Uncertainty Analysis for Heavy Simulations of Galaxy formation.

Ian Vernon*

Department of Mathematical Sciences, Durham University,
Science Laboratories, Durham, DH1 3LE, UK

September 6, 2017

1 Simulating Galaxies and Universes

"We have some good news", my collaborator announced as I wandered into his office one morning in early 2017. "We have been granted 60 million CPU hours to run possibly the largest hydrodynamic simulation of the Universe ever!". "Well, I guess that is good news" I said, uncertainly, "What's the bad news?". "Er, I didn't mention there was any bad,... well, OK, we kind of want you to choose the location in parameter space to run the model at", he said. This wasn't wholly unexpected. "How long in real time will it take to run 60 million CPU hours on the given facility?" I enquired, curiously. "Real time? Oh, about one and a half years...". Several unprintable expletives then followed.

The model in question is the EAGLE simulation (Schaye et al., 2015), which is indeed one of the most complex models of galaxy formation ever run. My collaborator is Prof Richard Bower, a member of the Institute of Computational Cosmology here at Durham University, and one of the core members of the EAGLE group and of the VIRGO consortium (see http://www.virgo.dur.ac.uk) that created and ran EAGLE. The facility in question is run by PRACE, the Partnership for Advanced Computing in Europe (see http://www.prace-ri.eu). And I am a Bayesian statistician, with a background in theoretical physics, who specialises in the Bayesian uncertainty analysis of computer models of complex physical systems - an area that overlaps with, and some more contentious than myself would say has a far wider and deeper scope than, the recently fashionable area commonly termed "Uncertainty Quantification".

1.1 The EAGLE Simulation

EAGLE stands for the Evolution and Assembly of GaLaxies and their Environments which, aside from implying that someone really wanted an acronym that spelt EAGLE, means that its purpose is to understand how large numbers of galaxies form, collide and evolve. The simulation models a cosmological volume of $(100 \text{ Megaparsecs})^3$, which is about $(326 \text{ million light years})^3$, a volume large enough to contain approximately 10,000 galaxies of the size of the Milky Way or larger. The simulation starts prior to the formation of any stars or galaxies, when the Universe was still very uniform, and uses nearly 7 billion particles in combination with well known fundamental physical laws for example of gravity and of hydrodynamics. It models the effects of dark matter (which

*i.r.vernon@durham.ac.uk. Many thanks to Richard G. Bower, Aaron Ludlow, Alejandro B. Llambay (ICC, Durham University), the EAGLE group, the VIRGO consortium and PRACE.

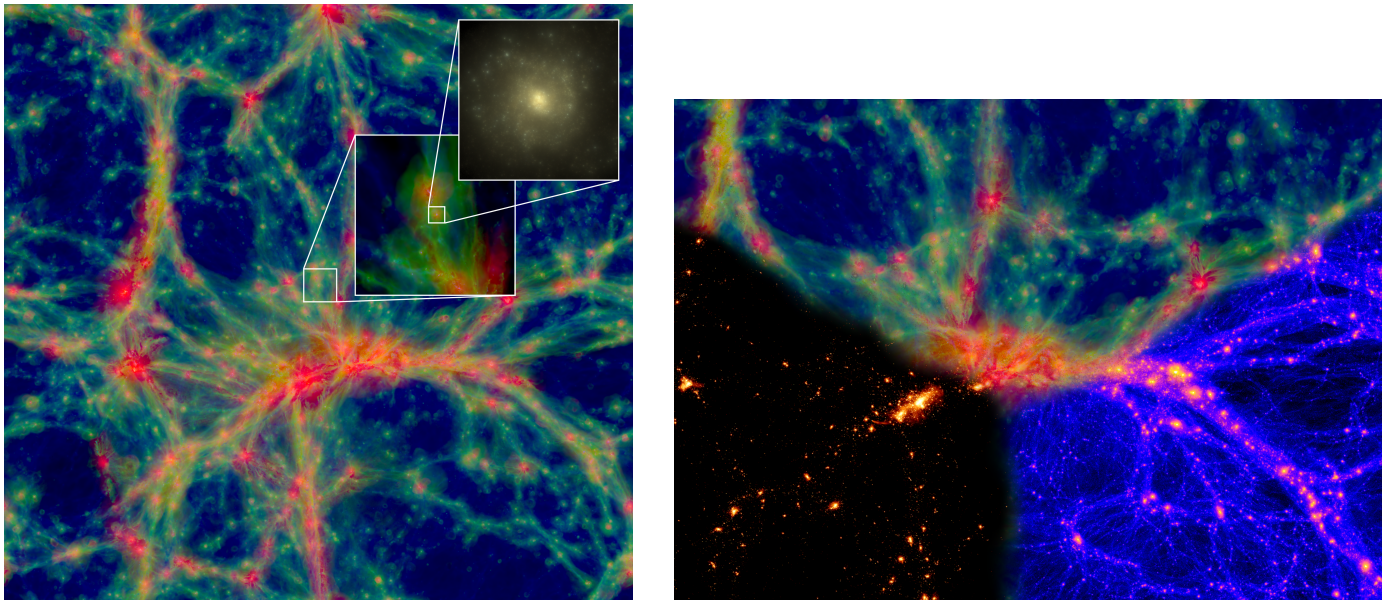

Figure 1: Left panel: a slice through the EAGLE simulation volume, showing the intergalactic gas colour coded from blue to red with increasing temperature. The inset zooms into a galaxy similar to the Milky Way, showing first its gas and then its stellar disc, which looks remarkably similar to observed spiral galaxies. Right panel: another slice through the EAGLE simulation showing the hot gas content (top), the dark matter density (bottom right) and what the simulation would look like in the visible spectrum (bottom left). Thanks to the VIRGO consortium.

allows large galaxy sized structures to grow) and baryonic matter (that forms stars) along with the cosmological constant (that causes cosmic acceleration). The end results of the simulation can then be compared to various detailed but complex observed data sets that measure a variety of galaxy features: common ones include the stellar mass function (the distribution of galaxies over stellar mass), and the distributions of galaxy sizes. Examples of the output from EAGLE can be seen in figure 1. See http://icc.dur.ac.uk/Eagle/ for more details, including some rather beautiful movies.

Some example scientific questions that EAGLE seeks to answer (there are many more) are:

- How do galaxies stop growing? Is it because of the activity of the central black hole? Is it because they collide and merge? Is it because they are in a crowded environment?
- How typical is our own Milky-Way? Are we in a normal galaxy in a normal part of the Universe or is there something special about where we live?
- How do the different gas flows affect the formation of galaxies?
- How does the presence of gas affect the observations of halo masses, lensing or dark matter?

A single run of the 100 Megaparsec (Mpc) model was performed in 2015, which took 1.5 months using 4064 processors (a substantial proportion of the VIRGO consortium's computational resources at the time). This showed that EAGLE is of sufficient accuracy to attempt to answer many of the above questions, and led to a large number of publications, the first of which (Schaye et al., 2015) has obtained over 580 citations, and was one of the most cited papers on astro-ph that year. Now,

the plan is to run even larger volumes: perhaps up to 15 times larger, as described in the slightly melodramatic opening paragraph above.

1.2 A Major Challenge

So, what in fact is the problem? Well, in a word: uncertainty. Now that the huge amount of work developing and efficiently coding up the current EAGLE version has been completed, we can perform a single model evaluation, using admittedly substantial computational resources and a lot of patience. This would be sufficient, were there only one possible way to run EAGLE. However, EAGLE features several uncertainties, many in the form of parameters related to hard to model 'sub-grid' processes. In short, galaxy formation critically depends on processes spanning wildly different scales: for example black holes at the centre of galaxies draw in gas on scales of 0.01 parsecs, but the energy produced by this process affects the whole galaxy and possibly its host halo up to a scale of 1 Megaparsec, effectively spanning 8 orders of magnitude in spatial resolution.

EAGLE itself, commendably one should say, spans over 5 orders of magnitude in resolution. To give some feel for this scale (although such comparisons should be treated with extreme caution, as there are many complexities here), if one managed a similar level of spatial resolution attempting to model the Earth's atmosphere, for use for example in a climate model, each cubic grid cell would be less than 26 metres across. However, EAGLE's impressive resolution is still nowhere near high enough to accurately represent either the effect of central black holes nor various other important small scale phenomena that affect galaxy formation, such as the impact of supernovas (massive stars that explode and drive gas out of the galaxy). Hence these processes have to be modelled via sub-grid scale models, that are parameterised using a modest number of physical input parameters representing uncertain aspects of the processes in question. EAGLE also possesses additional cosmological parameters, but these are little more understood and usually set to the values as measured to reasonably high accuracy by the Plank satellite. 7 sub-grid parameters of interest have been identified as strongly influential and hence form the core of the current study. The remaining parameters are thought to be somewhat sub-dominant, but their effects will be taken into account, in a less detailed form, in our analysis below.

Hence to really understand the scientific ramifications of EAGLE, one inevitably has to explore its uncertain behaviour over this 7-dimensional parameter space. As each step in this parameter space takes 1.5 months to complete, using 4064 processors, one can see the problem: standard search techniques are utterly infeasible. Just to reiterate this point: a 7-dimensional hypercube has 128 corners, so just to visit these would take the current version of EAGLE over 64000 years (on a single processor). But more detail would inevitably be required, for example a 7-dimensional grid of length 10, and therefore of size 10^7 may be sufficient, but this would take over 5 billion years to evaluate, which is somewhat ironically, well over a third of the current age of the Universe.

Critically, we must go even further: as EAGLE produces many different outputs that can be compared with a range of observed data sets, our real goal is to identify *all* the choices of the input parameters that will lead to acceptable matches between the model output and observed data (or to find that no such choices exist), hence requiring a detailed parameter search. Note that only finding a single acceptable match may be scientifically highly misleading. This is sometimes referred to as an inverse problem, a Bayesian calibration problem, or a history matching problem (we prefer the latter, for various somewhat subtle reasons: for details see Vernon et al. (2010)). Finally, we may want to use our understanding of the input parameter space, to choose the input parameters for a single future, even larger, EAGLE run, or perhaps to design a limited set of slightly smaller

runs chosen to be at highly informative locations across the parameter space. To address the above problems, we really require the use of Bayesian statistics.

2 Bayesian analysis of computer models of complex physical systems

The reason the above general problem structure, as faced by the EAGLE collaboration, is of interest to Bayesian statisticians is not just because of the fascinating scientific questions EAGLE hopes to answer, but because it has many of the attributes of a type of problem that is currently occurring in a wide variety of scientific disciplines. Due to the increase in mathematical modelling and corresponding computing power, many scientific areas are developing ever more complex, high-dimensional and computationally expensive models of physical systems. Helpfully, an area of (Bayesian) statistics has developed over the last 25 years, designed specifically to combat the challenges posed by this kind of problem, the general form of which we now describe.

A model is created for a particular real world system of interest, that describes how a vector of various system properties x affects a vector of system behaviour, given by the model as $f(x)$. So for example, for all of EAGLE's complexity, it is just a function $f(x)$ that maps a 7-dimensional x to a high-dimensional vector of galaxy property outputs f. The model or simulator is, however, imperfect, and the real system properties (suitably defined, an interesting question all by itself) are given by the vector y. We may wish to explicitly model the gap between reality y and the model $f(x)$ evaluated at its best input x^* for example via $y = f(x^*) + \epsilon$, where ϵ is now a random vector, with a possibly complex joint structure, representing the unknown structural deficiencies of the model. We can of course measure a subset of the system properties, but with error: these measurements are given by a vector of data z_p, and correspond to past system properties y_p with y partitioned as $y = (y_p, y_f)$, where y_f represents possible future properties of interest, that we may want to predict. Again, we may make the gap between measurements and real system explicit for example via say $y = z + e$, where e is a random vector representing measurement error.

We wish to answer various scientific questions, while accounting for all the uncertainties that exist in the above setup. For example we may wish to

1. Explore the model's behaviour $f(x)$ over a defined input space $x \in \mathcal{X}$.
2. Learn about acceptable values of x (or perform full Bayesian inference on x) by comparing the model $f(x)$ to observed data z.
3. Explore the accuracy of the model for reproducing various outputs, and hence assess its adequacy for the task at hand.
4. Use the model combined with past observations z_p to make predictions of future outputs y_f.
5. Use the model along with the assessed uncertainties in some decision theory calculation, for example, to help aid policy makers.

However, the model or simulator $f(x)$ is usually extremely computationally expensive to evaluate, relative to the dimension of x, preventing the evaluation of any of the above calculations. Hence we have some major problems which can be grouped roughly as follows:

- **The speed problem**: the model is far too slow to be used to explore its input parameter space in naive ways. For example, we cannot plug it into standard optimisers or more sophisticated algorithms such as MCMC, that usually require vast numbers of model evaluations.
- **The general uncertainty problem**: the answers to whatever scientific questions we wish to pose will critically depend upon the assessment of all the various uncertainties in the problem. In particular the multivariate nature of the structural discrepancy ϵ, the observational errors e, and input parameter uncertainty x may have a major impact.

2.1 Solving the Speed problem: Bayesian Gaussian Process Emulation

Firstly, we must acknowledge the underlying problem: except at a small number of input locations where we actually decide to run the model, we will always be uncertain as to the true value of the EAGLE function $f(x)$. In the Bayesian setting, we can incorporate this uncertainty naturally, by simply treating $f(x)$ at unevaluated x as another random vector. Secondly, we then ask what do we know about this uncertain function $f(x)$? For example, many physical functions are in some sense smooth, in that although small changes to the input parameters could in principle greatly effect the outputs, this may (in the domain expert's view) be deemed unlikely based on consideration of the fundamental equations, and the physical nature of the system under investigation. Such considerations facilitate the construction of Bayesian emulators, which are specifically employed to combat the speed problem. A Bayesian emulator is a fast statistical function built to mimic the behaviour of the EAGLE function $f(x)$ over the input space $\mathcal{X}$. The emulator provides both an expectation as to the value of $f(x)$ at an as yet unevaluated x, and critically an x dependent uncertainty statement as to the emulator's accuracy at this point, which can be naturally incorporated into a Bayesian analysis. Most importantly, the emulators are very fast to evaluate and are usually multiple orders of magnitude faster that the model itself. In this application, they are between $10^9 - 10^{12}$ times faster than EAGLE (depending on which version of EAGLE we compare to), the kind of speed increase that tends to turn heads in most scientific communities.

A popular statistical model for the Bayesian emulator for $f(x)$, which has individual outputs $f_i(x)$, $i = 1 \dots q$, is structured as follows (see for example Vernon et al. (2010) for details):

$$f_i(x) = \sum_j \beta_{ij} g_{ij}(x_{A_i}) + u_i(x_{A_i}) + w_i(x) \tag{1}$$

where the active variables x_{A_i} are a subset of the inputs x that are most influential for output $f_i(x)$. The first term on the right hand side of the emulator equation (1) is a regression term, where g_{ij} are known deterministic functions of x_{A_i}, a common choice being low order polynomials, and β_{ij} are unknown scalar regression coefficients. The second term, $u_i(x_{A_i})$ is a Gaussian process[1] over x_{A_i}, which means that if we choose a finite set of inputs $\{x_{A_i}^{(1)}, \dots, x_{A_i}^{(s)}\}$, the uncertain outputs $u_i(x_{A_i}^{(1)}), \dots, u_i(x_{A_i}^{(s)})$ will have a multivariate normal distribution with a covariance matrix constructed from an appropriately chosen covariance function, a popular form being:

$$\mathrm{Cov}(u_i(x_{A_i}), u_i(x'_{A_i})) = \sigma_{u_i}^2 \exp\left\{-\|x_{A_i} - x'_{A_i}\|^2 / \theta_i^2\right\} \tag{2}$$

[1]It is worth noting that Bayesian style Gaussian processes are now heavily used in the machine learning community, giving weight to the amusing, but perhaps unfair quip that "machine learning is just doing Bayesian statistics on a Mac".

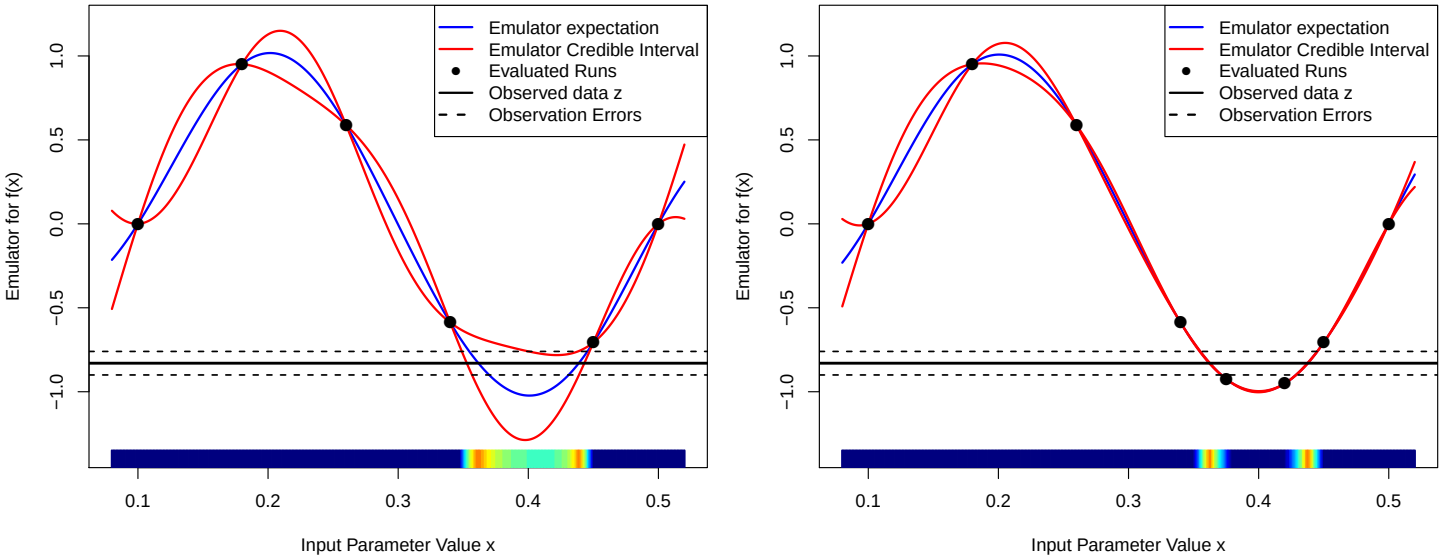

Figure 2: An example emulator for a 1-dimensional toy model where $f(x) = \sin(2\pi(x-0.1)/0.4)$, for the 1st wave, using just 6 runs (left panel), and for the 2nd wave, using 2 additional runs (right panel). The emulator's expectation $\mathrm{E}_D[f(x)]$ and credible intervals $\mathrm{E}_D[f(x)] \pm 3\sqrt{\mathrm{Var}_{D_i}(f_i(x))}$ are given by the blue and red lines respectively, with the observed data z that we wish to match to as the black horizontal line (with errors). The implausibility $I(x)$ is represented by the coloured bar along the x-axis, with dark blue implying $I(x) > 3$, light blue $2.5 < I(x) < 3$ and yellow ($I(x) < 1$). These emulators are for deterministic models, but stochastic equivalents of course also exist.

where $\sigma^2_{u_i}$ and θ_i are the variance and correlation length of $u_i(x_{A_i})$ which must be specified a priori. The third term $w_i(x)$ is a nugget, a white noise process with variance $\sigma^2_{v_i}$, uncorrelated with β_{ij}, $u_i(x_{A_i})$ and itself, that represents the effects of the remaining inactive input variables.

Given a set of n carefully chosen runs $D_i = (f_i(x^{(1)}), f_i(x^{(2)}), \ldots, f_i(x^{(n)}))$, we can update our prior beliefs about $f(x)$ at unevaluated x by D_i using either Bayes theorem (which requires full probability distributions) or the computationally efficient Bayes linear update (which only requires expectations and variances). The latter provides the adjusted expectation and variance of $f(x)$, denoted $\mathrm{E}_{D_i}(f_i(x))$ and $\mathrm{Var}_{D_i}(f_i(x))$. Figure 2 shows an example of a 1d emulator of a deterministic toy model (a simple sine function), with emulator's adjusted expectation $\mathrm{E}_{D_i}(f_i(x))$ and credible interval $\mathrm{E}_{D_i}(f_i(x)) \pm 3\sqrt{\mathrm{Var}_{D_i}(f_i(x))}$ as given by the blue and red lines respectively.

The speed of the emulators, allows us to comprehensively explore the input parameter space $\mathcal{X}$ and identify regions of $\mathcal{X}$ that may lead to acceptable matches to the observed data z. We do this by using implausibility measures, the simplest form of which is, for output i

$$I_i^2(x) \quad = \quad \frac{(\mathrm{E}_{D_i}(f_i(x)) - z_i)^2}{\mathrm{Var}_{D_i}(f_i(x)) + \mathrm{Var}(\epsilon_i) + \mathrm{Var}(e_i)} \tag{3}$$

Usually we perform the exploration in iterations or 'waves', using the emulators and implausibility measures to rule out parts of the current space $\mathcal{X}_k$ that are obviously poor (which have high $I_i(x)$ for a subset of the outputs), before performing further runs of the model in the not yet ruled out region $\mathcal{X}_{k+1}$ say, and reconstructing new, more accurate emulators defined only over $\mathcal{X}_{k+1}$. This divide and conquer approach is very powerful. The x-axis of figure 2 is coloured by implausibility, showing the obviously bad parts of the input space with high $I(x) > 3$ in dark blue, that correctly suggest $f(x)$ will be far away from the data z, given with error as the horizontal black lines.

2.2 Taming exceedingly slow simulators: Multilevel Emulation

Even given the above emulation technology, EAGLE at its current size of 100 Mpc, is still too slow to perform enough runs to construct even a moderately accurate emulator over 7-dimensional space. Things seem a little hopeless until we ask if there are faster, approximate versions of EAGLE available, that we can use for a process known as multilevel emulation. Helpfully there are, EAGLE can indeed be run over smaller volumes of the Universe, and has been set up to run on cubes of size 12.5 Mpc, 25 Mpc, 50 Mpc and the full 100 Mpc, which we will refer to as levels 1 to 4 respectively. Each level is thought to be approximately 8 times faster than the next, although levels 1 and 2 gain additional speed as they don't have to simulate very large galaxies.

There are, however, two important differences between the levels: a) levels 1 and 2 only model relatively small numbers of galaxies, and so we encounter noise in many of the outputs due to finite galaxy counts, b) the lower levels are *physically* different from the level 4 simulation, in that due to periodic boundary conditions the largest galaxies simply cannot form inside a 12.5 Mpc or even a 25 Mpc box, leading to possibly substantial systematic differences between runs at different levels for the same input x. Multilevel emulation can usually handle such issues. All we need is for the lower levels to be informative for the higher levels (so biases, systematic differences etc. are fine).

We begin by building an emulator $f^{(1)}(x)$ for level 1, summarised by the uncertain quantities $\{\beta_{ij}^{(1)}, u_i^{(1)}(x_{A_i}), w_i^{(1)}(x)\}$ from equation (1), based on a carefully chosen set of 60 runs. We then construct a prior emulator for level 2 by specifying a representation for $\{\beta_{ij}^{(2)}, u_i^{(2)}(x_{A_i}), w_i^{(2)}(x)\}$ based on their level 1 counterparts, say by modestly inflating the level 1 uncertainties and by including any additional physical structure or suspected systematic differences we are aware of. We now require far fewer level 2 runs (here we used 20) to update this relatively well informed prior level 2 emulator. We will then repeat the process for levels 3 and 4, but now focus on the parts of $\mathcal{X}$ that may yield good matches to observed data (so that we do not squander runs in uninteresting parts of the parameter space). We are currently in the process of designing the set of level 3 runs.

Figure 3 shows the results of the level 2 emulator and corresponding (maximised) implausibility measure based on the important stellar mass function outputs, over the full 7-dimensional space (shown as all possible 2-dimensional projections). This used 400000 emulator evaluations, completing in minutes. The dark blue areas will be ruled out as implausible. The light blue/green/red areas will need a second wave of runs to investigate further, but look likely to produce moderate to good fits to the observed data set. The pink dot is the location of the single 100 Mpc run performed in 2015. It can be seen that it is in a good part of the input space as judged from several 2-dimensional projections, however its location could be improved, especially in terms of the BlackHoleViscousAlpha input parameter (however, we should stress that the 100 Mpc run was chosen to match several additional data sets that could have therefore caused this disparity).

This project is ongoing, but once we have performed a small number of level 3 and 4 runs, we will be in a position to answer the original question and propose a suitable parameter location for a single massive 'level 5' run, or to suggest a set of locations for slightly smaller runs, designed to resolve some of the key scientific questions outlined above. Then we will just have to wait.

2.3 The General Uncertainty Problem: application to other disciplines.

Addressing the general uncertainty problem, for example, by assessing the properties of the structural discrepancy ϵ and observed errors e, that are combined with the emulator uncertainty in equation (3), is of course context dependent. However, we have successfully applied this style of

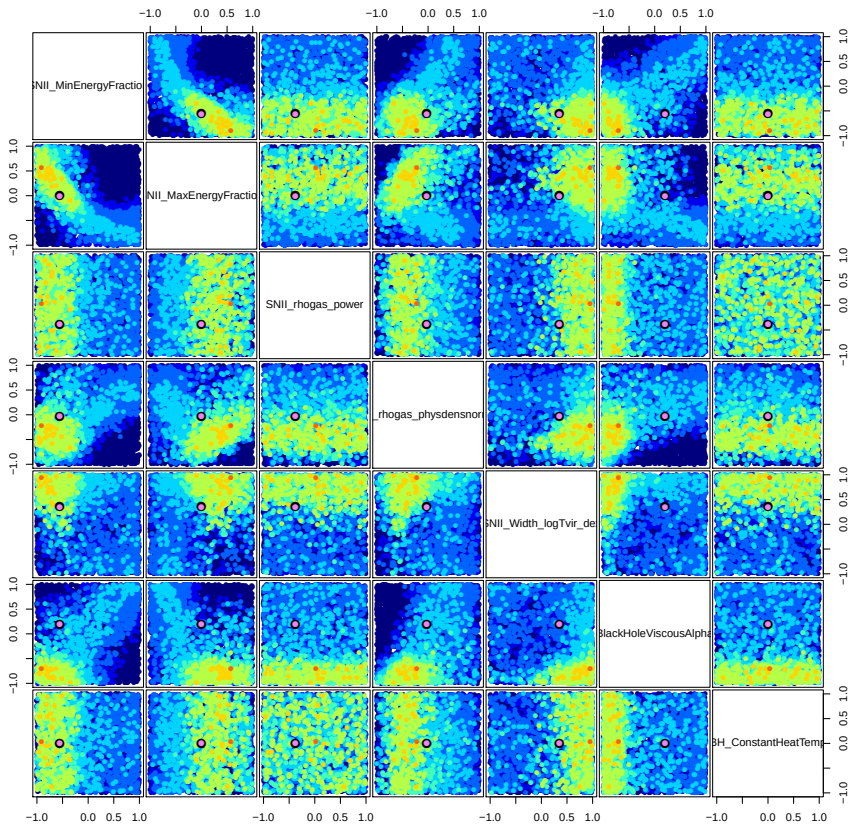

Figure 3: The implausibility of the 7-dimensional input space of the EAGLE simulation, shown as all possible two-dimensional projections (the 7 input parameters are named down the main diagonal: the first 5 describe supernova, and the last two central black holes). The colour scheme is consistent with figure 2 so that dark blue shows regions we would discard, light blue gives borderline regions ($2.5 < I(x) < 3$) that we would wish to explore further in the next wave, while the green/yellow regions suggest that the emulators currently think that good matches between the 25 Mpc level 2 version and the stellar mass function data could be found (but this may change with more runs). Note that the low implausibility points are plotted last, allow one effectively to see through the less interesting parts of the space. The pink dot is the location of the previous 100Mpc EAGLE run.

Bayesian uncertainty analysis across multiple scientific disciplines, including to semi-analytic galaxy formation simulations (Vernon et al. (2010)) in which we assessed the contribution of nine separate sources of uncertainty to a 17 input dimension model, a paper that was subsequently awarded the top prize in Bayesian statistics: the Mitchel Prize for the best applied Bayesian article worldwide by JASA/ISBA (see also Rodrigues et al. (2017)), and we are following this form of analysis for EAGLE. Other areas of application include for example epidemiology (Andrianakis et al. (2015, 2016)), involving models of HIV possessing up to 96 inputs; environmental models (Goldstein et al. (2013)), describing tools for detailed structural discrepancy assessments; systems biology models (Vernon et al. (2017)), and climate models (Williamson et al. (2013)).

References

Andrianakis, I., McCreesh, N., Vernon, I., McKinley, T., Oakley, J., Nsubuga, R., Goldstein, M., and White, R. (2016), "History matching of a high dimensional individual based HIV transmission

model," *Journal of Uncertainty Quantification*, to appear.

Andrianakis, I., Vernon, I., McCreesh, N., McKinley, T., Oakley, J., Nsubuga, R., Goldstein, M., and White, R. (2015), "Bayesian History Matching of Complex Infectious Disease Models Using Emulation: A Tutorial and a Case Study on HIV in Uganda." *PLoS Comput Biol.*, 11, e1003968.

Goldstein, M., Seheult, A., and Vernon, I. (2013), *Environmental Modelling: Finding Simplicity in Complexity*, Chichester, UK: John Wiley & Sons, Ltd, chap. Assessing Model Adequacy, 2nd ed.

Rodrigues, L. F. S., Vernon, I., and Bower, R. G. (2017), "Constraints to galaxy formation models using the galaxy stellar mass function." *MNRAS*, 466, 2418–2435.

Schaye, J., Crain, R. A., Bower, R. G., and et. al. (2015), "The EAGLE project: simulating the evolution and assembly of galaxies and their environments," *Monthly Notices of the Royal Astronomical Society*, 446, 521–554.

Vernon, I., Goldstein, M., and Bower, R. G. (2010), "Galaxy Formation: a Bayesian Uncertainty Analysis," *Bayesian Analysis*, 5, 619–670.

Vernon, I., Liu, J., Goldstein, M., Rowe, J., Topping, J., and Lindsey, K. (2017), "Bayesian uncertainty analysis for complex systems biology models: emulation, global parameter searches and evaluation of gene functions." *BMC Systems Biology*, in submission.

Williamson, D., Goldstein, M., Allison, L., Blaker, A., Challenor, P., Jackson, L., and Yamazaki, K. (2013), "History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble," *Climate Dynamics*, 41, 1703–1729.