Considerations, Coordination, and Sharing of Numerical Simulations for Astrophysics

Thematic area: Enabling Foundations for Research

Nick Battaglia[1] NBatta@astro.cornell.edu
Andrew Benson[2] ABenson@carnegiescience.edu
Andrew Hearin[3] AHearin@anl.gov
Katrin Heitmann[3] Heitmann@anl.gov
Shirley Ho[4,5,6] ShirleyHo@flatironinstitute.org
Alina Kiessling[7♦] Alina.A.Kiessling@jpl.nasa.gov
Zarija Lukic[8] Zarija@lbl.gov
Michael Schneider[9] Schneider42@llnl.gov
Elena Sellentin[10] Sellentin@strw.leidenuniv.nl
Joachim Stadel[11] Stadel@physik.unizh.ch

♦Corresponding Author
[1] Cornell University, Ithaca, NY 14853, USA
[2] Carnegie Observatories, Pasadena, CA 91101, USA
[3] Argonne National Laboratory, Lemont, IL 60439, USA
[4] Flatiron Institute, New York, NY 10010, USA
[5] Princeton University, Princeton, NJ 08540, USA
[6] Carnegie Mellon University, Pittsburgh, PA 15213, USA
[7] Jet Propulsion Laboratory/California Institute of Technology, Pasadena, CA 91109, USA
[8] Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[9] Lawrence Livermore National Laboratory, Livermore, CA 94550, USA
[10] Leiden Observatory, Leiden NL-2333, The Netherlands
[11] University of Zurich, Zurich 8057, Switzerland

Considerations, coordination, and sharing of numerical simulations for astrophysics

Foreword

The Tri-Agency Cosmological Simulations (TACS) Task Force was formed when Program Managers from the Department of Energy (DOE), the National Aeronautics and Space Administration (NASA), and the National Science Foundation (NSF) expressed an interest in receiving input into the cosmological simulations landscape related to the upcoming DOE/NSF Large Synoptic Survey Telescope (LSST), NASA/ESA's Euclid, and NASA's Wide Field Infrared Survey Telescope (WFIRST). The Co-Chairs of TACS, Katrin Heitmann and Alina Kiessling, invited community scientists from the USA and Europe who are subject matter experts and also members of one or more of the surveys to contribute. The following is a summary of the ~40-page report from TACS that was delivered to the Agencies in December 2018. The public version of the full report will be posted on the arXiv soon.

Motivation

Numerical simulations have become increasingly sophisticated over the last several decades and their role in cosmological surveys has correspondingly experienced enormous growth. Numerical simulations are now integral to forecasting and survey formulation, in addition to the eventual analysis of the observational data. Developing and exploiting the numerical simulations requires large computing and storage resources as well as people with specialized expertise to develop the modeling and analysis pipelines and database approaches. **Many of the numerical simulation tasks are common between the major cosmological surveys and it is therefore strongly advisable to evaluate common approaches, coordination, and resource sharing between the surveys.** Additionally, investigations of scientific gains that can be reaped from joint pixel analysis efforts have been initiated by the funding Agencies; such investigations rely on the availability of shared synthetic catalogs that can be used across the surveys and are based on the same underlying cosmological simulations.

The shift from Stage 3 to Stage 4 cosmological surveys has been underway for the last several years and during this time the role of cosmological simulations in the surveys has undergone a shift from being a research and development (R&D) effort to being a key element of the Stage 4 survey infrastructure. Elements that are considered part of the survey infrastructure are deemed as *essential to the success of the survey* and have traditionally included efforts like ground operations, analysis pipelines, and data management pipeline development, *but not cosmological simulations.* As a result, key work is difficult to undertake in a timely or planned manner due to the uncertainty of proposal selection. This has resulted in efforts to date being limited to the few groups that have been successful in securing short-term funding and resources for very specific tasks.

Added to this challenge is the reality that students and postdocs working on cosmological simulations and synthetic sky generation have historically had little success in securing permanent jobs in the field. Consequently, the number of people available to contribute to these efforts is consistently low and **the "next generation" are being lost to more secure and higher-paying jobs in data science. This issue should be recognized as a pervasive problem in the field that is deserving of more focused consideration, perhaps by encouraging US National Labs and**

the Agencies to develop a program for more long-term employment options for highly skilled simulators.

This white paper will begin by introducing "extreme-scale" simulations followed by large simulation campaigns, which are the two primary classes of simulation required for upcoming cosmological surveys. Next, the generation of synthetic sky maps and the challenges to this effort are discussed, followed by an analysis of how simulations are essential to investigating and mitigating systematic effects. The role of simulations in developing advanced statistical techniques is then investigated and the document concludes by presenting an argument for the development of a common archival infrastructure to share simulation products.

The purpose of this white paper is to summarize cosmological simulation efforts that are essential to the success of the upcoming Stage 4 cosmological surveys, particularly LSST, WFIRST, and Euclid. The white paper focuses on collaborative efforts that will benefit two or more of the surveys and should make it clear that coordination between the Agencies in providing joint resources between the surveys will enable efficient development and sharing of simulations and related analysis tools. However, the current support for a program of this nature is not well established since these activities are often viewed as survey infrastructure tasks rather than as a broader research and development activity. Consequently, **funding that, in particular, supports work across surveys (and therefore Agencies) is sparse. The most promising solution is a focused collaboration between the surveys and Agencies that will enable the most efficient use of resources and will facilitate rapid development in key areas that are currently experiencing only moderate progress due to this lack of support.**

1. Extreme-scale Simulations

Extreme-scale simulations include both very large, high-resolution N-body simulations (also called "grand challenge" simulations) that form the basis for synthetic sky maps and very detailed, cosmological-volume hydrodynamic simulations that are important to advancing our understanding of astrophysics systematics. These simulations require major computing allocations (in the U.S. for example, the DOE INCITE – Innovative and Novel Computational Impact on Theory and Experiment – Program provides opportunities to apply for such allocations at the Leadership Computing Facilities) in the tens of millions of hours (exact numbers depend on the supercomputer) and access to a supercomputer with a performance that is close to the top 10 supercomputers in the world[1]. By the very definition of extreme-scale, only a handful of these simulations will be available in the world at any one time due to the high cost of running the simulations and storing the outputs.

Gravity-only simulations: Gravity-only simulations at large volume (> 3 h^{-1} Gpc on a side) and high mass resolution (around $10^9\ h^{-1}\ M_\odot$, more optimal would be $10^8\ h^{-1}\ M_\odot$) are extremely important for LSST, Euclid, and WFIRST to enable the generation of detailed synthetic sky maps. Currently, two simulations are available – the Euclid Flagship simulation (Potter et al. 2017) and the Outer Rim simulation (Heitmann et al. 2019) – that are used for this purpose and are close to the ultimately required mass resolution and volume for generating these maps. With the advent of the next-generation supercomputers (e.g., Summit at the Oak Ridge Leadership Computing Facility in the winter of 2018/19) the remaining needed increase in resolution should be

---

[1] https://www.top500.org/lists/2019/06/

achievable relatively easily. The storage required to host the primary data products from these simulations' ranges from hundreds of terabytes to tens of petabytes. Currently two codes are being actively developed to carry out these extreme-scale simulations: PKDGRAV3 (Potter et al. 2017) and HACC (Hardware/Hybrid Accelerated Cosmology Code; Habib et al. 2016). Sharing the results from these simulations is highly desirable as these simulations are very computationally expensive to produce, analyze, and store, and there are very few people with the computational and astrophysical expertise to undertake this effort. **However, sharing will require an infrastructure support investment to enable sharing of the simulation data and to enable the collaborations to generate synthetic catalogs given the different approaches used by the two codes to carry out analysis tasks.**

Hydrodynamic Simulations: Unlike gravity-only simulations, hydrodynamic simulations are far from the ultimate goal with respect to achieving large, cosmological volume simulations at high resolution with reliable physics implementations. Not even the next-generation of supercomputers will rectify this situation, although some progress is being made to (at least) generate consistent results across codes at moderate scales. Currently, large-volume (> 100 h^{-1} Mpc on a side), high-mass-resolution (< 10^6 h^{-1} $M_\odot$) are largely out of reach due to challenges with developing the simulation software to undertake efficient load-balancing and scaling on the supercomputers. Most hydrodynamic codes do not scale efficiently to utilize the full machines available today. An even more serious concern arises due to uncertainties in the current sub-grid model implementations. The necessity of using relatively crude sub-grid models precludes truly first principle predictions and therefore makes it very difficult to use simulations for the purpose of understanding astrophysical systematics. These systematics will ultimately be the limiting factors to improving cosmological constraints in upcoming surveys like LSST, Euclid, and WFIRST. Therefore, it is crucial to have concerted support across the surveys for improving hydrodynamic simulation capabilities. Efforts are needed to help bridge the work carried out on the smallest scales to the larger-volume, cosmologically-relevant, simulations. Detailed studies of sub-grid models must also be carried out to improve our understanding of baryonic effects. The most effective studies will come from multi-wavelength comparisons including cross-correlations with observables for which hydrodynamic simulations make testable predictions. Sharing the results of hydrodynamic simulations is much easier than for the gravity-only simulations due to their current limitations in size. **Therefore, in order to make progress in the field of hydrodynamic simulations, emphasis should be placed on supporting code development efforts, the calibration of sub-grid models, and public access to the simulations to enable wide utilization and cross-comparisons.**

1.1 Recommendations for Extreme-scale Simulations

Gravity-only Simulations: It is important emphasize that numerical simulation cross-survey work is currently not explicitly supported by any of the funding Agencies and usually only occurs if the contributing scientists belong to more than one project. The demands that each survey puts on members of the simulation team are already very high and the efforts are not supported sufficiently within each survey to begin with. **Thus, we recommend that the funding Agencies coordinate on new funding lines for cross-survey infrastructure support that includes in-kind contributions of supercomputing time, storage space accessible across the collaborations, and**

people support to run and analyze the simulations and to develop an infrastructure that allows for easy data access.

Hydrodynamic Simulations: The tasks discussed above fall under the Agencies pre-existing R&D models for numerical projects. In addition, the Agencies existing grant and award solicitations are sufficient to support these efforts in the near-term. **However, we recommend that the Agencies emphasize such proposals in grant programs including, but not limited to, NSF-AST, NASA-TCAN, NASA-ATP, and career awards. We also encourage the Agencies to fund multiple proposals in these solicitations to diversify the code development, sub-grid modeling, and comparison efforts.** The initial funding selection for such efforts is critical to begin as soon as possible to have new sub-grid models tested and implemented. These hydrodynamic simulations are essential for the systematic mitigation and cross-correlation measurements for LSST and Euclid, and thus need to be completed by the time of first light for these surveys. A second round of funding will be necessary to further develop sub-grid models for WFIRST and to update them with the new observations and tests provided by LSST and Euclid.

2. Large Simulation Campaigns

Linking measurements of upcoming surveys to physical model parameters requires very demanding forward simulations, which evolve the universe from early times to the present day. Extracting precision cosmological information from surveys depends upon extending existing modeling capabilities further into the small-scale nonlinear regime as well as rigorous marginalization over currently unknown physics. In practical terms, it means that no single simulation can be sufficient for inferring new cosmological insights from observations, but that large simulation campaigns producing ensemble runs, while varying cosmological and other parameters, are needed. At this point, only a few such emulation projects have been carried out, mostly focusing on statistics that are easily extracted from N-body simulations, such as the matter density power spectrum and more recently galaxy-related statistics. In future, those emulators closer to direct observable statistics will become crucial, including galaxy (photometric)-shear, galaxy-galaxy (photometric) correlation, shear-CMB cross-correlation, shear-CMB lensing cross-correlation, galaxy-CMB cross-correlation, and others. While no simulation in the ensemble will be at the precision of the extreme-scale simulations discussed in the previous section, they are still computationally costly and require significant resource allocations on modern supercomputers. Data produced by those ensemble runs can be many petabytes in size, matching or even surpassing the data volume produced by the extreme-scale numerical simulations. In addition, the analysis of the suites of cosmological simulations is complex if the aim is to directly compare or apply them to the analysis of the observational data.

There are several challenges connected to generating large simulation ensembles. Some of these challenges are the similar to those for the extreme-scale simulations, but additional challenges arise due to the complexity of handling and organizing a large number of simulations. As for the extreme-scale simulations, securing computational resources, allocations as well as storage, to enable the runs themselves is difficult. However, the advantage is that each individual simulation is relatively small, so many more supercomputing facilities can be engaged to carry out such simulations. At the same time, if one wants to take full advantage of a range of computing resources, there are major challenges for the simulator related to running and monitoring the simulations across multiple facilities.

The major challenges for carrying out large ensemble runs are:

- Securing computational resources (allocations, storage) to enable the runs themselves.
- Developing analysis tools to efficiently extract a range of measurements from the simulations to enable the construction of emulators.
- Building workflows that enable management for running and analyzing very large numbers of simulations (potentially across multiple facilities with varying architectures and requirements).

2.1 Recommendations for Large Simulation Campaigns

LSST, Euclid, and WFIRST all require ensembles of simulations that span cosmological and nuisance parameters to be able to fully exploit the cosmological information available from each survey. **The funding Agencies should coordinate on a new investment to support the design, production, and storage/hosting of these simulation ensembles to reduce the overall cost and effort to each individual survey.**

3. Generation of Synthetic Sky Maps

There are a wide variety of methods for producing synthetic sky maps and there are many parallel efforts currently underway (using the same base simulation in many cases). Listed here are the broad categories of methods in descending order of the computational expense to generate a single synthetic sky:

1. Hydrodynamical simulations of cosmological volumes (for a recent review article, see Somerville and Davé, 2015, and references therein).
2. Semi-analytic models (SAMs; for a recent review article, see Somerville and Davé, 2015, and references therein) are grafted into gravity-only N-body simulations and require significant post-processing.
3. Empirical models (for a recent review article, see Wechsler and Tinker, 2018, and references therein) are also grafted into N-body simulations and are statistical in nature.
4. Approximate N-body methods employ various analytical techniques to circumvent the need for a full simulation.

Current and planned large-scale structure surveys most commonly employ empirical modeling and SAMs in the generation of synthetic skies supporting the survey. Synthetic sky map models are already able to meet the goal of being applied to the current generation of base gravity-only simulations, although in some cases this requires considerable computational resources. As the gravity-only simulations increase in volume and resolution, the sizes of the required synthetic sky maps increases, the demands for modeling of additional quantities increases (particularly for multi-survey modeling), and the computational demand of synthetic sky map production also increases significantly. While these demands will be met in part by the next generation supercomputers, significant investment of effort in code optimization, and development of statistical techniques to reduce computational demand will be crucial.

One rarely discussed issue is that **the funding structure of large cosmological surveys provides insufficient professional incentive to carry out the work of developing and generating synthetic sky catalogs.** Currently, individual groups within a survey compete with each other to provide the synthetic mock that is singled out as the “flagship” or “standard” catalog of the collaboration. As generating these synthetic sky catalogs is a fairly specialized scientific activity, ***it is common for the graduate students and postdoctoral researchers involved to struggle to***

advance to the next career stage within the field. This competition-based funding model has thus far resulted in closed-source software packages with only modest applicability beyond the specific survey for which each package was tailored.

3.1 Recommendations for Synthetic Sky Maps

Our assessment is that meeting the cross-survey goals for LSST, Euclid, and WFIRST requires a sustained effort to develop a scalable modeling platform with natural extensibility to multi-wavelength cosmological data. This platform needs to be developed in close contact with each survey's scientific working groups, and the code base also needs to be open-source and adaptable to suit the needs of the specialized analyses within each survey. We consider it unlikely that any such framework will emerge in the absence of a new channel of stable, long-term funding dedicated to supporting the effort and **recommend a coordinated investment into joint-survey synthetic sky catalog code development, generation, calibration, and validation from the funding Agencies.**

4. Astrophysical and Theoretical Systematic Effects

Astrophysical systematics are common across all surveys and developing the required systematics mitigation strategies to optimize the science return of LSST, WFIRST, and Euclid, requires an integrated effort that includes simulations, observations, and analytical descriptions. It is important to note that astrophysical systematics are both correlated with each other and with cosmological observables. **As a result, the common approach of developing mitigation strategies for each systematic independently will have very limited success.** To motivate a coordinated effort across the surveys in this area, only non-survey-specific systematics are considered in this white paper.

Observations from precursor surveys such as the Dark Energy Survey (DES), the Kilo Degree Survey (KiDS), and the Hyper Suprime Cam Survey (HSC) – in combination with CMB and spectroscopic surveys like the Baryon Oscillation Spectroscopy Survey (BOSS) and later the Dark Energy Spectroscopic Instrumet (DESI) – provide information on astrophysical systematics e.g. intrinsic alignments, baryonic effects, galaxy bias, non-linear evolution of structure formation, and projection effects. These systematics can be modeled through analytical expressions, which are then incorporated into numerical simulations in two ways: 1) via post-processing of N-body simulations as part of synthetic sky generation or 2) through fine-tuning of sub-grid physics in hydrodynamical simulations. The increased precision of these simulations will enable an improved interpretation of LSST, WFIRST, and Euclid data. This will be an iterative process that requires close interaction of observers, theorists, and simulators, which is necessary to avoid cosmological surveys being dominated by astrophysical systematics in future. It is important that this iterative process not double-count information, i.e. a thorough process must be developed to ensure that the data used to inform and improve the simulations is not also the data subsequently analyzed with said improved simulations. Ensuring that the information used in the systematics simulations remains independent of LSST, Euclid, and WFIRST data is critical.

Baryonic Effects: As optical and near-IR imaging surveys push the measurements of galaxy clustering and weak lensing into the non-linear regime, it is important to understand effects at smaller scales. For weak lensing, the signal is mostly concentrated in the smaller scales and thus **accounting for baryonic effects on the matter density power spectrum is critically important to provide unbiased cosmological parameter inference** (e.g. Semboloni et al., 2013, Zentner et al., 2013, Eifler et al., 2015). For galaxy cluster science, the need to characterize the clusters with baryonic physics is critical if they are to be used to provide unbiased cosmological constraints (see e.g. Bocquet et al., 2016, for a study on the impacts of baryons on the halo mass function). **More studies into the effects of baryons are needed and initiating a joint program across the surveys to tackle this question would enable detailed comparisons and studies** of the (very different) sub-grid physics models that are employed in these simulation efforts and how they affect the cosmological observables.

Intrinsic Alignments: Cosmic shear is typically measured through two-point correlations of observed galaxy ellipticities. In the weak lensing regime, the observed ellipticity of a galaxy is the sum of its intrinsic ellipticity and gravitational shear. If the intrinsic shapes of galaxies are not random, but spatially correlated, these intrinsic alignment (IA) correlations can contaminate the gravitational shear signal and lead to biased measurements if not properly removed or modeled (see Troxel and Ishak, 2015, and Joachimi et al., 2015, and references therein for reviews). **It is critical for the future to refine numerical simulations with the latest observations and to forecast the impact of IA on LSST, WFIRST, Euclid analyses**, and to further refine this iterative approach to improve IA modeling. In this context it is of particular interest to study the correlations between IA uncertainties and galaxy-halo and baryonic modeling uncertainties and to develop a joint description of these intertwined astrophysical phenomena.

The Nonlinear Regime: The nonlinear regime of structure formation holds a wealth of cosmological information (see e.g. Krause and Eifler, 2017). LSST, WFIRST, Euclid have great potential to exploit this information if accurate predictions are provided well into the nonlinear regime. This task is difficult, both due to baryonic physics that alter predictions on small scales and the challenges of generating high accuracy, gravity-only simulations across cosmologies. To this end, the nonlinear evolution of dark matter on large scales can be treated in different ways. One is using perturbation theory, which has been the default method when interpreting galaxy clustering in redshift surveys. Another method employs phenomenological fits to N-body simulations based on the halo-model (e.g. Takahashi et al. 2012), or emulators of the actual N-body power spectrum measurements (Lawrence et al., 2017). A third approach is full forward modeling, where simulations are rapidly produced (using fast approximate codes) and compared with observational datasets directly (e.g., Agrawal et al., 2017). A fourth approach includes using machine learning directly to predict cosmological parameters from the large-scale structure to very small scales (Ravanbakhsh et al., 2017). All of these methods must be refined to reach the accuracy required for upcoming surveys. **A joint effort to investigate the validity of these approaches, the most efficient implementation, and spatial reach at a given accuracy would be highly valuable across LSST, Euclid, and WFIRST**.

4.1 Recommendations for Systematic Effects

Accelerating progress in this area can already be achieved within the funding Agencies pre-existing R&D models. **We recommend that the funding Agencies prioritize investigations of systematics *and their cross-correlations*.** Multiple research teams should be competitively selected and these teams should include experts who are cross-institutional and cross-survey; experts from precursor surveys (e.g., BOSS, DES, eBOSS, HSC, KiDS, SDSS, VIPERS, etc) and from external data sets (e.g., CMB, X-ray, SZ); and experts on numerical simulations, analytical modeling, theory, and observations.

5. Advanced Statistical Methods

The analysis of cosmological data and simulations relies on using the most sophisticated statistical methods available today. The input to many of these methods are large numbers of gravity-only simulations, as discussed in Section 2. Due to the high cost of these simulations, **it is critical to study statistical methods that help to reduce the number of required simulations**. Examples include the development of emulators (predictions tools) from a limited set of high-quality simulations spanning a range of cosmological parameters or new modeling techniques for covariance estimates to reduce the number of realizations needed.

Emulators: The creation of each virtual universe – for a given set of cosmological and marginalization parameters, as well as the particular random realization of the initial density fluctuations – requires an extreme-scale simulation as discussed in Section 1. In order to make cosmological inverse problems practically solvable, constructing a computationally cheap surrogate model or an emulator is imperative. To meet future survey requirements, the next generation of emulators needs to exhibit progress in the following ways: (1) to have an iterative instead of a fixed design; (2) to be multi-fidelity capable, meaning the ability to combine simulations done at different fidelities; and (3) use multi-level emulation via separating design into "expensive" (e.g. cosmology parameters) and "cheap" parameters, like those appearing in postprocessing runs, responsible for predicting different luminosities or galaxy types from the density field.

Covariances and Likelihood Functions: Methods to obtain covariances can be broadly structured into 3 different categories: 1) analytic covariances, 2) covariances estimated from numerical simulations, and 3) covariances estimated from the data directly.

Analytic covariance matrices are computationally feasible for large, multi-probe data vectors (see e.g. Krause and Eifler, 2017), but there is an open question about whether higher-order moments of the density field are captured precisely enough for clustering, galaxy clusters, and other probes using analytic descriptions. For simulated covariance matrices, the choice of estimator is important. The common choice of a sample variance estimator assumes a Gaussian likelihood, which requires an inverse covariance. The inverse of an estimated covariance is not the same as the estimated inverse covariance and even minute residual noise can severely bias the inverse covariance. Hartlap et al. (2007) described a way to correct for this, but later studies by Taylor et al. (2013), Dodelson and Schneider (2013), and Taylor and Joachimi (2014) found that of order ~$10^6 – 10^8$ simulation realizations are required to achieve an inverse covariance with an acceptable precision. Recently new Hybrid estimators (combining analytic and simulations and data) have emerged (Friedrich and Eifler, 2018) and linear and nonlinear shrinkage estimators are being explored (e.g. Joachimi, 2017) which have substantially reduced these estimates and further reductions are possible via data compression.

The functional form of the likelihood being a multivariate Gaussian has been questioned in the literature, mostly in the context of weak lensing (Hartlap et al., 2009, Wilking and Schneider, 2013, Sellentin et al., 2018), but the same argument holds for galaxy clustering, galaxy-galaxy lensing and other largescale structure probes. The core argument is that summary statistics derived from a non-Gaussian field have no first principle reasons to follow a multivariate Gaussian likelihood.

Alternative approaches such as estimating the likelihood from simulations directly, or utilizing likelihood free analysis techniques such as Approximate Bayesian Computation are still in their early phase of exploration and require targeted research funding to mature fully as alternatives. **The necessity of abandoning the multivariate Gaussian likelihood function as an assumption needs to be established first.** Currently the literature does not conclusively state whether this approximation fails at the level of precision for LSST, WFIRST, and Euclid.

Discrepancy Metrics: Developing meaningful discrepancy metrics is a core element of interpreting cosmological data. The most prominent questions are: Is model A preferred over model B (LCDM vs wCDM in the simplest case)? Is dataset A in tension with dataset B (Euclid vs WFIRST vs LSST)? Before combining datasets, scientists must assess whether the data to be combined are in tension with one another in the context of a given cosmological and systematics model. Discrepancy metrics are also important for a joint simulation effort between LSST, WFIRST, and Euclid – namely in determining whether the simulations are sufficiently precise given the constraining power of the surveys individually and then jointly. This is not trivial since, in principle, the assessment requires an even more precise simulation of the survey(s). Even in the presence of such a fiducial high-precision simulation (e.g., see the Euclid Flagship simulation), the questions arise: what precision do the emulator simulations need, what precision do the covariance/likelihood simulations need, and what precision do the systematics simulations need? **In order to asses whether a simulation (campaign) is sufficiently accurate for the individual surveys LSST, WFIRST, and Euclid and additionally for their joint analysis, it is critical to unify the analysis choices for the survey simulations across the LSST, Euclid, and WFIRST communities.**

5.1 Recommendations for Advanced Statistical Methods

We recommend that the studies mentioned above either be undertaken by competitively selected teams within the funding Agencies pre-existing R&D models or by small, directed "Tiger Teams" that combine expertise in statistical methods and numerical simulations across the surveys. First and foremost, the surveys should share expertise and code, and develop a coordinated testing scheme for software implementations. The simulation resources required to implement some of the solutions on emulators, covariances/likelihoods, and discrepancy metrics should also be shared between the surveys.

6. Common Archival Infrastructure to Share Simulation Products

LSST, WFIRST, and Euclid are all looking at the same sky in a similar time-frame and they all have similar requirements for cosmological simulations. At the simplest level, it is a poor use of resources for the three surveys to produce largely redundant simulation suites individually. In addition, there are only a limited number of people in the world with the expertise to produce extreme-scale cosmological simulations and synthetic sky catalogs and also only a limited number of supercomputing facilities with the resources available to produce extreme-scale simulations or large suites of cosmological simulations. Given these limitations it is challenging for the surveys to realize their cosmological simulation needs individually. In practice, it is the same simulators being approached by the different surveys with slightly varying requests for cosmological simulations and their respective data products. A common archival infrastructure for sharing cosmological simulations will reduce the overall number of simulations that need to be produced, reducing the pressure on both the supercomputing facilities and the simulators.

In order to ensure the scientific success of LSST, WFIRST, and Euclid, it is clear that a common archival infrastructure for simulated data products should be made available. This includes hardware (e.g. storage space, data servers, fast connection and transfer links), as well as a common approach for data curation to make data products easily accessible to the community. It also includes expert support personnel (both for the simulations and the data hosting) who are actively engaged in developing and maintaining the infrastructure, in addition to supporting the users.

There are a number of challenges to developing a common infrastructure. There needs to be a plan for where the simulations are being run with some guarantees that those resources will be available for these efforts. Once the simulations have been completed, an initial analysis may be completed at a different facility, so rapid transfer capabilities of very large datasets need to be in place. Decisions need to be made about what data products are being stored and hosted and how those products are being curated to enable widespread use (i.e. does the data need to be stored sequentially or in a format that enables rapid ingestion by a database?). There are a range of solutions, from simply storing and hosting the flat files for direct download by scientists to analyze on a system that they identify themselves, to more sophisticated database solutions that include access to increasingly powerful analysis hardware at the data center.

6.1 Recommendations for a Common Archival Infrastructure

Every section of this white paper discusses efforts that require either the generation or utilization of cosmological simulations to ensure the scientific success of LSST, WFIRST, Euclid, and related projects. With limited resources and expertise available for each of the surveys, coordination between the surveys on which cosmological simulations and synthetic sky catalogs to produce and a common infrastructure to share the data will clearly contribute to the scientific success of each of the surveys. This approach will also save money in the long-term by reducing the overall number of required simulations and facilitating a common data curation approach that will increase user efficiency in accessing and utilizing the simulations. The work to flesh out the range of solutions for an LSST, WFIRST, Euclid simulation data archival infrastructure requires additional effort. This effort includes scoping and costing the hardware requirements, coordinating with the scientists to identify which data products should be stored and the best methods for curating the data, exploring the methods for accessing the data and options for interfacing with the data, scoping a range of support levels that a data archive center could provide and costing those options, and providing detailed proposals that show what capabilities and scientific return can be expected with specific levels of investment. **We recommend that the funding Agencies invest in a coordinated program to scope and implement a joint simulation archival infrastructure that could be utilized by many surveys and individual scientists that would benefit from the shared numerical simulation data products.**

Acknowledgements

AK was supported by JPL, which is run under contract by the California Institute of Technology for NASA. AK was also supported in part by NASA ROSES grant 12-EUCLID12-0004 and NASA grant 15-WFIRST15-0008. Argonne National Laboratory's work by KH and AH was supported under the U.S. Department of Energy contract DE-AC02-76SF00515.

References

Aniket Agrawal, Ryu Makiya, Chi-Ting Chiang, Donghui Jeong, Shun Saito, and Eiichiro Komatsu. Generating log-normal mock catalog of galaxies in redshift space. Journal of Cosmology and Astroparticle Physics, 2017(10):003, 2017. URL http://stacks.iop.org/1475-7516/2017/i=10/a=003.

S. Bocquet, A. Saro, K. Dolag, and J. J. Mohr. Halo mass function: baryon impact, fitting formulae, and implications for cluster cosmology. Monthly Notices of the Royal Astronomical Society, 456:2361-2373, March 2016. doi: 10.1093/mnras/stv2657.

S. Dodelson and M. D. Schneider. The effect of covariance estimator error on cosmological parameter constraints. Phys. Rev. D, 88(6):063537, September 2013. doi: 10.1103/PhysRevD.88.063537.

T. Eifler, E. Krause, S. Dodelson, A. R. Zentner, A. P. Hearin, and N. Y. Gnedin. Accounting for baryonic effects in cosmic shear tomography: determining a minimal set of nuisance parameters using PCA. MNRAS, 454:2451-2471, December 2015. doi: 10.1093/mnras/stv2000.

O. Friedrich and T. Eifler. Precision matrix expansion - efficient use of numerical simulations in estimating errors on cosmological parameters. Monthly Notices of the Royal Astronomical Society, 473:4150-4163, January 2018. doi: 10.1093/mnras/stx2566.

S. Habib, A. Pope, H. Finkel, N. Frontiere, K. Heitmann, D. Daniel, P. Fasel, V. Morozov, G. Zagaris, T. Peterka, V. Vishwanath, Z. Lukic, S. Sehrish, and W.-k. Liao. HACC: Simulating sky surveys on state-of-the-art supercomputing architectures. New Astronomy, 42:49-65, January 2016. doi: 10.1016/j.newast.2015.06.003.

J. Hartlap, P. Simon, and P. Schneider. Why your model parameter confidences might be too optimistic. Unbiased estimation of the inverse covariance matrix. Astronomy and Astrophysics, 464:399-404, March 2007. doi: 10.1051/0004-6361:20066170.

J. Hartlap, T. Schrabback, P. Simon, and P. Schneider. The non-Gaussianity of the cosmic shear likelihood or how odd is the Chandra Deep Field South? Astronomy and Astrophysics, 504:689-703, September 2009. doi: 10.1051/0004-6361/200911697.

K. Heitmann, H. Finkel, A. Pope, V. Morozov, N. Frontiere, S. Habib, E. Rangel, T. Uram, D. Korytov, H. Child, S. Flender, J. Insley, S. Rizzi. The Outer Rim Simulation: A Path to Many-Core Supercomputers. ArXiV Preprints: 1904.11970. April 2019

B. Joachimi. Non-linear shrinkage estimation of large-scale structure covariance. Monthly Notices of the Royal Astronomical Society, 466:L83-L87, March 2017. doi: 10.1093/mnrasl/slw240.

B. Joachimi, M. Cacciato, T. D. Kitching, A. Leonard, R. Mandelbaum, B. M. Sch�afer, C. Sif_on, H. Hoekstra, A. Kiessling, D. Kirk, and A. Rassat. Galaxy Alignments: An Overview. SSRv, 193:1-65, November 2015. doi: 10.1007/s11214-015-0177-4.

E. Krause and T. Eifler. cosmolike - cosmological likelihood analyses for photometric galaxy surveys. Monthly Notices of the Royal Astronomical Society, 470:2100-2112, September 2017. doi: 10.1093/mnras/stx1261.

E. Lawrence, K. Heitmann, J. Kwan, A. Upadhye, D. Bingham, S. Habib, D. Higdon, A. Pope, H. Finkel, and N. Frontiere. The Mira-Titan Universe. II. Matter Power Spectrum Emulation. The Astrophysical Journal, 847:50, September 2017. doi: 10.3847/1538-4357/aa86a9.

D. Potter, J. Stadel, and R. Teyssier. PKDGRAV3: beyond trillion particle cosmological simulations for the next era of galaxy surveys. Computational Astrophysics and Cosmology, 4:2, May 2017.

S. Ravanbakhsh, J. Oliva, S. Fromenteau, L. C. Price, S. Ho, J. Schneider, and B. Poczos. Estimating Cosmological Parameters from the Dark Matter Distribution. ArXiv e-prints, November 2017.

E. Sellentin, C. Heymans, and J. Harnois-D_eraps. The skewed weak lensing likelihood: why biases arise, despite data and theory being sound. Monthly Notices of the Royal Astronomical Society, 477:4879-4895, July 2018. doi: 10.1093/mnras/sty988.

E. Semboloni, H. Hoekstra, and J. Schaye. E_ect of baryonic feedback on two- and three-point shear statistics: prospects for detection and improved modelling. MNRAS, 434:148-162, September 2013. doi:10.1093/mnras/stt1013.

Rachel S. Somerville and Romeel Davé. Physical Models of Galaxy Formation in a Cosmological Framework. Annual Review of Astronomy and Astrophysics, 53:51-113, August 2015. ISSN 0066-4146. doi:10.1146/annurev-astro-082812-140951.

R. Takahashi, M. Sato, T. Nishimichi, A. Taruya, and M. Oguri. Revising the Halofit Model for the Nonlinear Matter Power Spectrum. Astrophys. J., 761:152, December 2012. doi: 10.1088/0004-637X/761/2/152.

A. Taylor, B. Joachimi, and T. Kitching. Putting the precision in precision cosmology: How accurate should your data covariance matrix be? Monthly Notices of the Royal Astronomical Society, 432: 1928-1946, July 2013. doi: 10.1093/mnras/stt270.

A. Taylor and B. Joachimi. Estimating cosmological parameter covariance. Monthly Notices of the Royal Astronomical Society, 442:2728-2738, August 2014. doi: 10.1093/mnras/stu996.

M. A. Troxel and M. Ishak. The intrinsic alignment of galaxies and its impact on weak gravitational lensing in an era of precision cosmology. PhR, 558:1-59, February 2015. doi: 10.1016/j.physrep.2014.11.001.

R. H. Wechsler and J. L. Tinker. The Connection between Galaxies and their Dark Matter Halos. ArXiv:1804.03097, April 2018.

P. Wilking and P. Schneider. A quasi-Gaussian approximation for the probability distribution of correlation functions. Astronomy and Astrophysics, 556:A70, August 2013. doi: 10.1051/0004-6361/201321718.

A. R. Zentner, E. Semboloni, S. Dodelson, T. Eifler, E. Krause, and A. P. Hearin. Accounting for baryons in cosmological constraints from cosmic shear. PhysRevD, 87(4):043509, February 2013. doi: 10.1103/Phys-RevD.87.043509.