, etc.). Its structure and galaxies have also been studied (de Filippis, Sereno & Bautz 2005; Lah et al. 2009), pointing towards a system undergoing a line-of-sight merger. In Fig. 6, we see evidence of this complexity: two possible BCGs and a large area over which lensed images are found. Indeed, this field has the second highest number of spectroscopically confirmed images, in part due to a recent MUSE survey (Lagattuta et al. 2017, 2019).

The large number of constraints on the field from the 90 GOLD images does seem to be able to combat the complexity, at least for the parametric models. The mass profiles of Fig. 6 show scatter at small radii, but most of the models agree very well at larger radii. Indeed, this field has the smallest 1σ scatter at an arcminute from the BCG out of all six fields: only 2 per cent. This is probably due in part to the wide area over which the image constraints are spread, similar to what was seen in MACS J0717 but with many more images.

However, it is interesting to look at the outlier case of the Bradač-Strait model, which is significantly higher than the other mass profiles. Recall, this team also employed weak lensing data in addition to the strongly lensed images to constrain the mass distribution at larger radii; none of the other teams did this. It is unclear whether this higher mass profile stems from the weak lensing alone, or also from their modelling methodology, but it is an interesting result.

4 MAGNIFICATION COMPARISON

4.1 Overview

In order to determine the intrinsic properties of a lensed galaxy, the amount of magnification must first be determined. This makes magnification the most important quantity in the search for and

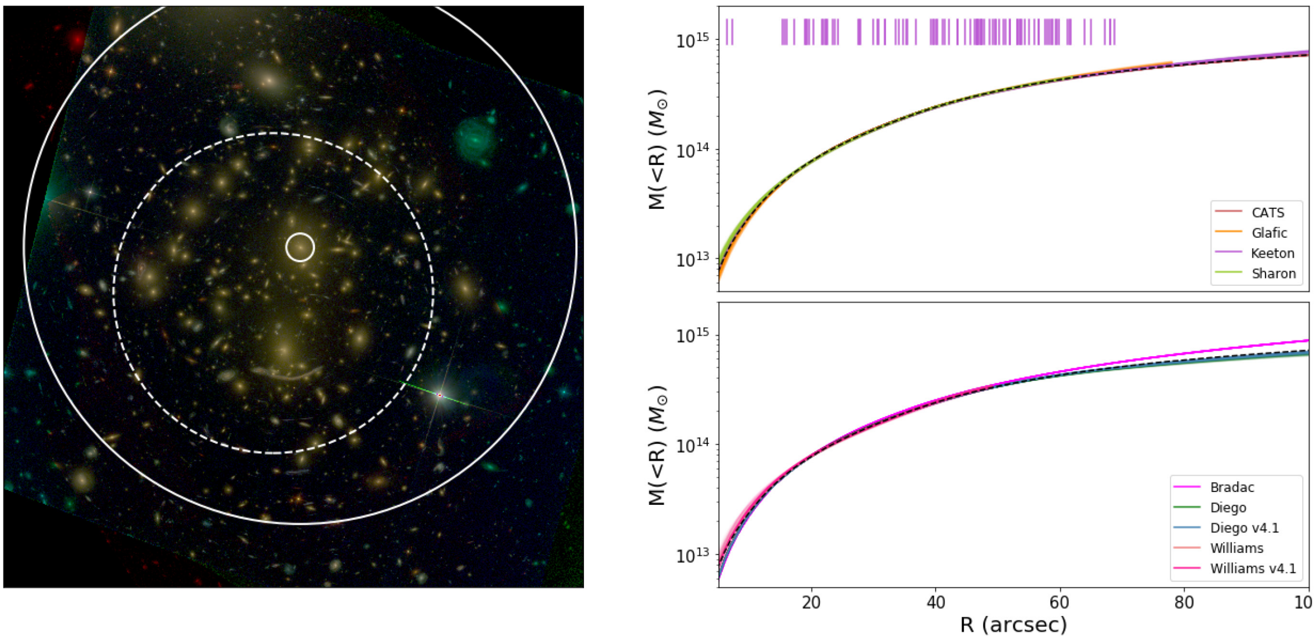

Figure 6. Similar to Fig. 1 for Abell 370.

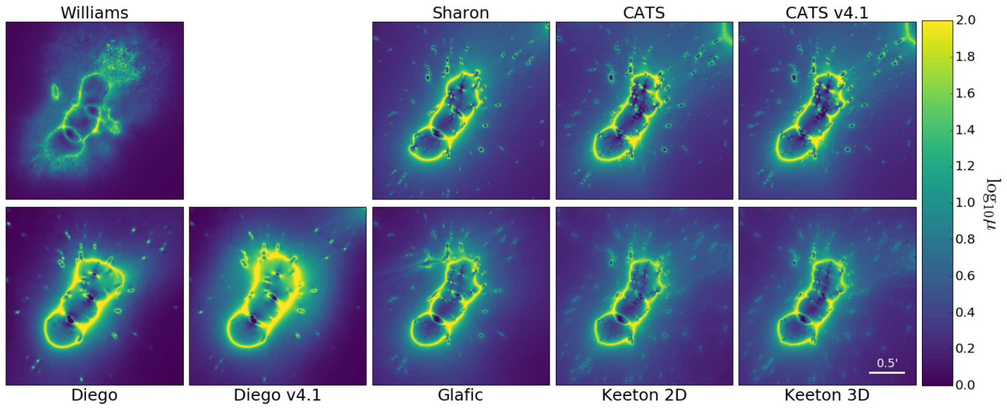

Figure 7. The half-sample mode (HSM) magnification maps for a source at redshift $z = 9$ from the suite of realizations for each model of Abell 2744; each panel is version 4 unless specified otherwise. Every plot covers the same area and is oriented such that up is North and left is East. The overall shape of the critical curves is seen to be mostly consistent between models, but a number of differences exist.

study of high redshift galaxies, but it can also be hard to constrain. It is highly non-linear and a small change in model parameters can produce large changes in the magnification at a specific point, especially if it is close to the critical curves (defined as where $\mu \to \infty$).

In this section, we seek to compare the magnification maps submitted by each team. To do this, we first find the largest area in common between the range maps of all teams and trim the maps to this area; this does sometimes exclude interesting regions that the team(s) with the smallest area did not model, but it is necessary to make a fair comparison overall. We then find the lowest spatial resolution, i.e. highest area per pixel, among the teams and resample all of the maps to this resolution. We use 2D linear interpolation to find values at the same locations in each map instead of rounding to the nearest pixel in order to prevent artefacts, specifically in the 2D histograms.

This yields a data cube comprising the range maps, now with a common area and resolution, for each model. It is not straightforward to analyse such a data set; we would want something that incorporates the errors but also does not ignore the spatial aspect of the maps. Priewe et al. (2017) tackled this problem in various ways for version 3 models of Abell 2744 and MACS J0416, namely looking at ∼200 pixels set in a grid around the cluster core and analyzing the spread in magnification histograms for various magnification bins. While this accomplished the goal of showing the increasing spread of magnifications across the field, the spatial context was mostly lost. That is, if one part of the map showed a higher spread than other parts (say, due to an interloping foreground galaxy), this would not be apparent in a magnification histogram.

4.1.1 Half-sample mode

We analyse the models in two separate ways, the first of which uses half-sample mode (HSM) maps. The HSM is a robust way to find the value of maximum likelihood of a random variable. This method finds the peak of a histogram, which may be non-Gaussian and/or have outliers. It is important that the estimator used be robust to outliers; near the critical curves, small shifts in the mass distribution can cause large shifts in magnifications.

The estimator is found by a recursive method that cuts a sample down to the smallest interval that encloses half of the data until the mode is found (see Bickel & Fruehwirth 2006). This is done for each model, pixel by pixel, across the range of all realizations for that model. In this way, a data cube is condensed to a single map, but errors are still somewhat included. We can then show both the HSM map and a ratio between two HSM maps of different teams to highlight variations.

We note that the HSM technique does introduce a "fuzziness" artefact in the maps, specifically with models that show significant scatter. For example, our models include scatter in the mass-luminosity relation. This causes variations in the critical curves around the galaxies, which can manifest as washed out features in the HSM maps. It is also a prevalent feature in the free-form maps of the Williams team. However, this is also useful: their maps are particularly free outside of the strong lensing region due to the freedom in their methodology but fairly well constrained within this region, which is highlighted by the HSM maps.

4.1.2 2D histograms

Another way to visualize the difference between the models is a 2D histogram. With it, we depict the joint probability distribution $P(\mu_1, \mu_2)$ that model 1 predicts μ_1 and model 2 predicts μ_2 taken across all pixels and between 1000 pairs of maps sampling the range. This is particularly useful in that we naturally sample from the complete set of realizations and thus get a sense of the full range of the models. Since the maps are ~250 pixels on a side, the histograms then have roughly 7×10^7 pixels in the 1000 pairs. We note that many teams have 100 or more realizations of each model, thus the 1000 pairs undersample the full suite, but the results show little to no change if the number of pairs is increased.

It is easy to pick out by eye which models are relatively similar to each other in a 2D histogram. Models with many pixels in common will show high density along the one-to-one line with varying scatter depending on how tightly constrained the parameters are in a given model; if they are not tightly constrained, they fall in a cloud around $\mu_y = \mu_x$. Other differences in the models can result in more interesting features in the 2D histograms. For example, if a model has bimodal characteristics and the realizations fall within two classes, this might appear in the 2D histogram as another track of relatively high density, as opposed to a cloud due to scatter.

4.1.3 Presentation of results

In the following subsections, we first show the HSM magnification maps for the area in common for all of the models. This allows us to look at broad stroke similarities and differences and to compare the overall shape of the models. We try to keep a common structure to the plots for the fields, but there will be some variations due to some teams not modelling all of the fields. The second plot for each cluster shows both the HSM ratio maps and the 2D histograms for easy comparison. The ratio panels are arrayed such that the HSM of the team denoted on the *x*-axis is divided by that of the team on the *y*-axis. Thus, a panel showing mostly red, i.e. positive ratios, indicates that the magnifications in the model of the team on the *x*-axis are higher than those of the team on the *y*-axis.

The 2D histograms fill in the rest of the space left from the ratio map triangle plot nicely. It offers the same combinations of model comparisons, except transposed: e.g. the left-most column corresponds to the bottom row. Having these plots next to each other is quite useful: areas of red, positive values in the spatial maps correspond to the area above the one-to-one line in the 2D histogram. The 2D histograms along the diagonal from the bottom left to the top right show self-comparisons, i.e. both data sets making up the 1000 pairs of realizations come from the same model. This allows us to see what the statistical scatter of a given model is and compare it to the scatter seen among the teams.

4.2 Abell 2744

Six teams produced nine models of this field. The HSM maps of each model are shown in Fig. 7. For this field, the difference between the two CATS models is the same as that between the Diego models: v4 used only GOLD constraints, while v4.1 used GOLD+SILVER+BRONZE.

Based on the HSM maps shown in Fig. 7, all models seem relatively consistent, especially near the core of cluster. This is where one would expect them to be most similar since it is where the bulk of the images are. Some form of a double band structure can be seen in all of the models, caused by the two bright, large cluster members seen in Fig. 1, whose influence is important enough to be captured by the free-form models.

These similarities are encouraging, but there are also clear differences. For example, some models have a halo off to the north-west (upper right in Fig. 7), which the Williams, Diego, and Glafic models do not require. This halo does not seem to be in much agreement among the models which do have it. The two CATS models have a halo with a large critical curve while the models of Sharon and Diego v4.1 prefer a halo with smaller critical curve. Our two models both place a halo in this region with similar Einstein radius, though the 3D model finds one that is more diffuse, sometimes not even producing a critical curve. Both of our models place the halo due west of the top of the cluster critical curve while the other teams put it to the north-west. We find that halo to have a wider range in parameters than the other two large-scale haloes, which causes the blurry edges seen in the HSM map. The two CATS models also put in another halo to the north-east (upper left) which is cut off in the maps shown here. The Glafic model appears to sometimes have a quite elongated halo near one of the galaxies to the east of the top of the cluster critical curve.

To see how these differences compare quantitatively, we show ratio maps in Fig. 8. Immediately, a number of trends can be seen. The free-form versus parametric model comparisons at the upper left of the figure all seem to have a red base, even away from the cluster core. This is due to the free-form models having lower magnifications away from, even if there are higher magnifications near, the core of the cluster, as shown by the Diego models. Recall that in the mass profiles, the Diego v4 and Williams models were lower than the median profile. While the Diego v4.1 model agreed very well with the median profile and those of the parametric models, it is clear that this added mass causes the magnification maps to look very different.

The parametric models show slightly less variations, though the haloes outside of the core affect the ratios. It can be seen in the Sharon versus CATS and CATS v4.1 panels that, though both teams predict a halo to the north-west, there is disagreement in its parameters. It is clear that the Glafic model has no halo to the

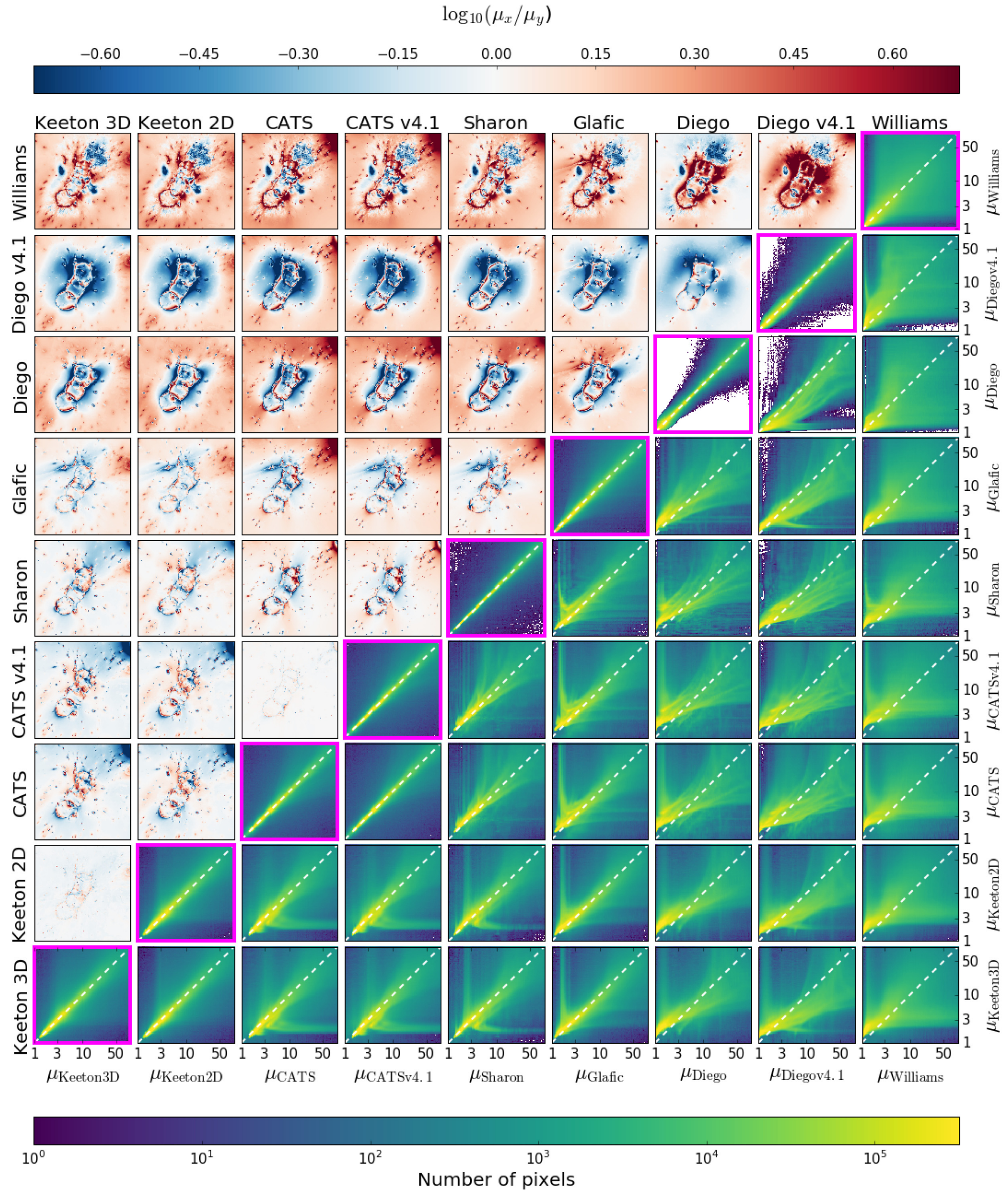

Figure 8. *Upper left:* Ratios of HSM maps for every pair of models of Abell 2744, arranged as Model$_x$/Model$_y$ such that, e.g. the first row is all models divided by the Williams model. Note the colour scale: we use logarithmic values due to the wide range in magnifications. *Lower right:* Two-dimensional histograms showing the probability distribution $P(\mu_1, \mu_2)$ that Model$_x$ predicts μ_1 and Model$_y$ predicted μ_2. Note that for each panel, this is calculated across 1000 pairs of models drawn from the submitted realization maps. Models that are very similar to one another will have a high density along the one-to-one line (white, dashed). Model self-comparisons, i.e. a model versus itself, are plotted along the diagonal and outlined in magenta. The various structures seen in the plots can be explained by differences in mass structures in the models, as described in the text. All panels assume a source redshift of $z = 9$.

west of the cluster, and their elongated halo to the east stands out more clearly here than in Fig. 7.

To determine how the full suites of realizations compare among the models, we also show the 2D histograms in Fig. 8. We see that two models that were very similar in the ratio maps, e.g. CATS v4 and v4.1, produce a 2D histogram that is heavily populated, as expected, along the one-to-one line (dashed white). Some comparisons do not fall along the one-to-one line at all, e.g. Diego 4.1 versus CATS v4.1; others may vaguely fall along this line but have large spreads, e.g. Williams versus Diego.

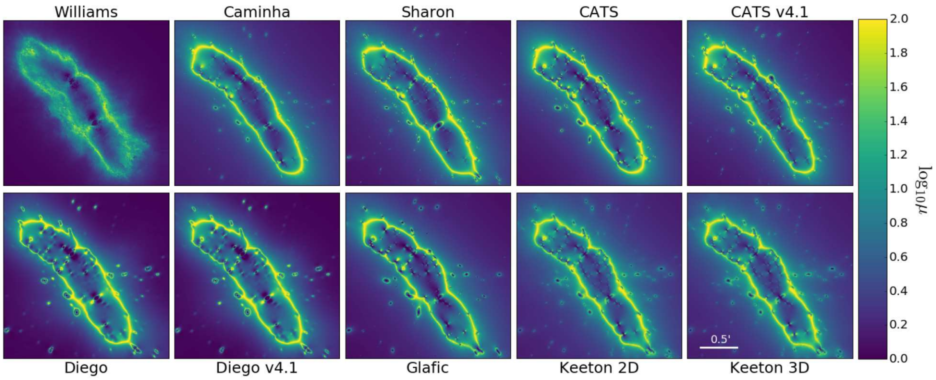

Figure 9. Similar to Fig. 7 for MACS J0416.

The structures that appear in these panels are also informative about the models themselves. For example, in the Sharon versus Keeton 2D histogram panel, there are horizontal and vertical branches that correspond to the extra haloes that the two models include. If only one model has a halo at a certain position, then the model without the halo will have constant low magnification, while the other model will show increasing magnifications around the critical curves. Since the Keeton and Sharon teams have both haloes, but in different places, this creates two branches.

These plots are also important in that they show that, even at low magnifications, the models do not necessarily agree. The panels showing parametric versus parametric models are well populated around $\mu_y = \mu_x$ at low magnifications, but this is not true for the free-form versus parametric model comparisons. This is not surprising, given what we see in the ratio panels; it is also important to point out that much of this is caused by the region outside of the critical curves.

4.3 MACS J0416.1-2403

Fig. 9 shows the HSM maps for the common area among the 10 models produced for this field. It is clear that the teams can agree fairly well on where the critical curves sit. This cluster has the highest number of spectroscopically confirmed lensed images out of the six HFF clusters; most models use around 100 images, though Glafic also includes images without spectroscopic redshifts for a total of 202 images. Just as we saw in Abell 2744, the free-form models here have similar structure to the parametric models, specifically in that they find a bend at the northern BCG.

One of the obvious differences in the models comes from their treatment of the cluster members. The number of members included, for example, varies between the teams, as does how mass is assigned to them. For example, galaxies in the Diego models have larger critical curves than the galaxies in the Caminha model. We also see a difference in cluster members between the two CATS models. In Abell 2744, the difference between the CATS v4 and v4.1 models was the rank of constraints used; in MACS J0416, the difference was which galaxies were included in the model. It is clear from Fig. 9 that CATS v4.1 model included galaxies out to a larger radius and indeed, the v4 model has 98 galaxies while v4.1 includes 178. We see the effects of this choice in the ratio panels of Fig. 10. The CATS versus CATS v4.1 panel shows that there are small differences between these two models, particularly at the northern and southern ends of the cluster, leading to shifts in magnifications.

In Fig. 10, we see that, unlike in Abell 2744, the two Diego models for this cluster agree very well as indicated by the mostly white ratio panel and the very tight 2D histogram. Those models also agree more with the parametric models here than they did in Abell 2744. The parametric models here, other than ours, show interesting dipole patterns in their ratio distributions between the northern and southern ends of the cluster. Nonetheless, they overall agree more with each other than with ours or the free-form/hybrid models.

This is not true when compared to our models, which predict lower magnifications at the northern edge and higher magnifications everywhere else. In Raney et al. (2019), we saw that a model without LOS galaxies was biased low as compared to the 3D model, and here we see that other modelling teams indeed have lower magnifications. This is also borne out in the 2D histograms. When comparing our models against the other parametric models (bottom two rows), the histograms are populated above the one-to-one (white dashed) line; this is not seen in the other panels comparing parametric models.

An obvious feature present in the 2D histograms of Fig. 10 is the vertical or horizontal lines in many of the panels. Something similar was seen in Abell 2744, though with thicker lines; it was caused primarily by differences in the position of a large-scale halo outside of the cluster core. The features here are produced by a similar cause but a different source: galaxy-scale haloes. This causes the features to be more numerous since there are more galaxies than large-scale haloes and thinner due to the typical use of scaling relations when assigning mass.

For example, there are more lines seen in the CATS row than that of the CATS v4.1 due to the former having 80 fewer galaxies. The lines are at different magnifications due to the galaxy's position relative to the cluster's critical curves and thus differing base magnification. Further, the fact that we see this feature only in MACS J0416 and not in other fields, which of course also have galaxies, points to how well the models agree with one another. That is, the features are not getting washed out by differences in the large-scale haloes, as they are in the other fields.

4.4 MACS J0717.5+3745

This field is very complex, as was seen previously in the disagreements among the modelling teams of the mass profiles. Still, we see an overall structure to the magnification maps that is at least somewhat consistent among the models in Fig. 11. The critical curves are vaguely mitten-shaped, with all models agreeing on an arm stretching off to the north-west that aligns with the

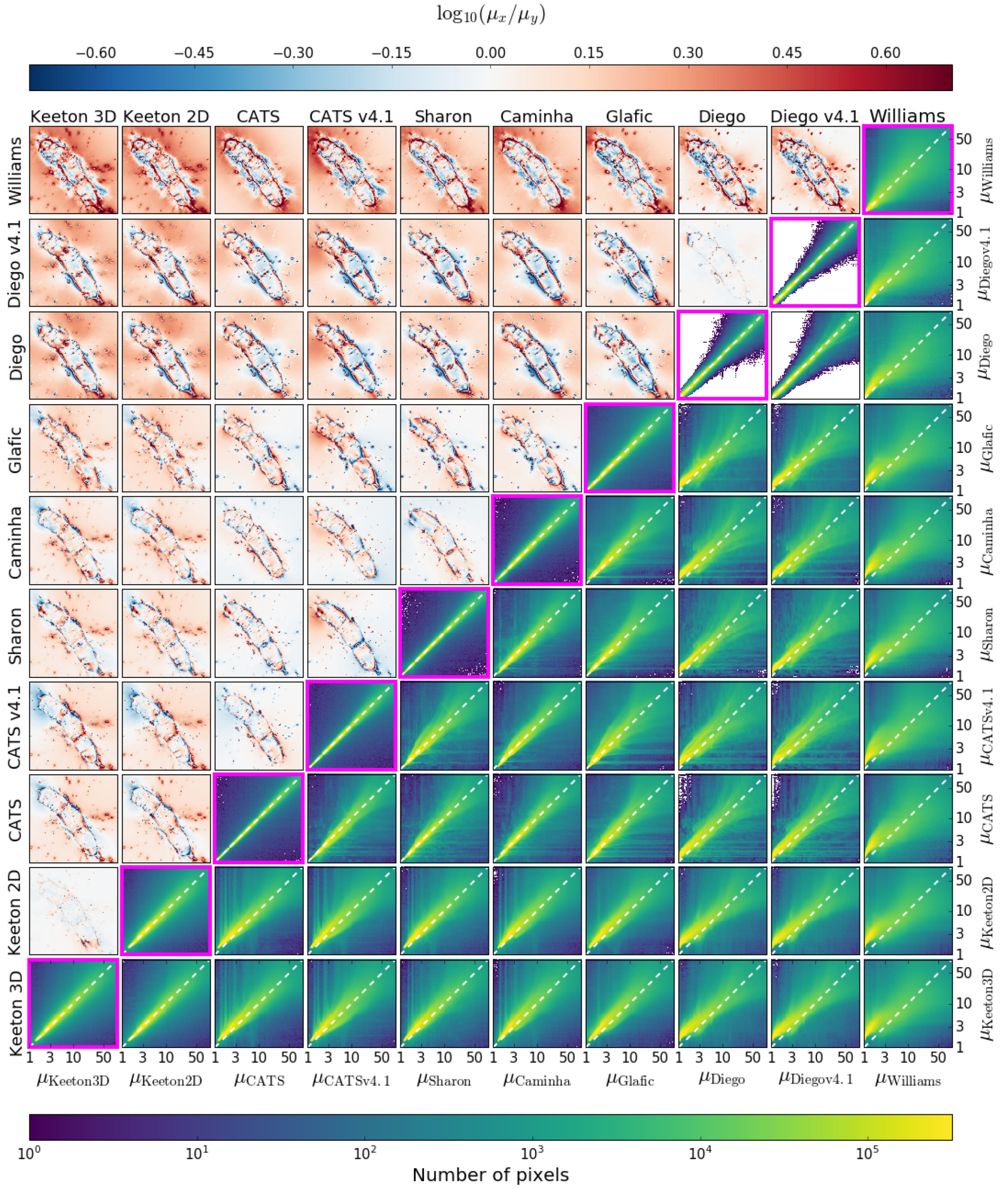

Figure 10. Similar to Fig. 8 for MACS J0416.

possible filament seen in Fig. 3. Contrary to what was seen in the magnification maps of previous clusters, the core of the cluster is not well constrained or agreed upon. This is not surprising, given the large disagreement in mass profiles at smaller radii. Indeed, different models show clear offsets between the positions and number of the main haloes. All of the models except for those from the CATS team place a massive structure in the middle north of the cluster, though with varying importance. Recall: the CATS models vary from one another in whether the main haloes are (v4) or are not (v4.1) cored.

Another clear difference is seen in the galaxy populations. The CATS team included only the most prominent galaxies in their models, while other teams included more to varying degrees. The size of the critical curves for these galaxies also varies greatly among the models. This could be either due to differing placements of the large-scale haloes or by the varying mass prescriptions used by the teams. The area of low magnification to the south-east of the cluster core in the Diego models is centred on a bright foreground galaxy, which causes further differences in the models.

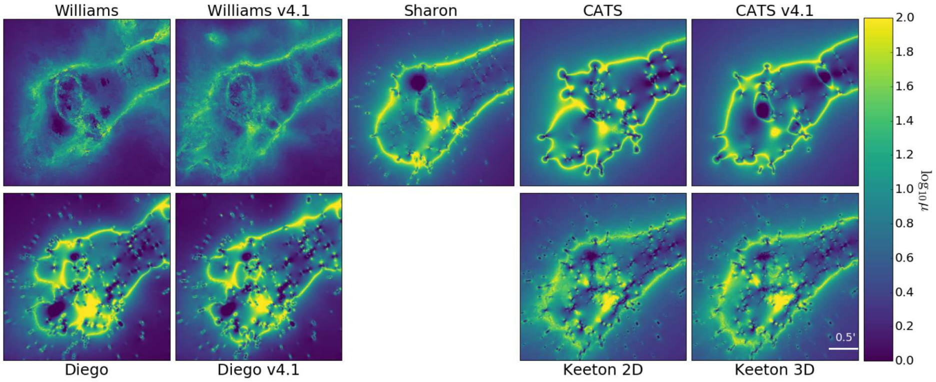

Figure 11. Similar to Fig. 7 for MACS J0717.

The ratio maps, shown in the top left triangle of Fig. 12, are expectedly messy near the cluster core. It appears that our models agree more with the cored CATS model than the non-cored v4.1 model, though the Sharon model seems to disagree with both. The two Diego models disagree more with each other in this field than in MACS J0416 but interestingly not as much as in Abell 2744.

The 2D histograms of the magnifications in Fig. 12 are the broadest of any field, save perhaps for Abell 370. For this field, offsets in haloes do not produce clear structures in the panels, e.g. like the ones seen in Abell 2744. This is due to the fact that the haloes, though they show clear offsets between teams, are still in the cluster core. We saw the structures in Abell 2744 because the haloes of one model fell in a region where the other model did not predict large mass; thus, there was a constant small magnification. If both haloes are offset but overlapping, this will not be the case and instead will cause the 2D histograms to fall in a cloud rather than in nice linear structures. There is also a varying number of haloes between each teams that further smears the histograms out.

4.5 MACS J1149.5+2223

This field had only eight models from five teams in the latest round of modelling, likely because the available data did not change much compared to the previous round. Most of the models for this field agree on the broad strokes: the mass distribution is somewhat complicated, with spurs to the north and south off of a vaguely elliptical structure, as shown in Fig. 13. All models except for Diego v4 agree that this southern region is elongated to some degree, though the Diego v4.1 model shows a more rounded structure than the other models. The Sharon model has a highly concentrated mass component in that area leading to a large area of low magnification. The northern spur is similarly varied, with the Diego model preferring a more rounded structure, while the CATS models have an area of high magnification not seen in the other models.

In the HSM comparison panels of Fig. 14, we see that magnifications outside the critical curves are essentially one for the free-form models, leading to the red box when comparing the free-form versus parametric models, as we saw before. The difference in the southern prong between Sharon and the other teams is clear, leading to areas of high-magnification ratios. The northern region with high magnification in the CATS models likewise shows a clear divergence from other models, which do not have such an area.

The locations of the SN Refsdal images are close to the core of the cluster, near the southeastern edge of the ''belt' of the critical curves. This is in part why the models all agree reasonably well in the middle. It is important to note, however, that those images can only constrain the model at a few points. These models are very complex and can compensate in various ways such that, even if one has images near a dark matter halo at the cluster core to constrain it reasonably well, the models may still disagree on large scales.

4.6 Abell S1063

This field has very small scatter among the mass profiles, and we see this trend continue into the magnification maps shown in Fig. 15. Certainly, the position angle and ellipticity are well constrained, as is the placement of the 'belt' at the position of the BCG, even for the free-form models. Of those, the Diego model matches the shape of the parametric models most closely, though with very large critical curves around their galaxies. The Williams v4.1 model has a larger area of low magnification at the core of the cluster than any of the other models.

The Williams model shows an elongation of the critical curves to the north-east; this horn feature is in the same direction as the elongation seen in the Glafic and, to a somewhat lesser extent, Sharon models. This feature seems to be due to a clustering of member galaxies that are located just out of the bounds of the map; this clustering was also part of the argument by Gómez et al. (2012) for a recent merger, thus making it particularly interesting that the models would differ in their treatment of it.

We note that ours and the two CATS models do not show such an elongation; these models also have only two large-scale haloes, whereas at least the Glafic model includes three. This elongation is further evident in Fig. 16. The Sharon ratio panels show high magnifications compared to all of the other models except for Glafic. Our own models somewhat split the difference between the clustering of galaxies the Sharon, Williams, and Glafic models pick out and galaxies more to the north, similar to, though less drastic than, Diego v4.1. Evidence of this can be seen when comparing our models to the CATS models, which are otherwise very similar in shape.

The CATS models are interesting in that they have lower magnifications outside of the critical curves than the Sharon models or ours. This is also seen in the 2D histogram panels of Fig. 16 as a shift away from the one-to-one line. The free-form versus parametric panels exhibit this behaviour, as in the other clusters, though in this

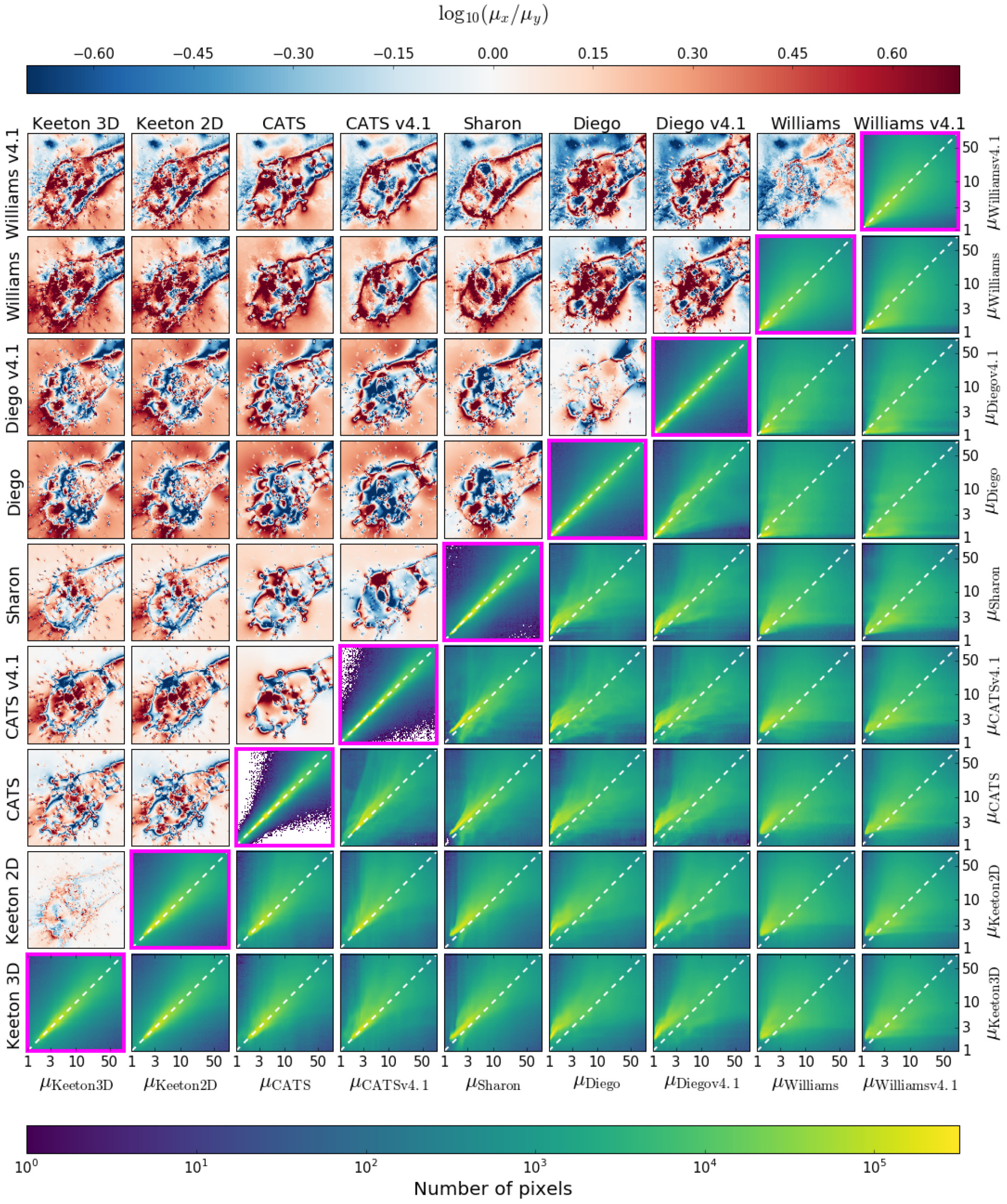

Figure 12. Similar to Fig. 8 for MACS J0717.

case the Diego models also appear to be higher than the Williams models.

4.7 Abell 370

The defining characteristic shared among all of the models of this field is the double-core, as seen in Fig. 17. This duality is caused by the cluster's two BCGs, which are very similar in both size and luminosity. This cluster has perhaps the most spread in the Williams models; recall, since the plots shown in Fig. 17 are HSM maps, the 'fuzzy' nature of a plot indicates that there is wide variation in magnifications in that area among the realizations of that model. The models from Diego and Bradač/Strait do not share this quality and are tightly constrained, though the Diego models are unique in that they do not have the areas of low magnification near the two BCGs. The Bradač/Strait model has only low magnifications in the southern lobe, somewhat similar to the CATS model. The Diego team also did not include as many galaxies in its model of

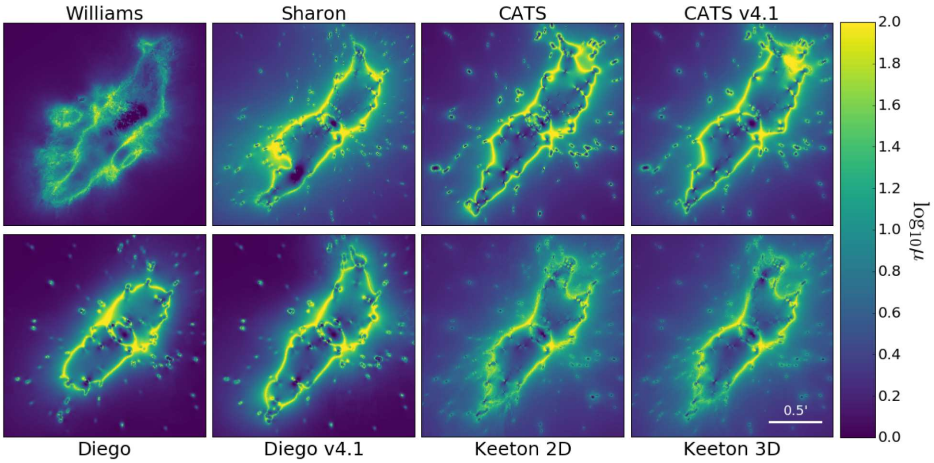

Figure 13. Similar to Fig. 7 for MACS J1149.

this cluster as in other clusters. Our models are different from all others in that they split the southern lobe into two subsections.

The 'crown' of galaxies in the northern region is asymmetric in the Diego models. The CATS model shows a similar bump caused by the critical curves stretching to a background galaxy with a redshift from GLASS of $z = 0.82$ (Schmidt et al. 2014; Treu et al. 2015), as does the Glafic model. We did not include this galaxy due to its distance from images. The Sharon model varies in this region, leading to the fuzzy nature of the HSM map. There is a bright foreground galaxy to the north just out of frame, which our 2D model extends up to while our 3D model does not. The Bradač/Strait model has smoother critical curves in the northern region due to not explicitly including galaxies, though there is a knot in the north-east near a clump of galaxies.

Our models, along with that of the Sharon team, have a larger high-magnification region on the eastern side of the critical curves. This region has such high magnification in our models due to a clustering of galaxies, some of which are foreground galaxies at the same redshift ($z = 0.33$), though it is unclear if they are physically related. It is interesting that all four of the Williams and Diego models place a structure extending to the east of the cluster, which is not really seen in the parametric models or the Bradač/Strait model. This could be a stand in for the cluster members extending off to that side of the cluster, or could perhaps be hinting at some kind of LOS structure that the other parametric models are not taking into account.

The large area of the high magnifications leads to a wide range in the ratio panels of Fig. 18. This is similar to what was seen in MACS J0717, which also had broad swaths of fairly high magnifications. It is important to note that the large differences in the Williams panels are more an artefact of our HSM maps than their modelling process. It is interesting that the Bradač/Strait model is not part of the red block of the other free-form models that we have seen in every field. It could be due to their different modelling process; recall that they employ weak lensing constraints, which would affect the model at large radii.

The 2D histograms offer a similar view of the differences in the models. An interesting characteristic about this cluster is the lack of structure in most of the histograms. This is partially due to the messiness of the cluster, as well as the size, both of which will cause a wide spread in magnifications that leads to a smearing out of the 2D histograms. This was also seen in MACS J0717, another very messy and large cluster. However, that cluster also had the least number of constraints whereas Abell 370 has the second highest number, just under MACS J0416. Yet, the other clusters, barring MACS J0717, have higher agreement between the models. Interestingly, the Bradač/Strait model has a very tight self-comparison 2D histogram; in addition, they have virtually no pixels below a certain magnification, leading to a lot of white in their histograms. We see similar behaviour in the Diego histograms at low magnifications, though not quite to the same extent.

There is more spread in the 2D histograms and structure in the ratio plots for this field than for some of the others. It is clear that this field posed somewhat of a challenge to model, though it is not immediately obvious why. All of the fields in this sample show evidence for a recent or ongoing merger, as evidenced by X-ray studies and/or the fact that they have more than one BCG; Abell 370 is certainly not unique in this regard. However, it is notable that this mass distribution is physically wider than the other fields. For example, Abell 2744, MACS J0416, and MACS J1149 are all fairly thin on the short axis. MACS J0416, which has two BCGs just as Abell 370 does and about as many lensed images, is about an arcminute on its short axis; Abell 370 is around twice that.

5 DISCUSSION

5.1 Mass profiles

One of the ways in which strong lensing can be a tool is in determining how mass is distributed within a system. It is an interesting exercise to see how well lens models can reproduce distributions that are complicated: for example, a cluster in a state of merging. Further, it is important to see how the results differ between modelling teams who use different techniques and density profiles to assign mass to their haloes. Meneghetti et al. (2017) studied this by creating two mock clusters and asking numerous teams to model them as a way to perform a controlled test. The fields were created using two different methods, though it is important to note that both used the light-traces-mass assumption, which is also often employed by parametric modelling methods. They found that the teams using parametric methods were able to reproduce the true mass profiles to within ±2−10 per cent, while the free-form teams had slightly higher scatter of ±5−15 per cent. This was true even

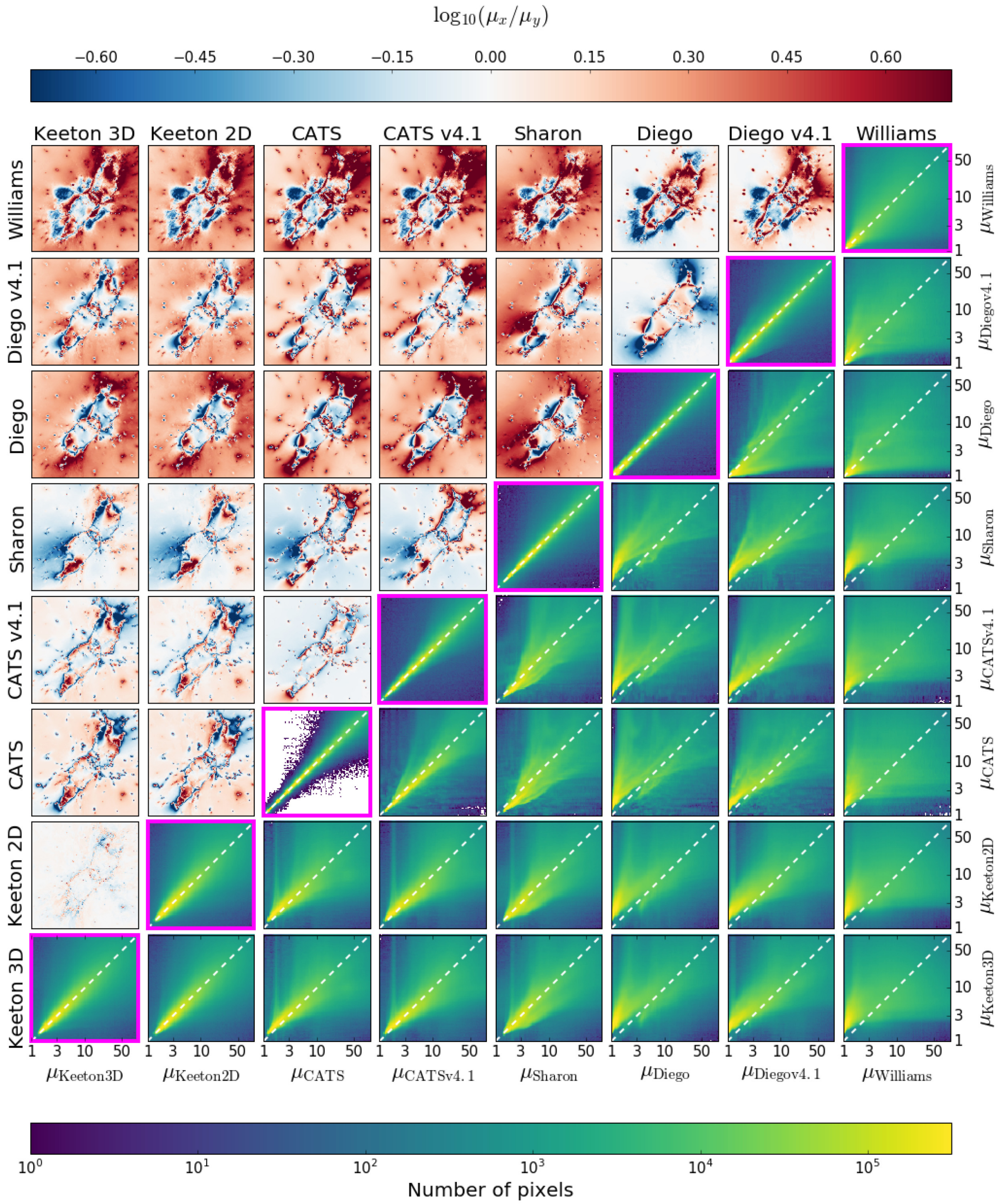

Figure 14. Similar to Fig. 8 for MACS J1149.

though some teams did not use the same density profile as the input mock model did.

In Section 3, we showed the mass profiles for submitted models of each field. From the plots, it was clear that the models were, for the most part, well constrained and showed little scatter between the models. We quantify this in the left-hand panel of Fig. 19 by showing the per cent scatter across all the realizations for each field out to an arcminute from the BCG. This is found by taking the half-width of the 68 per cent confidence interval across all realizations and dividing by the median. The features seen in the mass profiles are also borne out here. For example, MACS J0717 clearly has the largest scatter at low radii, partially due to the cored versus non-cored models of the CATS team, which fit the data equally well. Abell 370 also showed high scatter at low radii, but it quickly falls off to the lowest values of all six fields. At larger radii, MACS J1149 has a scatter that is more than twice the other fields, likely because the area spanned by the lensed images is the smallest of the sample.

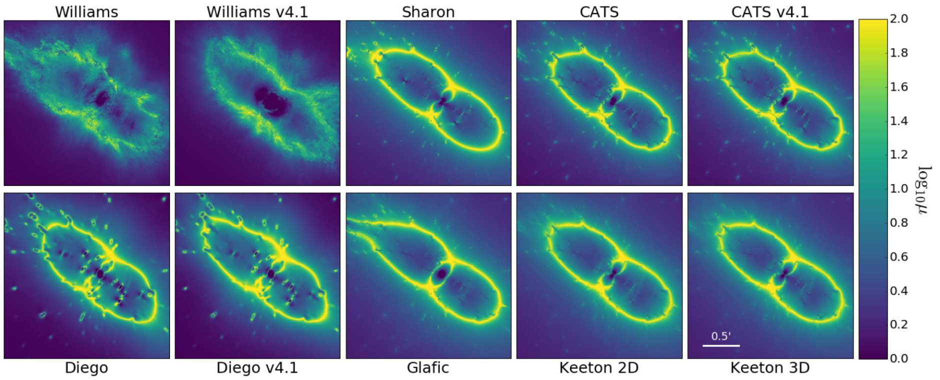

Figure 15. Similar to Fig. 7 for Abell S1063.

Nonetheless, we find that the scatter is quite often below 5 per cent, which is somewhat remarkable given that these clusters are very complicated and often in various stages of merging. In the left-hand panel of Fig. 19, we also show the mean statistical error for all six fields in the black dashed line found by averaging error in a given model using the realizations. Though some of the other curves get quite close to this line, most are indeed above it. This suggests that systematics between the models are more important than statistical uncertainty, which has been a known problem in cluster lensing and which we will again see among the magnifications.

With these models, we can also ask how the mass profiles of the clusters compare to one another. In the middle panel of Fig. 19, we show median mass profiles (across all submitted models) now as a function of physical radius in kiloparsecs, along with 1σ error bars. The error bar is quite large at low radii for MACS J0717, which is unsurprising, given the left-hand panel. However, as radius increases and thus more lensed images are included within the radius, the error shrinks. Across all of the six fields, at 100 (200) kpc from the BCG, the mean enclosed mass is $0.668 \times 10^{14}\,\mathrm{M}_{\odot} \pm 11$ per cent ($1.96 \times 10^{14}\,\mathrm{M}_{\odot} \pm 12$ per cent).

Past studies of simulations (Diemer & Kravtsov 2014) have shown that clusters should be self-similar and thus should also have very similar mass profiles, specifically when scaled by $M_{200\mathrm{c}}$ and $R_{200\mathrm{c}}$. Indeed, a recent study by Caminha et al. (2019) examined clusters from the Cluster Lensing And Supernova survey with *Hubble* (CLASH; Postman et al. 2012) and found just that among profiles of seven clusters, the scatter was only 5–6 per cent.

We sought to test this with our own profiles, as shown in the right-hand panel of Fig. 19. While the two clusters included in Caminha et al. (2019) are quite similar (MACS J0416 and Abell S1063), the others are fairly different. This causes a slight decrease in the average enclosed mass we find as compared to values reported in Caminha et al. (2019). We also find an increase in scatter: around 15 per cent. Interestingly, the scatter is slightly larger in this case as opposed to the unscaled case. We note that this does not include the error in the $M_{200\mathrm{c}}$ or $R_{200\mathrm{c}}$ measurements, which can be ∼ 25 per cent.

5.2 Magnification maps

Among the magnifications, we often do not find the remarkable similarity seen in the mass profiles. Since the goal of the HFF program was to find high redshift galaxies, understanding magnification errors is vital, given that these errors may propagate into luminosity function calculations. Multiple teams were invited to model the fields so that the error in magnification could be estimated by considering the various models. It is important to note that most cluster lenses do not have the same modelling effort behind them. We can then use the HFF models to ask how we might be biasing our magnification estimates by using only one lens model of a given field.

Essentially, we want to find the conditional probability $P(\mu\,|\,\mu_{ref})$ of finding a magnification μ across all models, given that one model predicts a magnification of μ_{ref}. For this analysis, we take a given realization of a model as our reference and find all pixels in that map that have a certain magnification, say $\mu_{\mathrm{ref}} = 3$. We then look at magnifications for that set of pixels across all realizations of the other models. We can repeat this procedure, changing which model and realization we use as our reference, creating a distribution of magnifications. If the models all agree with each other, i.e. if one model has high predictive power for the other models, then the distribution should be tightly constrained around μ_{ref}.

In the left-hand panel of Fig. 20, we show the median of this distribution across all models of the six fields versus the reference magnification from any one given model. The black, dashed line is one-to-one and illustrates magnifications from one model perfectly agreeing with median magnifications across the other models. We find that at low magnifications, one model can predict the median magnification fairly well. However, it does start diverging at higher magnifications. Different fields are affected at different times: e.g. Abell 2744, MACS J0717, and MACS J1149 are farther away from the one-on-one line at $\mu = 10$ than the other three fields. At large magnifications, the difference is large for all six fields.

We note that the curves in the left-hand panel of Fig. 20 are mostly below the one-to-one line, suggesting that, at a given pixel with a high magnification in one model, the other models will predict a lower magnification. This has to do with the non-linear nature of magnification and, specifically, the critical curves. Since magnification drops off quickly as one moves away from a critical curve, you have many more low magnification pixels than high magnification, which causes this bias towards lower magnifications. The effect grows with magnification as well, which causes the flattening of the curves. We explore this further using a toy model in Appendix A.

In the middle and right-hand panels of Fig. 20, we use violin plots to depict the full distribution of magnifications for reference values of 3 and 10. For comparison, we also isolate the statistical scatter via the unfilled violin plots. That is, we now look at all pixels where $\mu = 3$ or 10 for a model and consider the distribution that consists

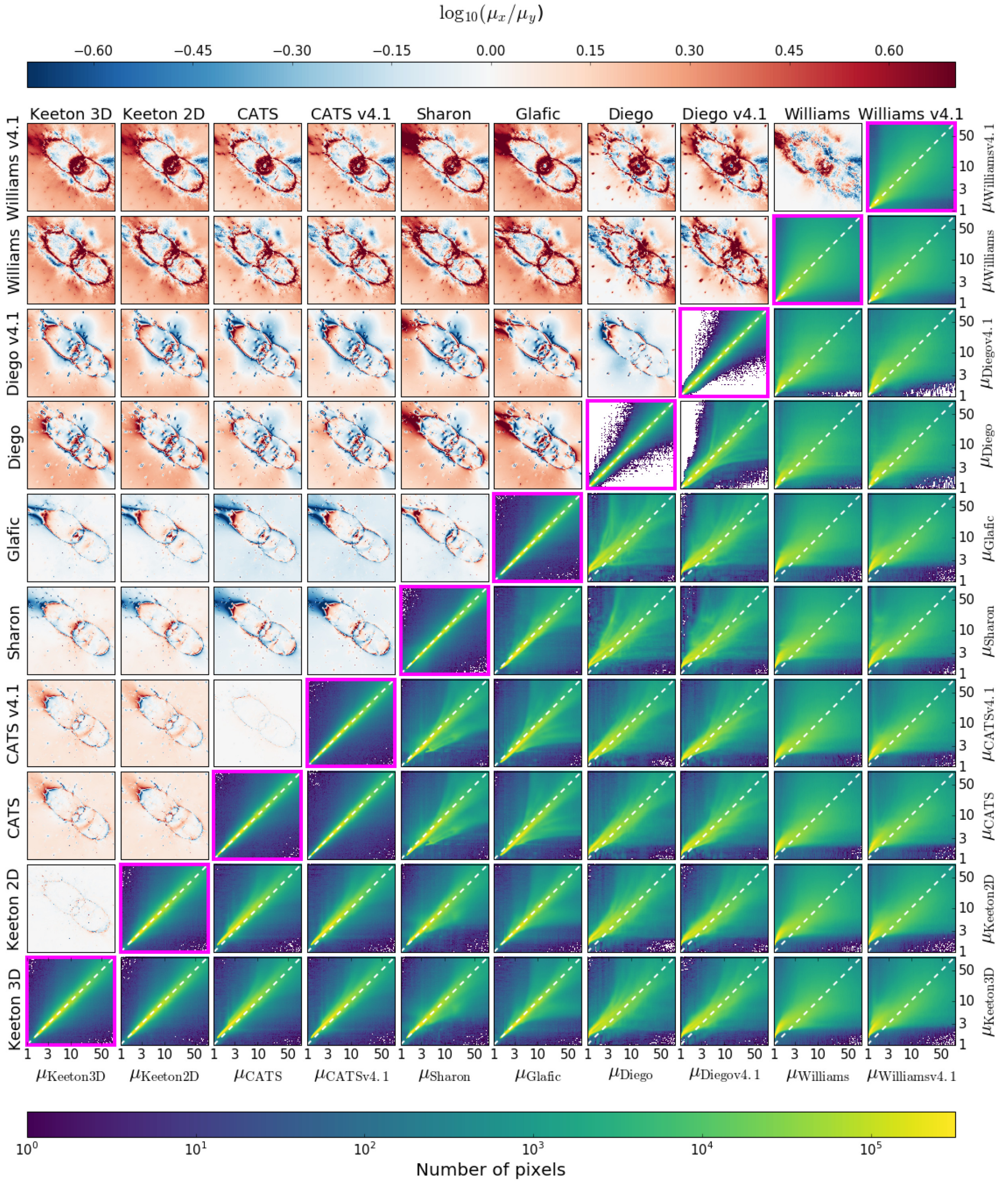

Figure 16. Similar to Fig. 8 for Abell S1063.

of the magnifications at those pixels across only the realizations of that model, as opposed to the realizations across all models (which are shown in the filled violin plots).

As we saw in the left-hand panel, the medians are further away from the correct value at $\mu = 10$ than at $\mu = 3$. One can also see that the scatter is much larger in the higher magnification case. Priewe et al. (2017) did a similar analysis of the results for two fields, Abell 2744 and MACS J0416, from the v3 round of modelling. They found a scatter of 30 per cent at low magnifications ($\mu \sim 2$), which increased to 70 per cent at higher magnifications ($\mu \sim 40$). We find a similar amount of scatter for these fields, along with Abell S1063 for our low magnification case of $\mu = 3$. Abell 370 has a slightly higher amount of scatter at 35 per cent, but the largest scatter lies in our highest redshift clusters, MACS J0717 and MACS J1149, which both show 49 per cent scatter at low magnification. We also note that the average statistical scatter across all six clusters is significantly lower at $\sim$ 6 per cent.

For the higher magnification case, $\mu = 10$, the amount of scatter is, unsurprisingly, even higher. The lowest scatter is seen in MACS J0416 and Abell S1063 at $\sim$ 45 per cent, while the highest is in

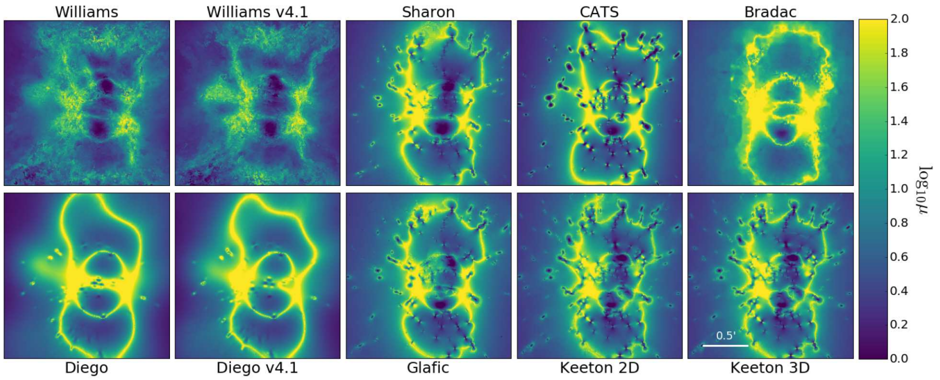

Figure 17. Similar to Fig. 7 for Abell 370.

MACS J0717: 82 per cent. The other three clusters range from 59 ∼ 67 per cent. Statistical scatter is still far below these values, though it does increase: for $\mu = 10$, $\sigma_{\rm syst} = 4.1 \times \sigma_{\rm stat}$, as opposed to $5.7 \times \sigma_{\rm stat}$ for the $\mu = 3$ case.

It is interesting that our results agree with those of Priewe et al. (2017), given that there were significant changes to the constraints of Abell 2744 and MACS J0416 between v3 and v4. Specifically, two surveys utilizing VLT/MUSE greatly increased the number of spectroscopic constraints for the fields. In Abell 2744, the number of image families with spectroscopic redshifts went from 5 to 29 (Mahler et al. 2018) and from 15 to 37 in MACS J0416 (Caminha et al. 2017).

However, this does seem to be in line with the work of Johnson & Sharon (2016), which considered how the number and type of constraints impacted model fits for the two mock clusters presented in Meneghetti et al. (2017). They found that there was a limit to how much additional constraints decreased magnification error in the models; specifically, the decreasing error levelled off around 25 image systems. Further, magnification bias or variance did not correlate with fraction of images with spectroscopic redshifts as long as the constraints included at least five spectroscopically confirmed systems. Those models without any spectroscopic constraints had magnifications biased low; this could be explained by an increase in model variation, which we have previously shown will decrease magnifications. They also found that exactly which image systems are used as constraints can be a bigger source of systematic error than number of spectroscopic redshifts. This could be a important part of the systematic error we see here, given that there is such a wide range in constraint selection between the teams.

Other works have looked at how these errors propagate into luminosity functions, finding various results. For instance, Livermore, Finkelstein & Lotz (2017) found that magnification uncertainties did not have a large effect on the luminosity function, while Bouwens et al. (2017) found that a large uncertainty could produce an artificial steepening of the slope. Atek et al. (2018) used the submitted models of each team to get error bars on their measurements, though we note that this technique would not be possible if there were not multiple modelling teams.

5.3 LOS effects

In Raney et al. (2019), we showed that there can be systematic effects produced when galaxies along the line of sight to a cluster are either not included in the model or their effects are approximated to the cluster lens plane. Specifically, not placing the galaxies at their true redshift could cause a bias in magnifications on the level of 5 per cent or could cause an increase in the scatter of the magnifications. We argued that, while these effects were non-negligible, they were also quite small and unlikely to be the dominant source of error in current models.

In this magnification analysis of this work, we included both the models where galaxies are approximated to the cluster lens plane (Keeton 2D) and the case where these LOS galaxies are placed at their true redshift and thus the model has multiple lens planes (Keeton 3D). We find that the results from our previous work are again supported here. Some small differences can be seen between these two models. For example, in the HSM ratio Keeton 2D versus 3D panel of MACS J0416 (see Fig. 10), there is a knot in the southern part of the cluster where magnifications are quite different that coincides with the location of a bright foreground galaxy. Another example can be seen in the HSM ratio Keeton 2D versus 3D panel of Abell 2744 (see Fig. 8) where there is a very slight blue tinge across the plot; this corresponds to the 2D model predicting lower magnifications than the 3D model, as we saw in our previous work.

Abell 2744 is an interesting case as well because there is actually more of a difference between the Keeton 2D and 3D models than there is between the CATS v4 and v4.1 models. Recall, the difference between the two CATS models is their lensing constraints. This could again support the previous assertion that adding additional constraints does not significantly change the magnifications of the model. If we are indeed in the regime where additional constraints are not useful in further constraining models, it is then worrying that the different models among the teams are not more similar. This could be a problem for future cluster lensing surveys, which will likely not have the same modelling effort the HFF did. We note that this seems to only be true of the parametric models of Abell 2744: the two Diego models, which also vary in constraints used, do show many differences in their magnification maps.

6 CONCLUSIONS AND IMPLICATIONS

The HFF program was a tremendous effort by many. It took a significant amount of observing time, with both *HST* and other telescopes, in order to conduct the photometric and spectroscopic surveys needed. Also, the different lensing teams put in the effort

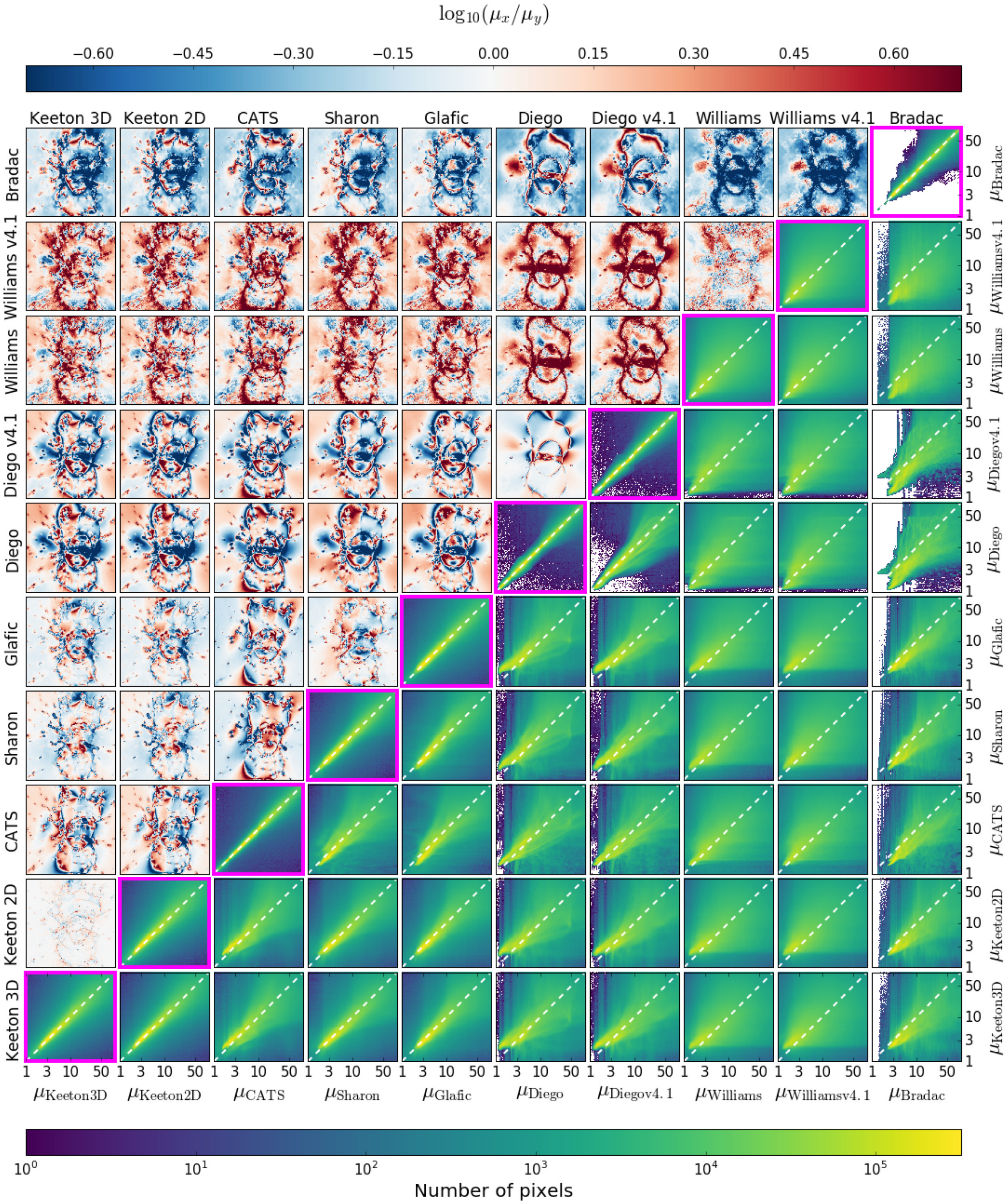

Figure 18. Similar to Fig. 8 for Abell 370.

to find and rank the possible lensed images and, of course, model the fields. Thus, it serves as a wonderful opportunity to compare the results of the models of these fields and see what the state of the lensing field is. Though the HFF program has finished, further cluster lensing surveys are underway: the Reionization Lensing Cluster Survey (RELICS; Coe et al. 2019) and the Beyond Ultra-deep Frontier Fields and Legacy Observations (BUFFALO; Steinhardt et al. 2020), the successor to HFF. We can then use the results from the HFF modelling effort to make improvements going forward.

We chose to compare the models in two ways in this work: circularly averaged mass profiles, derived from the surface density maps, and magnifications. These models came from eight teams using a variety of different methodologies and making various decisions in the modelling process. We considered not just the fiducial models but also the realizations that each team submitted. In this way, we were able to get an idea of how systematic errors compare to the statistical errors of each team. The conclusions we drew can be summarized as follows:

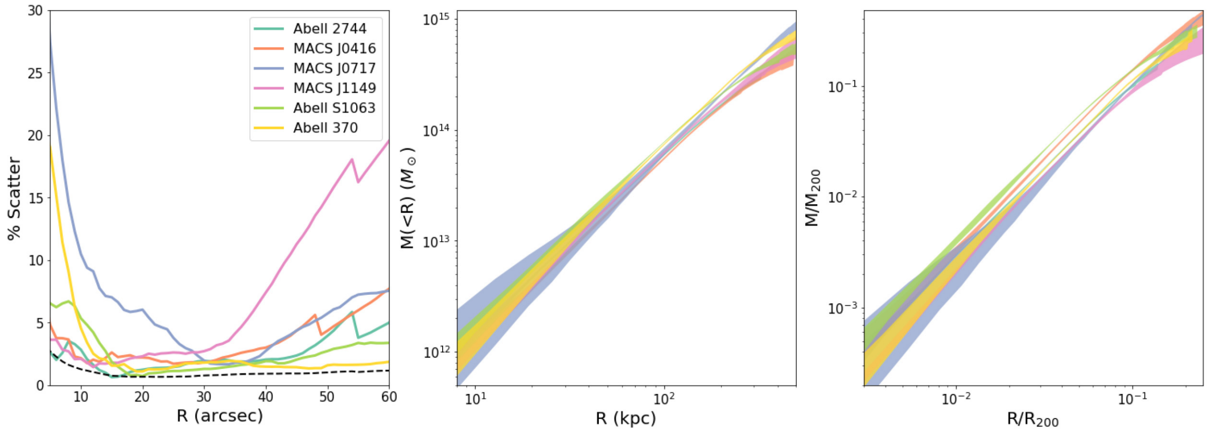

Figure 19. *Left:* Scatter given by the 1σ error across all realizations of the submitted models for each field, as a function of radius in arcseconds. The black dashed line is the average statistical scatter across all models and all fields. *Middle:* Median enclosed mass (given in solar masses) as a function of physical radius in kiloparsecs. The 1σ error bars are taken across all submitted models for a given field. Colours are the same as in the left-hand panel. *Right:* Similar to the left-hand panel, but now scaled by M_{200c} and R_{200c} values. These were obtained from Medezinski et al. (2016) for Abell 2744, Umetsu et al. (2011) for Abell 370, and Umetsu et al. (2016) for the rest.

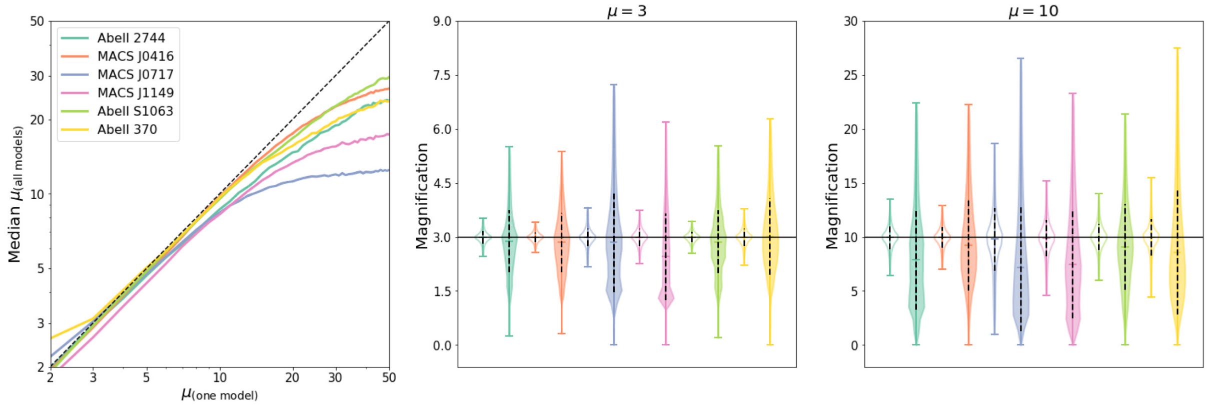

Figure 20. *Left:* For a given magnification, we show how well the magnification predicted by one model tracks the magnification predicted by all models. Specifically, we look at the pixels from one model that have a given magnification ($\mu_{\rm (one\,model)}$) and then find the distribution of magnifications for those pixels across all realizations of the other models, i.e. $\mu_{\rm (all\,model)}$. The plotted curve is then the median of this distribution. If a field has many models with similar magnifications, then its curve will fall close to the one-to-one line (black, dashed). *Middle:* We show the full distribution of $\mu_{\rm (all\,model)}$ for a magnification of 3. The median values will be the same as the left-hand panel, but now the full distribution is shown as well. Unfilled violin bodies are when the magnifications of one model are compared against itself, analogous to the panels along the diagonal in the 2D histograms. Dashed black lines show the 1σ error. Colours are the same as in the left-hand panel. *Right:* Similar to the middle panel, but for a magnification of 10.

(i) The circularly averaged mass profiles are remarkably similar across the models with 1σ scatter often < 5 per cent. This systematic scatter across all models is larger than the statistical error for a given model, though in some cases it is quite close.

(ii) The mass profiles across fields are also notably similar to one another when plotted as a function of physical radius in kiloparsecs, with a scatter of only about 13 per cent. They become less similar when scaled by $M_{200\rm c}$ and $R_{200\rm c}$, and the scatter becomes 20 per cent.

(iii) Magnification maps often show significant differences between teams. If one assumes that a single model is correct and compares magnifications at a given pixel, results will be biased low due to the non-linear nature of magnifications maps. This bias is fairly small at low magnifications, where the median magnification averaged across the six fields is 2.82 for $\mu = 3$. However, the bias increases with magnification: $\mu = 10$ gives a median magnification of 8.22.

(iv) Further, the scatter in these magnifications can be quite high: $30 \sim 50$ per cent at low magnifications and $45 \sim 82$ per cent at higher magnifications. This large uncertainty may propagate into quantities derived using magnification, i.e. intrinsic luminosity or size of the lensed galaxies.

Is it worrying that, even with dozens to hundreds of lensed images per field, the models still show clear disagreements? It certainly suggests that statistical uncertainties have decreased to the point that they are smaller than systematic effects in lens modelling. This is an important lesson because it is not likely that future surveys will have 5+ lens modelling teams to sample the systematic effects. We need to use this opportunity provided by the HFF program to thoroughly understand the systematics in cluster lens modelling and ensure that uncertainties in future surveys are not underestimated.

At the same time, we believe that it is still impressive that such complicated systems can be modelled with the precision seen. Perhaps another lesson involves the choice of systems for detailed study. The models of Abell S1063, a fairly simple cluster, show the most agreement among the teams. While larger clusters such as MACS J0717 may have larger areas of high magnification, they are much harder to study and to constrain the lens models, leading to higher error bars on the luminosities of any high redshift galaxies found. It is an interesting question for the future of cluster lensing: should we focus more on those fields that are large and massive (thus very likely to have elongated areas of high magnification) even though they also may be very complicated due to mergers? Or, instead, would it be better to look at neater fields that are easier to model, even if they lack the lensing power seen in the more complicated systems?

There is much work that could be done in the future. Particularly, there are many sources of systematic error that have not been studied in great detail. Further, which systematic biases are most important (and the strength of those biases, as we showed in our previous work) may depend on the particular cluster. Thus, any study of systematics should ideally be done for more than one or two fields. The next generation of telescopes will be promising for cluster lensing, e.g. *JWST* and *WFIRST* with their IR capabilities to find high redshift galaxies and, in the latter case, a wide FOV to study mass in the cluster outskirts. With more and more data, work into quantifying systematic errors will become vital if we are to use these fields to their full potential as ways to detect and study galaxies from the early Universe.

ACKNOWLEDGEMENTS

We recognize the other Hubble Frontier Fields version 4 modelling teams whose work was used in this paper: Bradac & Strait; Caminha; Richard, Natarajan & Kneib (CATS); Diego; Ishigaki, Kawamata, and Oguri (Glafic); Sharon & Johnson; and Williams. We acknowledge support from Hubble Frontier Field Lensing Support through contract STScI-49745 from the Space Telescope Science Institute, which is operated by NASA under contract no. NAS5-26555. C. A. R. and C. R. K. also acknowledge support from the U.S. National Science Foundation through grant AST-1909217. All version models were obtained via the Mikulski Archive for Space Telescopes (MAST) and the web-based lens model tool. We thank Dan Coe and Keren Sharon for organizing the useful discussions among all modelling teams and contributors. We thank John Hughes and Peter Doze for their helpful discussions about cluster mass profiles. We acknowledge Nicholas James for his work on the 2D histograms, which served as an initial step for some of the plots shown in this paper.

REFERENCES

Abdelsalam H. M., Saha P., Williams L. L. R., 1998, MNRAS, 294, 734
Abell G. O., Corwin H. G., Jr., Olowin R. P., 1989, ApJS, 70, 1
Acebron A., Jullo E., Limousin M., Tilquin A., Giocoli C., Jauzac M., Mahler G., Richard J., 2017, MNRAS, 470, 1809
Andrade K. E., Minor Q., Nierenberg A., Kaplinghat M., 2019, MNRAS, 487, 1905
Atek H., Richard J., Kneib J.-P., Schaerer D., 2018, MNRAS, 479, 5184
Bartelmann M., Maturi M., 2017, Scholarpedia, 12, 32440
Bayliss M. B., Johnson T., Gladders M. D., Sharon K., Oguri M., 2014, ApJ, 783, 41
Bézecourt J., Kneib J. P., Soucail G., Ebbels T. M. D., 1999, A&A, 347, 21
Bickel D. R., Fruehwirth R., 2006, Comput. Stat. Data Anal., 50, 3500
Bouwens R. J., Oesch P. A., Illingworth G. D., Ellis R. S., Stefanon M., 2017, ApJ, 843, 129
Bradač M., Allen S. W., Treu T., Ebeling H., Massey R., Morris R. G., von der Linden A., Applegate D., 2008, ApJ, 687, 959
Bradač M., Schneider P., Lombardi M., Erben T., 2005, A&A, 437, 39
Bradač M. et al., 2009, ApJ, 706, 1201
Caminha G. B. et al., 2017, A&A, 600, A90
Caminha G. B. et al., 2019, A&A, 632, A36
Chirivì G., Suyu S. H., Grillo C., Halkola A., Balestra I., Caminha G. B., Mercurio A., Rosati P., 2018, A&A, 614, A8
Cibirka N. et al., 2018, ApJ, 863, 145
Clowe D., Bradač M., Gonzalez A. H., Markevitch M., Randall S. W., Jones C., Zaritsky D., 2006, ApJ, 648, L109
Coe D. et al., 2012, ApJ, 757, 22
Coe D. et al., 2013, ApJ, 762, 32
Coe D. et al., 2019, ApJ, 884, 85
de Filippis E., Sereno M., Bautz M. W., 2005, Adv. Space Res., 36, 715
Diego J. M., Protopapas P., Sandvik H. B., Tegmark M., 2005a, MNRAS, 360, 477
Diego J. M., Sandvik H. B., Protopapas P., Tegmark M., Benítez N., Broadhurst T., 2005b, MNRAS, 362, 1247
Diego J. M., Tegmark M., Protopapas P., Sandvik H. B., 2007, MNRAS, 375, 958
Diego J. M. et al., 2015, MNRAS, 446, 683
Diemer B., Kravtsov A. V., 2014, ApJ, 789, 1
Dunham S. J. et al., 2019, ApJ, 875, 18
Ebeling H., Edge A. C., Henry J. P., 2001, ApJ, 553, 668
Ebeling H., Barrett E., Donovan D., 2004, ApJ, 609, L49
Ebeling H., Barrett E., Donovan D., Ma C.-J., Edge A. C., van Speybroeck L., 2007, ApJ, 661, L33
Gómez P. L. et al., 2012, AJ, 144, 79
Grillo C. et al., 2015, ApJ, 800, 38
Harvey D., Massey R., Kitching T., Taylor A., Tittley E., 2015, Science, 347, 1462
Hoekstra H., Bartelmann M., Dahle H., Israel H., Limousin M., Meneghetti M., 2013, Space Sci. Rev., 177, 75
Host O., 2012, MNRAS, 420, L18
Ishigaki M., Kawamata R., Ouchi M., Oguri M., Shimasaku K., Ono Y., 2015, ApJ, 799, 12
Jauzac M. et al., 2012, MNRAS, 426, 3369
Jauzac M. et al., 2014, MNRAS, 443, 1549
Jauzac M., Harvey D., Massey R., 2018, MNRAS, 477, 4046
Johnson T. L., Sharon K., 2016, ApJ, 832, 82
Johnson T. L., Sharon K., Bayliss M. B., Gladders M. D., Coe D., Ebeling H., 2014, ApJ, 797, 48
Johnson T. L. et al., 2017, ApJ, 843, L21
Jullo E., Kneib J.-P., Limousin M., Elíasdóttir Á., Marshall P. J., Verdugo T., 2007, New J. Phys., 9, 447
Kawamata R., Oguri M., Ishigaki M., Shimasaku K., Ouchi M., 2016, ApJ, 819, 114
Kawamata R., Ishigaki M., Shimasaku K., Oguri M., Ouchi M., Tanigawa S., 2018, ApJ, 855, 4
Kelly P. L. et al., 2016, ApJ, 819, L8
Kempner J. C., David L. P., 2004, MNRAS, 349, 385
Kneib J.-P., Natarajan P., 2011, A&A Rev., 19, 47
Lagattuta D. J. et al., 2017, MNRAS, 469, 3946
Lagattuta D. J. et al., 2019, MNRAS, 485, 3738
Lah P. et al., 2009, MNRAS, 399, 1447
Liesenborgs J., de Rijcke S., Dejonghe H., Bekaert P., 2007, MNRAS, 380, 1729
Limousin M. et al., 2016, A&A, 588, A99
Livermore R. C. et al., 2012, MNRAS, 427, 688
Livermore R. C., Finkelstein S. L., Lotz J. M., 2017, ApJ, 835, 113
Lotz J. M. et al., 2017, ApJ, 837, 97
Lynds R., Petrosian V., 1989, ApJ, 336, 1
Mahler G. et al., 2018, MNRAS, 473, 663
Mann A. W., Ebeling H., 2012, MNRAS, 420, 2120
McLeod D. J., McLure R. J., Dunlop J. S., 2016, MNRAS, 459, 3812

Medezinski E., Broadhurst T., Umetsu K., Oguri M., Rephaeli Y., Benítez N., 2010, MNRAS, 405, 257

Medezinski E. et al., 2013, ApJ, 777, 43

Medezinski E., Umetsu K., Okabe N., Nonino M., Molnar S., Massey R., Dupke R., Merten J., 2016, ApJ, 817, 24

Meneghetti M. et al., 2017, MNRAS, 472, 3177

Merten J. et al., 2011, MNRAS, 417, 333

Mohammed I., Liesenborgs J., Saha P., Williams L. L. R., 2014, MNRAS, 439, 2651

Molino A. et al., 2017, MNRAS, 470, 95

Murata R. et al., 2019, PASJ, 71, 107

Oesch P. A., Bouwens R. J., Illingworth G. D., Labbé I., Stefanon M., 2018, ApJ, 855, 105

Oguri M., 2010, PASJ, 62, 1017

Owers M. S., Randall S. W., Nulsen P. E. J., Couch W. J., David L. P., Kempner J. C., 2011, ApJ, 728, 27

Postman M. et al., 2012, ApJS, 199, 25

Priewe J., Williams L. L. R., Liesenborgs J., Coe D., Rodney S. A., 2017, MNRAS, 465, 1030

Raney C. A., Keeton C. R., Brennan S., 2019, MNRAS, 492, 503

Rau S., Vegetti S., White S. D. M., 2014, MNRAS, 443, 957

Remolina González J. D., Sharon K., Mahler G., 2018, ApJ, 863, 60

Richard J. et al., 2014, MNRAS, 444, 268

Rodney S. A. et al., 2016, ApJ, 820, 50

Salmon B. et al., 2018, ApJ, 864, L22

Schmidt K. B. et al., 2014, ApJ, 782, L36

Schneider P., Ehlers J., Falco E. E., 1992, Gravitational Lenses. Springer, Berlin

Shu C., Zhou B., Bartelmann M., Comerford J. M., Huang J. S., Mellier Y., 2008, ApJ, 685, 70

Soucail G., Fort B., Mellier Y., Picat J. P., 1987, A&A, 172, L14

Steinhardt C. L. et al., 2020, ApJS, 247, 64

Strait V. et al., 2018, ApJ, 868, 129

Treu T. et al., 2015, ApJ, 812, 114

Treu T. et al., 2016, ApJ, 817, 60

Umetsu K., Broadhurst T., Zitrin A., Medezinski E., Hsu L.-Y., 2011, ApJ, 729, 127

Umetsu K. et al., 2014, ApJ, 795, 163

Umetsu K., Zitrin A., Gruen D., Merten J., Donahue M., Postman M., 2016, ApJ, 821, 116

Vega-Ferrero J., Diego J. M., Bernstein G. M., 2019, MNRAS, 486, 5414

Williams L. L. R., Sebesta K., Liesenborgs J., 2018, MNRAS, 480, 3140

Wong K. C., Ammons S. M., Keeton C. R., Zabludoff A. I., 2012, ApJ, 752, 104

Zheng W. et al., 2012, Nature, 489, 406

Zitrin A., Broadhurst T., 2009, ApJ, 703, L132

APPENDIX A: MAGNIFICATION DISTRIBUTIONS WITH SCATTER

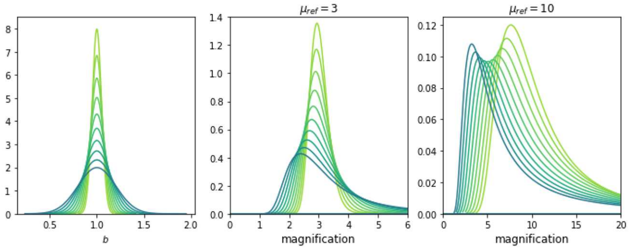

Figure A1. *Left*: Probability distribution function for the Einstein radius b with standard deviation $\sigma_b = (0.05, 0.2)$. *Centre*: Magnification distribution for a reference magnification of 3. As σ_b increase, the distribution shifts towards smaller values and becomes wider. *Right*: Magnification distribution for a reference magnification of 10. We see behaviour similar to the middle panel but more dramatic.

The result in Fig. 20 is perhaps counterintuitive: regardless of which model is chosen as the reference, all other models tend to predict a lower magnification at the reference pixels. To explain this, we consider a toy model of an isothermal sphere. In general, the magnification for an SIS with Einstein radius b is $\mu(r)$ such that

$$\mu^{-1} = 1 - \frac{b}{r}. \tag{A1}$$

Let the reference model have Einstein radius b_0 and consider the radius where the magnification is μ_0. Then for another model with Einstein radius b, the magnification at that same radius is

$$\mu^{-1} = 1 - \frac{b}{b_0}\left(1 - \mu_0^{-1}\right). \tag{A2}$$

Now let b be drawn from a Gaussian distribution with varying standard deviation σ_b, as shown in Fig. A1. The resulting magnification distributions are shown for two reference magnifications, 3 and 10, in the middle and right-hand panels, respectively. As the error in the Einstein radius increases, the magnifications shift towards smaller values. Further, the effect is stronger for the higher magnification case, as was seen in Fig. 20.

For the models of the HFF clusters, there are multiple parameters that have varying error associated with them, not just the Einstein radius parameter. However, this toy model with one source of error still provides a valuable result. Namely, the increase in the error of the parameter disproportionately affects higher magnifications: the highest σ_b (0.20) results in a shift of the median for the $\mu_{\rm ref} = 10$ case to $\mu = 4.1$, while the lower magnification case still has a median magnification of $\mu = 3$.

This paper has been typeset from a TEX/LATEX file prepared by the author.