Bayesian Analysis (2010) **5**, Number 4, pp. 619–670

Galaxy Formation: a Bayesian Uncertainty Analysis

Ian Vernon*, Michael Goldstein† and Richard G. Bower‡

Abstract. In many scientific disciplines complex computer models are used to understand the behaviour of large scale physical systems. An uncertainty analysis of such a computer model known as Galform is presented. Galform models the creation and evolution of approximately one million galaxies from the beginning of the Universe until the current day, and is regarded as a state-of-the-art model within the cosmology community. It requires the specification of many input parameters in order to run the simulation, takes significant time to run, and provides various outputs that can be compared with real world data. A Bayes Linear approach is presented in order to identify the subset of the input space that could give rise to acceptable matches between model output and measured data. This approach takes account of the major sources of uncertainty in a consistent and unified manner, including input parameter uncertainty, function uncertainty, observational error, forcing function uncertainty and structural uncertainty. The approach is known as History Matching, and involves the use of an iterative succession of emulators (stochastic belief specifications detailing beliefs about the Galform function), which are used to cut down the input parameter space. The analysis was successful in producing a large collection of model evaluations that exhibit good fits to the observed data.

Keywords: computer models, uncertainty analysis, model discrepancy, history matching, Bayes linear analysis, galaxy formation, galform

1 Introduction

Current theories of cosmology suggest that the Universe began in a hot, dense state approximately 13 billion years ago, and that it has been expanding rapidly ever since. However, observations of galaxies imply that there must exist far more matter in the Universe than the visible matter that makes up stars, planets and us. This is referred to as 'Dark Matter' and understanding its nature and role in the evolution of galaxies is one of the most important problems in modern cosmology. The Galform group, based at the Institute of Computational Cosmology, Durham University, is the world leading group in the study of Galaxy Formation in the presence of Dark Matter (see Bower et al. (2006) and references therein). Over the last 13 years, they have developed a detailed computer model, known as Galform, which simulates the creation and evolution of

*Department of Mathematical Sciences, Durham University, Science Laboratories, Durham, UK, mailto:i.r.vernon@durham.ac.uk

†Department of Mathematical Sciences, Durham University, Science Laboratories, Durham, UK, mailto:michael.goldstein@durham.ac.uk

‡Department of Physics, Durham University, Science Laboratories, Durham, UK, mailto:r.g.bower@durham.ac.uk

 DOI:10.1214/10-BA524

approximately one million galaxies from the beginning of the Universe until the present day. The simulation produces various physical features of each of the galaxies which can be compared to observed galaxy survey data.

The Galform model requires many input parameters to be specified in order to run the simulation. It is therefore necessary to explore the input parameter space and find the set of all input configurations that give rise to acceptable matches between model output and observed data. As the model run time is significant, this is a challenging task. Further, even to assess what constitutes an acceptable match, we must consider all of the uncertainties that are involved in the comparison between model and reality, including input parameter uncertainty, function uncertainty, observational error, forcing function uncertainty and structural uncertainty. Such a detailed level of uncertainty quantification has never been attempted for such a cosmological model.

This case study describes a collaboration between members of the Statistics group and the Galform group, at Durham, to carry out such an uncertainty analysis for Galform. Our aim is to identify all choices of input parameters that generate consistent physical models in the sense that they would yield sufficiently good matches to certain important features of observational data, when we have taken into account all relevant sources of uncertainty. In particular, it is of fundamental interest to know whether this set of acceptable inputs is non-empty.

In order to treat all uncertainties in a consistent and unified manner, we use general techniques related to the Bayesian treatment of uncertainty for computer models for large scale physical systems. In addition to the uncertainty associated with the Galform function itself, we elicit all of the other sources of uncertainty which must be addressed in order to make meaningful comparisons between Galform output and observations.

Our approach is based on the construction of an emulator for Galform, this being a stochastic function that represents our beliefs about the behaviour of the simulator. We use the emulator and the model uncertainties to define implausibility measures over the input parameter space for Galform, based on a Bayes Linear analysis. We exclude regions of input space by imposing cutoffs on our implausibility measures. We proceed iteratively, making function evaluations over the full range of the input space, emulating Galform over this space, using implausibility measures to remove a part of the space, making a further collection of evaluations of Galform in the reduced space, re-emulating within the reduced space, re-evaluating our implausibility measures over this subspace and therefore removing a further portion of the space and continuing in this fashion. We performed this cycle four times, in each case making a substantial further reduction to the allowable input space. We then made a final set of runs to check that we did have many acceptable matches between Galform output and observations over a range of input parameter choices within the final reduced space.

This is a significant contribution toward understanding the Galform model, as previously no knowledge of the shape and extent of the acceptable region of input space existed. Further, the previous best matches to the primary data set of interest were not compatible with other secondary, but important, observational data sets. Our analysis demonstrates that, by making realistic assessments of structural uncertainty, we are

indeed able to simultaneously match data sets that were previously thought to be incompatible, contradicting authors who suggested that the Universe is "anti-hierarchical" (see section 2), and that such a match is impossible. Thus this work should be viewed as consistent with the hypothesis that galaxies formed in the presence of large amounts of Dark Matter, and in particular via hierarchical merging.

This collaboration began in an informal fashion. Members of the statistics group were interested in applying various techniques that they had developed for the analysis of large scale computer models. The Galform group offered the use of their model and some of their computing facilities. Over time, it became clear that such an analysis was a useful tool for understanding various scientific issues related to the model, and merited a serious collaborative effort to pursue these questions. This account is a description of the results of the collaboration, described more or less as it has evolved.

The Case Study paper is structured as follows. In section 2 we discuss the physical motivation for the study of galaxy evolution and give a general description of the Galform model. Section 3 describes the Computer Model methodology that we will employ, and highlights all the relevant uncertainties that must be considered. In section 4 we describe the construction of the Wave 1 emulator. In section 5 we assess all remaining uncertainties relevant to the analysis and in section 6 we perform the first iteration of the History Matching process. Section 7 deals with the remaining iterations, and the results are reported in section 8. We conclude with discussions regarding physical insight gained in section 9.

2 A universe full of galaxies

The night sky is full of stars. Yet the stars that are visible to the human eye are only an unimaginably tiny fraction of the stars in the universe as a whole. Equipped with telescopes, astronomers have discovered that at great distances beyond our own galaxy lie millions of millions of other galaxies, each with their own populations of stars.

Because of the finite speed of light, such distant galaxies are seen when the universe was much younger. Astronomers can use this time delay to observe the build up and formation of galaxies. The most distant galaxies identified to date are seen only 10^9 years after the big bang, when the universe was less than $1/10^{\text{th}}$ of its current age. Such observations reveal some puzzling results: they suggest that a large proportion of the most massive galaxies are present quite early in the history of the universe. This is in seeming contradiction to the theoretical predictions of the popular Cold Dark Matter model, which suggests galaxies form through a process of 'hierarchical aggregation': small galaxies form early in the history of the universe, building larger and larger galaxies through gravitational collapse and collision. Explaining this apparent contradiction is a key motivation in the development of the Galform model described below. See Appendix A for more details, and for an introduction to galaxy formation.

2.1 Modelling Galaxy Formation: The GALFORM Model

There are essentially two approaches to modelling the formation of galaxies (see appendices A and B). The first is the "numerical simulation" method, a simple and direct approach which uses fundamental physical equations only. This suffers greatly from resolution issues and cannot model many critical features of galaxy formation, for example the winds produced by stars at their deaths, or the formation and effect of black holes.

"Semi-analytic modelling" represents the alternative approach. Rather than tackling the whole problem in a single numerical integration, it is broken down into its separate components or modules. For example, one component of the model is the growth and merging of *dark matter haloes* (each visible galaxy is thought to be contained within a massive halo of dark matter, which dominates the galaxy's motion). This can be computed by running a numerical calculation known as the Millennium simulation that only includes mass and the force of gravity (see section 2.2). As the various processes involved in galaxy formation are not precisely understood, the modules include a number of uncertain, adjustable input parameters x, the exploration of which is the main goal of this case study. Because of the intrinsic complexity of the galaxy formation problem, "semi-analytic models" currently offer the best avenue for progress.

Galform is a world-leading example of such a semi-analytic galaxy formation model (see Bower et al. (2006)). The principle modules within the Galform code track: [1] the gravitational collapse and build-up of dark matter haloes; [2] the cooling and accretion of gas; the formation of stars, stellar evolution and "feedback" from supernova explosions; [3] galaxy mergers and instabilities in stellar disks; [4] the formation of black holes and the associated feedback; [5] the effects due to re-ionisation of the universe by the ultra-violet radiation field.

The computer code for each section implements astrophysically motivated algorithms, each process drawing on the inputs provided by each of the other modules. The modules link together to form a network of non-linear equations that are integrated in time to trace the evolving properties of the galaxy population. In total the model uses over 50,000 lines of computer code. Further details are described in appendix B, while Baugh (2006) presents an introduction to the internal workings of the code.

2.2 Galform and Dark Matter

In order to run, Galform requires a forcing function[1] that represents the positions, and subsequent collisions, of all lumps of dark matter containing galaxies, at all times (referred to as the "merger histories of the Dark Matter Haloes"). This information is extracted from the Millennium simulation (a large Dark matter simulation described in appendix B), and, with it, Galform can model the far more complicated behaviour of baryonic (i.e. normal) matter. It is the baryonic matter that is responsible for the more

[1] We use the term 'forcing function' in the differential equation sense, referring to a function that appears in a network of differential equations and which depends on time only. Knowledge of this function is required before one can attempt to numerically solve or integrate the system of equations.

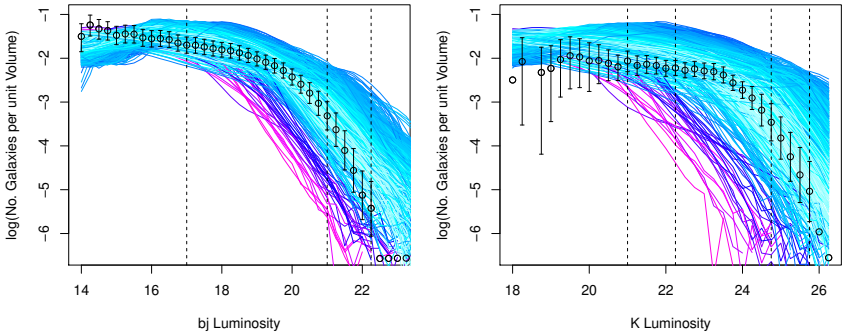

Figure 1: The bj (left) and K (right) Luminosity Functions giving the (log) number of galaxies per unit volume, binned by luminosity. The data are shown as the black points, along with 2 sigma intervals representing all relevant uncertainties identified in section 5. The coloured lines are the Galform outputs from 993 Wave 1 runs of the model described in section 4.2, none of which were found to be acceptable as they didn't pass sufficiently close to all of the observed data points. The vertical lines show the 7 outputs chosen for emulation also described in section 4.2.

intricate processes involved in galaxy formation.

As the Millennium simulation covers a substantial volume (1.63 billion light years cubed), its results are split into 512 sub-volumes, each of which can be used as a forcing function to the Galform model. The run time for one evaluation of the Galform model on a single sub-volume is approximately 30 minutes. The Galform group provided shared access to a cluster of 256 processors. Previous attempts by the cosmologists to calibrate Galform focussed on the first 40 sub-volumes out of 512, and we follow this approach here while taking account of the uncertainty this generates.

2.3 Inputs

Each module of Galform has associated input parameters, which define the workings of the module. The Galform model has a total of 17 inputs that relate to various uncertain physical processes involved in galaxy formation. We denote the vector of 17 input parameters as x. In order to run the code, the astrophysicist must specify values for each of these input parameters. Some parameters are quite well defined by numerical experiments or targeted observational data, but others are highly uncertain.

All 17 inputs x, along with their considered ranges, and associated modules are shown in table 1. Also shown are the variables that are initially considered, and those varied in Wave 1 of our analysis: this will be discussed in section 4.3.

Input Parameters	symbol	min	max	Initial Variables	Varied in W1 ($x_{[B]}$)	Process Modelled
vhotdisk	$V_{\rm hot,disk}$	100	550	x	x	SNe feedback
vhotburst	$V_{\rm hot,burst}$	100	550	x	x	·
alphahot	$\alpha_{\rm hot}$	2	3.7		x	·
alphareheat	$\alpha_{\rm reheat}$	0.2	1.2	x	x	·
alphacool	$\alpha_{\rm cool}$	0.2	1.2	x	x	AGN feedback
epsilonEdd	$\epsilon_{\rm Edd}$	0.004	0.05			·
epsilonStar	$\epsilon_\star$	10	1000	x	x	Star Formation
alphastar	$\alpha_\star$	-3.2	-0.3			·
yield	$p_{\rm yield}$	0.02	0.05		x	·
tdisk	$t_{\rm disk}$	0	1			·
stabledisk	$f_{\rm stab}$	0.65	0.95	x	x	Disk stability
tau0mrg	$f_{\rm df}$	0.8	2.7			Galaxy Mergers
fellip	$f_{\rm ellip}$	0.1	0.35			·
fburst	$f_{\rm burst}$	0.01	0.15			·
FSMBH	$F_{\rm bh}$	0.001	0.01			·
VCUT	$v_{\rm cut}$	20	50			Reionisation
ZCUT	$z_{\rm cut}$	6	9			·

Table 1: Table of Parameter Ranges (which were converted to -1 to 1 for the analysis), including the initial variables considered and those that are possibly active and analysed in Wave 1 (referred to as $x_{[B]}$). Parameters are grouped by physical process.

2.4 Outputs

Galform provides several different sets of output data related to various physical characteristics of the simulated galaxies. Observational data of differing degrees of accuracy are available for comparison with the Galform model output, the most important of these being the bj and K Luminosity Functions. These Luminosity functions give the (log) number of galaxies per unit volume, binned with respect to luminosity. We consider two types of luminosity function: "bj" and "K" which correspond to blue and infrared wavelengths of light, and which are more representative of younger and older galaxies respectively (but not exclusively: most galaxies emit measurable amounts of both wavelengths of light). Figure 1 shows the bj and K luminosity functions in the left and right panels respectively. The 993 model runs used in wave 1 are given by the coloured lines (see section 4.2). Observed data is given by the black dots, with 2σ error bars representing all the uncertainties described in section 5. Such data was gathered by the 2dFGRS sky survey (Colless et al. (2001)): telescopes sweep across sectors of the sky measuring hundreds of thousands of galaxies and their properties. In order to convert the data into the form given in figure 1 (current day, absolute luminosities), the data must be heavily processed to correct for several factors, for example, the redshifting of light emitted by galaxies due to their velocity away from Earth, and the time delay in receiving such light which implies we are effectively measuring distant galaxies billions

of years in the past (this is discussed in section 5.2). See Norberg et al. (2002) and Cole et al. (2001) for the details of such processing for the bj and K data respectively.

The Luminosity Function data set represents the most accurately measured observational data available and is seen as the benchmark by which models of galaxy formation are judged (Norberg et al. (2002)). Even if a particular galaxy formation model performs well with respect to other data sets, if it does not match the Luminosity function to an acceptable level then that model will be discarded. For these reasons, it was decided to focus our analysis on identifying the regions of input space that give rise to matches between the model output and the bj and K observed luminosity functions. Additional data sets could then be used at a later date to restrict the input space further.

3 Uncertainty Analysis for Computer Simulators

3.1 Uncertainty in complex models

Our aim is to identify that region of the input space of the Galform simulator for which certain aspects of Galform output match closely to observations. As such, this study falls within the general area of the analysis of uncertainty arising when we study complex physical systems by means of mathematical models typically implemented as computer simulators. The general version is as follows. A computer simulator f takes as input the vector x, which represents certain physical properties of a system of interest. The simulator output vector, $f(x)$, corresponds to certain aspects of the behaviour of the system. For given inputs, this behaviour is determined by equations embodying the relevant theoretical knowledge relating system properties to system behaviour. This approach is common to many areas of science. We can talk of an emergent methodology because, despite the enormous differences between each of the individual models, all such problems of physical modelling confront a similar collection of basic uncertainties.

[1] Parameter uncertainty. We do not know the appropriate values of the inputs x to the simulator. In some cases, we may not even know whether there is any appropriate choice for the inputs. Galform is a case in point. If we have misrepresented the underlying physics, for example if the role of Dark Matter is not supported by observational evidence, then the meaning of the model and the interpretation of the parameters will be called into question. In particular, were we to discover that there were no choices of inputs for which Galform output matched observations in our universe, then that might provide part of the evidence which would call the current account of cosmology into question.

[2] Simulator uncertainty. For any choice of inputs, x, the output of the Galform code $f(x)$ is a deterministic computer function. However, many computer simulators are very expensive, in time and resources, to evaluate, for any choice of inputs. In practice, it is appropriate to consider that the output values of such a simulator are unknown except at the input choices at which the simulator has been evaluated. An important stage in the analysis, therefore, is the construction of a statistical representation or **emulator** for the simulator. The emulator represents our uncertainty about

the value of the function at each possible input choice, and therefore acts both as an approximation to the function and as an assessment of the uncertainty introduced by the approximation. Much of the literature on **computer experiments** is concerned with efficient methods for building emulators; see for example Sacks et al. (1989); Santner et al. (2003); Currin et al. (1991). For our Galform investigations, we have made many evaluations of the simulator. Even so, emulation has proved essential in extending our uncertainty description from the function evaluations to the remainder of the input space.

[3] Structural uncertainty. However carefully we have constructed our model, there will always be a difference between the system and the simulator. Inevitably, there will be simplifications in the physics, based on features that are too complicated for us to include, features that we do not know that we should include, mismatches between the scales on which the model and the system operate, and simplifications and approximations in solving the equations determining the system. Often, understanding this structural uncertainty will be one of the most challenging aspects of the analysis.

[4] Observational error. This uncertainty arises when we match our model to system observations. Often, in complex physical systems, the observations are themselves somewhat indirect, being assessed on the basis of extensive preprocessing based on various additional theoretical constructs, that are external to the theory underlying and represented by, the computer model. Further, the measurements may not directly correspond to the outputs of the simulator and therefore require an extra layer of interpretation and analysis before the model predictions and the system observations can be compared. The observational error in Galform is of a particularly complex form, requiring considerable processing to transform the system observations to a comparable spatio-temporal resolution to the simulator outputs.

[5] Initial condition and forcing function uncertainty. This corresponds to all of the other aspects of the simulator which need to be specified before the model may be evaluated. For example, the Galform simulator requires a full spatial specification of the arrangement of Dark Matter at all times in the development of the universe, and so we need to account for the uncertainty introduced as we do not know this configuration. Often such specifications are too complex to be treated in the same manner as parameter uncertainty, as is discussed by House et al. (2009).

In this study, we will describe how we address each of these sources of uncertainty for the Galform project. We aim to be careful and thorough, but we must also recognise that, for a complex model such as Galform, uncertainty modelling is a process which is similar in many ways to the physical modelling process on which we are building. Quantifications of uncertainty depend on complex scientific judgements over which different experts may have different views. Further, expert knowledge is held collectively over a wide community of experimenters, observationalists, theoreticians and modellers. Therefore, it is as misleading to talk of a definitive assessment of the uncertainty associated with Galform as it would be to talk of a definitive form for the Galform model itself as experts would not currently agree on the precise form the Galform code should take. Assessment of uncertainty is an ongoing process for models which are, themselves,

undergoing continuous development. Our account documents one iteration in this ongoing process, albeit one for which the uncertainty analysis is carried out to a much greater level of detail than is usual in this field (or indeed in most analyses of complex physical models in any area of application of which we are aware).

3.2 Linking the simulator with the system

We now introduce the general structure describing the relationship between the simulator and the physical system. We describe this link for the Galform simulator, but the ingredients are common to many simulator analyses. We denote by z the vector of observations that we use for this study. Our choice for z will be the observed numbers of galaxies of various degrees of luminosity, assessed separately for younger and for older galaxies and expressed on the log scale. We describe the relationship between the observations, z, and the vector of true physical system values, y, as

$$z = y + \epsilon_{obs} \tag{1}$$

where ϵ_{obs} is the experimental error, which we judge to be uncorrelated with y (ϵ_{obs} represents uncertainty of type [**4**] in section 3.1).

Is the theoretical understanding of Galaxy formation, as embodied in Galform, consistent with observations z? Galform is represented as a function, which maps the inputs x to the outputs $f(x)$. The theoretical description involves the notion that when we evaluate Galform at the actual system properties, x^* say, then we should reproduce the actual system behaviour y. This does not mean that we would expect perfect agreement between $f(x^*)$ and y. Although Galform is a highly sophisticated simulator, it still offers a necessarily simplified account of the evolution of galaxies, and approximates the numerical solutions to the governing equations. The simplest way to view the difference between $f^* = f(x^*)$ and y is to express this as

$$y = f^* + \epsilon_{md}, \tag{2}$$

where we consider that ϵ_{md} is uncorrelated with f^* and with x^*. Expressing our judgements about the likely size of the *model discrepancy*, ϵ_{md}, determines how close a fit between model output, f^*, and observation y we require for an acceptable level of consistency between theory and observation. ϵ_{md} represents uncertainties [3] and [5] from section 3.1. Note that ϵ_{md} is a vector of length equal to the number of observations considered, and may have a rich covariance structure across the different outputs (see section 5.1).

We search for choices of input x for which the output $f(x)$ is sufficiently close to y that we would declare the observed output to be compatible with the predictions of the model, when we allow for model discrepancy. In practice, all that we can compare is $f(x)$ and z, which we do by combining (1) and (2). Achieving an acceptable match, for a particular input choice x, does not mean that the model is "correct" or that a parameter choice which achieves the match corresponds to the "true" value of the parameters, but simply that this version of the model will have met the challenge of

reproducing an important observational aspect of the galaxy formation study within our agreed tolerance level. Similarly, identifying the whole collection of possible choices of inputs x which achieve an acceptable match and the subsequent analysis of such a collection, is informative as regards the model and the galaxy formation process itself, as is discussed in section 8.

The form (2) is simple and intuitive, and is widely used in computer modelling studies. In our case, this corresponds to the natural approach in which we ask whether we could view Galform, with appropriate choice of inputs, as adequately reproducing the observed universe, within the tolerance set by the model discrepancy. In this account, we therefore ignore all of those additional aspects of our uncertainty modelling which would correspond to a more sophisticated analysis of model discrepancy, based, for example, on informed expert judgements as to the ways in which the Galform simulator is likely to evolve over the coming years. A detailed specification of such features would potentially be highly insightful, and might result in a much richer correlation structure across the elements of the discrepancy vector; see Goldstein and Rougier (2009). However, as we shall describe, it is challenging even to make a meaningful order of magnitude assessment of discrepancy variation, and so, as a first uncertainty quantification for Galform, we chose to focus on the most important large scale components of uncertainty.

3.3 Bayes Linear Analysis

In this case study, we follow the *Bayes linear* approach to uncertainty quantification and analysis. This approach is relatively simple in terms of belief specification and analysis, as it is based only on mean, variance and covariance specifications which, following de Finetti, we take as primitive; see De Finetti (1974, 1975). The appropriate updating rules for expectations and variances for a vector B, given a vector D are

$$\mathsf{E}_D[B] = \mathrm{E}(B) + \mathrm{Cov}(B, D)\mathrm{Var}(D)^{-1}(D - \mathrm{E}(D)), \tag{3}$$

$$\mathsf{Var}_D[B] = \mathrm{Var}(B) - \mathrm{Cov}(B, D)\mathrm{Var}(D)^{-1}\mathrm{Cov}(D, B). \tag{4}$$

$\mathsf{E}_D[B]$ and $\mathsf{Var}_D[B]$ are termed the *adjusted mean and variance of B given D* (Goldstein (1999); Goldstein and Wooff (2007)).

In this formulation, the probability of an event is the expectation of the corresponding indicator function. Conditional expectation may be viewed as the special case of belief adjustment where we base the adjustment on the indicator functions for a collection of events which constitute a partition. There are many areas of similarity between full Bayes and Bayes linear analyses. In particular, a full Gaussian specification for all of the relevant quantities would lead to similar updating formulae. For a detailed treatment, see Goldstein and Wooff (2007). An overview of the approach is given in Goldstein (1999).

There are two basic interpretations that we may give for a Bayes linear analysis. The first view arises as, when we attempt to carry out a full Bayesian analysis, we may face

two kinds of difficulty. Firstly, the probability specification that we may be required to make in order to describe fully all of the uncertainties for the problem may be extremely complex and subtle, and secondly the full Bayesian analysis may be technically very challenging to carry out and highly non-robust, as the posterior judgments may depend on aspects of the prior specification which we are unable to specify with confidence, given constraints of time, resource and knowledge.

In such circumstances, the Bayes linear analysis that we propose may be viewed as a pragmatic compromise to this full analysis. The task of probability specification is simplified as we only need to give means, variances and covariances for all of the random quantities involved in the analysis. The subsequent analysis is simplified as, rather than carrying out a full posterior assessment given the observations, we may carry out a Bayes linear update, as determined by equations (3) and (4). The adjusted expectation for B given D is the best linear fit for B, using the elements of D, in terms of minimising expected squared error loss - this minimum expected loss is the adjusted variance. This minimisation depends only on the second order specification that we have made. We may expand the Bayes linear analysis by using whichever collection of functions of the elements of D that we wish, in order to assess the adjusted expectation, as long as, for each chosen function of the elements of D, we are prepared to assess the full corresponding second order specification. If we introduce all functions of the elements of D, then we are effectively adjusting B by the partition on D, and we retrieve the conditional Bayesian expectation given the observed outcome for D.

In many problems, the extra effort and complications involved in assessing and analysing the full probabilistic structures over the complex and high dimensional input and output spaces which are required in order to carry out the full Bayes analysis may not be rewarded by a corresponding improvement in accuracy. For example, in our experience, it is usually reasonable to suppose that we can elicit expert judgements about the order of magnitude of quantities such as model discrepancy terms which are sufficient for us to make variance and covariance assessments across the various components of the model. The qualitative structure that we impose in order to make such elicitations is based on relations such as (1) and (2), which impose the requirement that the two terms on the right hand side of each equation are uncorrelated. This already is a strong assertion, and we might well be reluctant to extend this to a judgement of full probabilistic independence between the corresponding terms. Therefore, we may prefer that our analysis should only depend on those properties which follow directly from this specified lack of correlation, rather than relying on an infinite number of further joint orthogonality constraints as required by full probabilistic independence.

Even were we able to make a meaningful full probabilistic elicitation for the problem described in this paper, then, due to the iterative nature of our approach, which repeatedly reduces the volume of the space which we are exploring by imposing a series of very complicated and highly non-linear constraints, it would be enormously challenging even to construct a tractable Bayesian MCMC analysis. In contrast, the Bayes linear analysis is comparatively straightforward and is sufficient to achieve the study objective of identifying a class of good matches to observed history. Therefore, while it would be of interest to compare our results with the full Bayes version of the calculations that

we will describe, the technical challenges in doing so would be very great, and this is quite apart from the inherent non-robustness of the iterative version of the full Bayes analysis, due to the extremely complex, multi-modal form of the likelihood function. Having said this, if such a full Bayes analysis were feasible, it would contain additional information about all of the relationships within the problem compared to the Bayes linear approach, and so would allow more detailed inferences to be drawn.

The reason that the Bayes linear approach is successful as a surrogate for the full Bayes analysis derives from the second, and more fundamental, interpretation of the meaning and value of this analysis. We consider that a Bayesian analysis has value largely because we view it as the appropriate way to combine expert judgement and observations to give appropriate posterior judgements. However, there are two problems with this view. Firstly, due to the complexity of the problems that we face, often we are unable to make a prior specification which adequately represents our true state of uncertainty, and so the posterior analysis unavoidably inherits this lack of accuracy. Secondly, the operation of conditioning, itself, does not offer a complete description as to how judgements should be modified given new evidence. This is a larger issue than we have space to address here. For those interested in the fundamental reasons why we view the Bayes linear analysis as appropriate for complex uncertainty problems, under partial belief specification, we refer to the discussion in section 4 of Goldstein (2006) where the temporal relationships between Bayes linear adjustment and posterior beliefs are described and the foundational properties of standard Bayesian approaches are shown to be inherited from their status as Bayes linear adjustments. The role of such foundational considerations in the interpretation of the Bayes linear analysis for problems arising in the study of large computer models is discussed in detail in Goldstein (2010) where issues arising in the treatment of the Galform model are used as an illustration.

3.4 Emulation

We are interested in the behaviour of Galform over the whole of its specified input space. The substantial run time and the high dimensional input space combine to make direct exploration by model runs alone infeasible. We express our beliefs about Galform outputs at all locations in the input space by constructing an *emulator* (see uncertainty type [2] in section 3.1). An emulator is a stochastic belief specification for a deterministic function (Craig et al. (1996, 1997); O'Hagan (2006); Oakley and O'Hagan (2002); Conti et al. (2009); Higdon et al. (2004), and for a cosmology application Heitmann et al. (2009)). The emulator is much faster to evaluate than the simulator, so that we may explore the input space using the emulator, while taking into account the extra uncertainty that we have introduced by substituting emulator for simulator evaluations.

We construct our emulator for output i of the function $f(x)$ to have the form

$$f_i(x) = \sum_j \beta_{ij}\, g_{ij}(x) + u_i(x), \tag{5}$$

where $B = \{\beta_{ij}\}$ are unknown scalars, g_{ij} are known deterministic functions of x and $u(x)$, uncorrelated with B, is a weakly stationary stochastic process with constant vari-

ance. The regression term on the right hand side of equation (5) expresses the global behaviour of the function while $u(x)$ represents localised deviations from this global behaviour near to x.

In the Bayes Linear approach, the emulator specification requires a mean vector and a variance matrix for B and values for the mean, variance and correlation function of u. A simple specification for $u(x)$ is to suppose, for each x, that $u_i(x)$ has zero mean with constant variance and where $\mathrm{Corr}(u_i(x), u_i(x'))$ is a function of $\|x - x'\|$.

With high dimensional input spaces, it is common to find, for any output, f_i say, that a subset $x_{[i]}$ of the inputs has the most influence in explaining the variation in the value of $f_i(x)$, where $x_{[i]}$ varies with i. We reform the emulator as

$$f_i(x) = \sum_j \beta_{ij}\, g_{ij}(x_{[i]}) + u_i(x_{[i]}) + w_i(x), \tag{6}$$

where $u_i(x_{[i]})$ has zero mean, and covariance structure given by $\mathrm{Cov}(u_i(x_{[i]}), u_i(x'_{[i]})) = \sigma^2_{u_i} \exp(-||x_{[i]} - x'_{[i]}||^2/\theta_i^2)$ (this is the commonly used Gaussian form, see section 4.3 for more details). Here $w_i(x)$ is a "nugget term" with constant variance $\sigma^2_{w_i}$ over x, zero mean and $\mathrm{Cov}(w(x), w(x')) = 0$ for $x \neq x'$. The collection $x_{[i]}$ is often called the *active variables* for f_i, and $w_i(x)$ expresses all of the variation in $f(x)$ which arises if we view the emulator $f(x)$ simply as a function of $x_{[i]}$.

We can use the emulator to evaluate the expectation and variance of the function, for any input x and the covariance between the values of f at any pair of points x, x'. From (6), these are

$$\mu_i(x) = \mathrm{E}(f_i(x)) = \sum_j \mathrm{E}(\beta_{ij})\, g_{ij}(x_{[i]}), \tag{7}$$

$$\begin{aligned} \kappa_i(x, x') &= \mathrm{Cov}(f_i(x), f_i(x')) \\ &= \mathrm{Cov}(\sum_j \beta_{ij}\, g_{ij}(x_{[i]}), \sum_j \beta_{ij}\, g_{ij}(x'_{[i]})) + \sigma^2_{u_i} \exp(-||x_{[i]} - x'_{[i]}||^2/\theta_i^2) + \sigma^2_{w_i} I_{x,x'} \end{aligned} \tag{8}$$

where $I_{x,x'} = 1$ if $x = x'$, and zero otherwise. In the Bayes linear approach this is all that is required to be able to update our beliefs in terms of expectation and variances, about the model output $f_i(x)$ *at a new, unevaluated input point* x, given a set of model evaluations, as we now describe.

Say we have performed n model evaluations over some space filling design. We write the locations of the n runs in input space as $x^{(k)}$ with $k = 1, .., n$ where each $x^{(k)}$ represents the vector of inputs for the kth run. Similarly $x^{(k)}_{[i]}$ is defined to be the vector of Active Variable inputs for the kth run. We define $D_i = (f_i(x^{(1)}), f_i(x^{(2)}), ..., f_i(x^{(n)}))^T$, that is the column vector of the n evaluation outputs for output i, the prior expectation of which $(\mathrm{E}(D_i))$ can be found using equation (7). Replacing the random quantity B in the Bayes Linear update equation (3) with the unknown output $f_i(x)$ at input x, and

replacing D with D_i, gives the adjusted expectation $\mathrm{E}_{D_i}[f_i(x)]$ to be:

$$\begin{aligned} \mathrm{E}_{D_i}[f_i(x)] &= \mathrm{E}(f_i(x)) + \mathrm{Cov}(f_i(x), D_i)\mathrm{Var}(D_i)^{-1}(D_i - \mathrm{E}(D_i)) \\ &= \sum_j \mathrm{E}(\beta_{ij})\, g_{ij}(x_{[i]}) + t(x)^T A^{-1}(D_i - \mathrm{E}(D_i)), \end{aligned} \tag{9}$$

where, by equation (7), $t(x) = (\kappa_i(x, x^{(1)}), \kappa_i(x, x^{(2)}), ..., \kappa_i(x, x^{(n)}))^T = \mathrm{Cov}(f_i(x), D_i)^T$ is the column vector of covariances between the new and known points, and A is the matrix of covariances between known points: an $n \times n$ matrix with elements $A_{jk} = \kappa_i(x^{(j)}, x^{(k)})$. The Adjusted Variance $\mathrm{Var}_{D_i}[f_i(x)]$ can similarly be found from equations (4) and (7) giving:

$$\begin{aligned} \mathrm{Var}_{D_i}[f_i(x)] &= \mathrm{Var}(f_i(x)) - \mathrm{Cov}(f_i(x), D_i)\mathrm{Var}(D_i)^{-1}\mathrm{Cov}(D_i, f_i(x)), \\ &= \mathrm{Var}(\sum_j \beta_{ij}\, g_{ij}(x_{[i]})) + \sigma^2_{u_i} + \sigma^2_{w_i} - t(x)^T A^{-1} t(x). \end{aligned} \tag{10}$$

The adjusted expectation and variance, $\mathrm{E}_{D_i}[f_i(x)]$ and $\mathrm{Var}_{D_i}[f_i(x)]$, represent our updated beliefs about the output of the Galform function $f(x)$ at input x given a set of n model runs D_i, and are used in the implausibility measures described in section 3.5.

There is some debate in the computer experiment literature as to whether it is preferable to put a lot of effort into constructing the regression terms in the emulator or whether it is better to construct a simple mean function and to place more weight on the residual process $u(x)$. We prefer, where possible, to put as much detail as is feasible into the mean function, for the following reasons.

Firstly, many physical models, and Galform in particular, exhibit strong and physically interpretable monotonicities which are naturally expressed through the mean function. Secondly, it is easier for the expert to assess whether the emulator formulation is consistent with informed scientific judgement about the behaviour of the function if a large proportion of the variability is expressed through regression terms. Thirdly, if much of the structure of the emulator is encoded in the regression function, then this simplifies various of the calculations that we need to make when comparing the model to observations and suggests very cheap approximations to calculations which would otherwise be very expensive. Finally, in our experience, the form of local process, $u(x)$, can be difficult to assess, even with large numbers of function evaluations. Partly, this is because there is a fundamental confounding between the location of the mean function, the size of the residual variance ($\sigma^2_{u_i} + \sigma^2_{w_i}$), and the strength of the residual correlation, parameterised by θ_i. Partly, also, this is because any form of correlation function that we fit necessarily approximates the different degrees of smoothness of the function across different areas of the input space, and many methods of estimating smoothness parameters are potentially non-robust when applied to processes which do not fit exactly to the assumptions that are used to generate the fitting algorithms. Therefore, we prefer to model as much of the variation in the function as we can by the regression form, to reduce the residual variance as much as is feasible, and then to be fairly conservative in choosing the length of correlation θ_i that we shall impose.

In general computer experiments, we choose our form for the emulator by a combination of expert judgement based on physical intuition and experience with earlier versions of the model and, where appropriate, by preliminary experiments with fast approximate version of the simulator. In our case, we were able to make a collection of evaluations of the simulator, based on a Latin Hypercube design, which was sufficiently large to allow us to fit the emulator directly from our functional evaluations. Therefore we proceeded as follows, for each output that we chose to emulate.

Firstly, we carried out statistical model fitting, given the collection of runs, to select the deterministic functions g_{ij}, to assess the values of the coefficients B and to assess the residual variance and covariance function, $u(x)$ and, where appropriate, to identify active subsets $x_{[i]}$. We then used equations (9) and (10) to update our beliefs about the function $f(x)$, obtaining the adjusted expectation and variance $\mathrm{E}_{D_i}[f_i(x)]$ and $\mathrm{Var}_{D_i}[f_i(x)]$, for any input of interest x. We then checked that the form of the emulator was physically meaningful. Finally, we carried out a diagnostic analysis on our emulator. We will give details of these stages below.

3.5 History Matching

The aim of this study is to estimate the set of input values $\mathcal{X}^*$ for which the evaluation of $f(x)$ gives an acceptable match to the observations z, and to obtain a substantial collection of realised evaluations of the function which actually do yield acceptable matches and which may then be used to explore the match between other aspects of the Galform output and the corresponding observational information.

We refer to the process of identifying the collection $\mathcal{X}^*$ as *history matching*. This terminology is common in various applications (e.g. Raftery et al. (1995)), and in particular in oil reservoir modelling (Craig et al. (1996, 1997); Cumming and Goldstein (2009)), where it refers to the process of adjusting the inputs to a simulator of an oil reservoir until the output closely reproduces features such as the historical oil production and pressure profiles at all of the wells. The emphasis on identifying all of the possible matches to observation is ours. (Pragmatically, reservoir engineers often stop when a few matches, or even just one, have been obtained.)

History matching may be compared to model *calibration* in which we suppose there is a single "true but unknown" value x^* and our objective is to make probabilistic statements as to this value, based on a prior specification for x^*, the collection of model evaluations and the observed history (Kennedy and O'Hagan (2001); Higdon et al. (2008); Goldstein and Rougier (2006)). Calibration and history matching are thematically related, but fundamentally different. For example, calibration will always result in a proper posterior distribution over the input space, while history matching might lead to the conclusion that the collection of acceptable matches was empty. It would be of great interest to find that $\mathcal{X}^*$ was empty in the Galform study, as that might suggest possible defects in the general theory underlying the simulation process. However, in this study, we do find a collection of good fits to the observations. Our view is that history matching, as a form of model checking, is always of interest for

assessing computer models and calibration sometimes is. Even when we wish to carry out a model calibration, it is often good practice first to carry out a history match, partly to see whether such a match is achievable, and partly to reduce the size of the input space over which the calibration exercise will need to be performed.

Our approach to history matching is based on the assessment of certain *implausibility measures* (Craig et al. (1996, 1997)). An implausibility measure is a function defined over the input space which, when large, suggests that the match between model and system would exceed our stated tolerance. We may build this up as follows, for a single output $f_i(x)$, where i labels the output. For a given choice, x^*, we would like to assess whether the output $f_i(x^*)$ differs from the system value y_i by more than the tolerance that we allow in terms of model discrepancy. Therefore, we would assess the standardised distance

$$\frac{(y_i - f_i(x^*))^2}{\mathrm{Var}(\epsilon_{md:i})}$$

In practice, we cannot observe y_i and so we must compare $f_i(x^*)$ with the observation z, introducing measurement error, with corresponding standardised distance

$$\frac{(z_i - f_i(x^*))^2}{\mathrm{Var}(\epsilon_{md:i}) + \mathrm{Var}(\epsilon_{obs:i})} \tag{11}$$

However, for most values of x, we are not able to evaluate $f(x)$ so we use the emulator and compare z_i with $\mathrm{E}(f_i(x))$. Therefore, the implausibility function is defined as

$$I^2_{(i)}(x) = \frac{(\mathrm{E}(f_i(x)) - z_i)^2}{\mathrm{Var}(\mathrm{E}(f_i(x)) - z_i)} = \frac{(\mathrm{E}(f_i(x)) - z_i)^2}{\mathrm{Var}(f_i(x)) + \mathrm{Var}(\epsilon_{md:i}) + \mathrm{Var}(\epsilon_{obs:i})} \tag{12}$$

When $I_{(i)}(x)$ is large, this suggests that, even given all the uncertainties present in the problem, we would be unlikely to view as acceptable the match between model output and observed data were we to run the model at input x. Therefore, we consider that choices of x for which $I_{(i)}(x)$ is large can be discarded as potential members of the set $\mathcal{X}^*$. We discard regions of the input space by imposing suitable cutoffs on the implausibility function in that we discard x unless $I_{(i)}(x) < c$. The choice of cutoff c comes from consideration of the fraction of space removed, and from general unimodality arguments, as follows. Regarding the individual univariate Implausibility Measures $I_{(i)}(x)$, if we consider that for fixed x the appropriate distribution of $(\mathrm{E}(f_i(x^*)) - z_i)$ is both unimodal and continuous, then we can use the 3σ rule (Pukelsheim 1994) which implies quite generally that if $x = x^*$, then $I_{(i)}(x) < 3$ with a probability of greater than 0.95. This result applies even, for example, for highly skew or heavy tailed distributions. Values higher than 3 would suggest that the point x could be discarded.

In our comparisons, we have a separate implausibility function, given by equation (12), for each output that we use for history matching. We may choose to make some intuitive combination of the individual implausibility functions as a basis of eliminating portions of the input space. The simplest of these is obtained by maximising $I_{(i)}(x)$ over the considered outputs and we hence define the Maximum Implausibility Measure

$$I_M(x) = \max_i I_{(i)}(x). \tag{13}$$

This measure is used in later waves of our analysis and it represents a major part of the definition of an acceptable match. It is, however, sensitive to problems concerning the inaccuracies of individual emulators, and so we define the Second and Third Maximum Implausibility Measures $I_{2M}(x)$ and $I_{3M}(x)$ as:

$$I_{2M}(x) = \max_i(\{I_{(i)}(x)\} \setminus I_M(x)), \tag{14}$$

$$I_{3M}(x) = \max_i(\{I_{(i)}(x)\} \setminus \{I_M(x), I_{2M}(x)\}), \tag{15}$$

that is defining $I_{2M}(x)$ and $I_{3M}(x)$ to be the second and third highest value out of the set of univariate measures $I_{(i)}(x)$ respectively. These were used in the first wave as they were considered relatively safe measures in that they were less sensitive to the possibility that one of the emulators was inaccurate. We also construct the natural multivariate analogue of the implausibility, for later waves, which takes the form:

$$I(x) = (z - \mathrm{E}(f(x)))^T (\mathrm{Var}(z - \mathrm{E}(f(x))))^{-1} (z - \mathrm{E}(f(x))) \tag{16}$$

The multivariate form is more effective for screening the input space, but it does require careful consideration of the covariance structure for the various quantities.

History matching is an iterative process. We begin by emulating Galform over the whole input space. We evaluate our implausibility measures over the whole space and remove from the space all input choices for which the implausibility measure is large. We then re-sample within the remaining input space, denoted $\mathcal{X}_1$, and re-emulate Galform within this reduced space. This is termed *refocusing*, and we proceed to employ this process iteratively as represented by the following algorithm. At each iteration or Wave:

1. A design for a set of runs over the current non-implausible volume $\mathcal{X}_i$ is created, using a latin hypercube design with a rejection strategy based on each of the preceding implausibility measures.
2. These runs are used to construct a more accurate emulator defined only over the current non-implausible volume $\mathcal{X}_i$.
3. The implausibility measures are then recalculated over $\mathcal{X}_i$, using the new emulator.
4. Cutoffs are imposed on the Implausibility measures and this defines a new, smaller non-implausible volume $\mathcal{X}_{i+1}$ which should satisfy $\mathcal{X}^* \subset \mathcal{X}_{i+1} \subset \mathcal{X}_i$.
5. Unless the emulator variance is now small in comparison to the other sources of uncertainty, or unless computational resources are exhausted, return to step 1.
6. Generate a large number of acceptable runs from the final non-implausible volume.

The reasons that we may hope to further reduce the acceptable space at each iteration are firstly that we produce a higher relative density of runs at each stage, so that emulation is more effective, secondly that we may expect the function to become smoother and so easier to emulate as we reduce the area of the input space, and thirdly because, when we have accounted for much of the uncertainty related to the most important active variables, then variables which did not account for much of the variability in the original emulation may take on larger importance and therefore allow us to resolve

more of the uncertainty of the function. In this study, we refocused four times, and then carried out a fifth set of evaluations which produced a large number of runs which gave good matches to observations. This continued refocusing is very useful, but it also brings its own complications, as the only way in which we can determine whether an input value lies within our retained collection of potential history matches is by applying each implausibility function in turn and seeing whether each such evaluation is small enough for the input choice to be retained. This raises practical computational issues, which makes it important to have fast approximate methods to screen the input space, and also raises basic questions about practical visualisation methods to help us to represent and interpret the shape of the input space which we have retained.

All of these complications reflect the enormous difficulty of carrying out a fully Bayesian history matching exercise over a corresponding number of waves to that of our study. Rather than constructing the full probabilistic edifice, we have identified certain key aspects of the subjective judgements relating to the function, the model discrepancy and the observational error and used these to construct an event with low subjective probability for $x \in \mathcal{X}^*$ and much higher, though not explicitly evaluated, probability otherwise, which we have used to progressively filter the input space. The successes of this method, for example in this study we do identify a rich space of acceptable fits, suggests that we are indeed exploiting the probabilistic judgments in a meaningful way, but this does raise the basic question as to whether there is some tractable intermediary between our version of history matching and the full Bayesian solution that would be even more effective in achieving our goals.

4 First Wave Analysis

4.1 General Designs for Computer Model Experiments

We have to explore the high-dimensional input space of the Galform model, which takes a significant amount of time to run. Therefore the design for the set of inputs where models will be evaluated is very important: (Currin et al. 1991; Sacks et al. 1989; Santner et al. 2003). The design should be space-filling (to maximise coverage of the space), and approximately orthogonal (where possible) as we will be fitting polynomials to the outputs when constructing the emulator. Various designs have been discussed in the Computer Model literature (Santner et al. 2003), with a popular choice being the Maximin Latin hypercube design. An n point Latin Hypercube design is constructed by dividing the range of each of the input variables into n equal intervals. Points are placed so that one point will occupy each of the n intervals, for each input variable. Maximin Latin Hypercube designs are constructed by generating many Latin Hypercube designs and selecting the one that has the maximum 'minimum distance' between points. They are approximately orthogonal designs and suffer no projection issues as any lower dimensional projection remains a Latin Hypercube.

4.2 The Wave 1 Design

The first stage in the collaboration concerned History Matching using a smaller number of input variables than were present in the full Galform model, in order to demonstrate the methodology in a simplified version of the problem. As the collaboration progressed we extended our aims to include an analysis of the full model with all 17 input parameters. This evolution in priorities has had an impact on the general structure of the analysis, as will be noticeable from the initial design choices described here.

When considering the initial design, expert judgements were used to identify a subset of the 17 inputs which would have either significant effects on the bj and K luminosity function outputs, or be of physical interest to the cosmologists (expert judgements in this study were made by Richard Bower). These 6 inputs are shown in the 'Initial Variables' column of table 1. When the Galform project began, it was impossible to run the model while varying more than 11 input parameters simultaneously due to technical issues with the code. Therefore, we constructed two maximin Latin Hypercube designs: the first over the 6 inputs identified as important, and the second over the 11 inputs thought to be less significant. An initial analysis of the first set of runs, suggested that acceptable matches could, most likely, only be found for extremely low values of the 5th input parameter epsilonStar, with the Galform function decreasing rapidly at such values. This made intuitive sense as the relevant physical process is dependent upon the inverse of epsilonStar (see appendix B). We therefore reparameterised this input as epsilonStar^{-1} for all subsequent analysis. Comparison of the variance of the outputs in each data set implied that one parameter (alphahot) out of the 11 initially discarded inputs, had a clearly significant effect on the luminosity functions, and after careful consultation, this input was promoted into the active group. At this point, the cosmologists requested that the parameter "yield" also be promoted, as recent physical evidence had suggested that the value assigned to this parameter in previous analyses (0.02) was too low, and hence the cosmologists were interested in finding acceptable matches with a higher yield value. This meant that for the Wave 1 analysis the inputs were now divided into a group of 8 possibly active and 9 inactive variables respectively, as is shown in table 1.

Next, we constructed two 1000 point Latin Hypercube designs: the first over the 8 possibly active variables, and the second over the 9 inactive variables. The first of these was used to construct the Wave 1 emulator (see the next section), and the second was required to assess the uncertainty due to the set of 9 inactive parameters (see section 5.1). Due to runs crashing (for computational reasons), only 993 of the first batch of runs were completed, while all 1000 of the second batch finished successfully. For illustration, Figure 2 shows the main effects plots for the bj outputs at luminosity 17, for the first batch of 993 runs against the 8 possibly active input parameters. Note the clear effect of inputs vhotdisk and alphahot (one of the promoted inputs): these along with epsilonStar, alphareheat and vhotburst were eventually chosen as the active variables for this output (see section 4.3).

To perform a History Match for Galform, we do not need to analyse every output of the model. At each stage, we remove parts of the parameter space if the outputs fail to

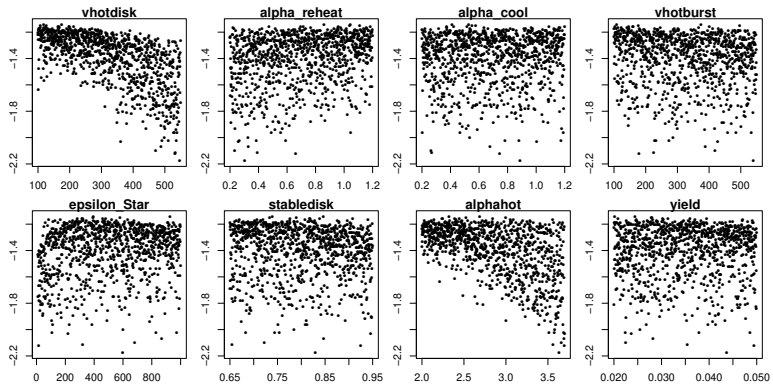

Figure 2: Main effects plots found by plotting the 993 bj outputs (corresponding to luminosity = 17, i.e. the first vertical line in the bj luminosity plot of figure 1) obtained from the Wave 1 runs, against the 993 values of each of the 8 possibly active inputs. Note the clear effect of inputs vhotdisk and alphahot.

match a carefully chosen subset of the observations. At the final stage, we will need to check that our acceptable matches are also in adequate agreement with those features of the output which haven't been used to achieve the history match. Therefore, we chose a subset of 7 of the outputs that are straightforward to emulate at a sufficient accuracy, are informative regarding the inputs in that they can be used to discard large regions of the input space, and that captured the main features of the luminosity function. These are shown as vertical lines in figure 1 along with the full bj and K luminosity outputs from the first batch of 993 runs over the 8 active parameters. The specific luminosity values of each of the 7 outputs are given in the top row of table 2. In later waves more outputs were used.

4.3 The Wave 1 Emulator

We now describe the construction of the 7 univariate emulators corresponding to the 7 luminosity outputs identified in the previous section. As we have many runs, we construct our emulators using a combination of data analytic techniques, checked against physical intuition, and using the Bayes linear update discussed in section 3.3.

The collection of 17 input parameters was split into a group of 8 possibly active parameters (x_B in table 1) and a group of 9 inactive parameters (x_{B^c}). 993 runs for each of the first 40 sub-volumes were completed from a Latin Hypercube design over the group x_B, and these were used to construct the wave 1 emulators. The quantity of interest is the mean output over the first 40 sub-volumes (see section 2.2). Writing

$f_i^{(j)}(x)$ as the ith output from the jth sub-volume, we define:

$$f_i(x) \quad = \quad \frac{1}{40}\sum_{j=1}^{40} f_i^{(j)}(x). \tag{17}$$

We emulate $f_i(x)$ using only the x_B inputs. We add the uncertainty due to sampling only 40 sub-volumes, and the uncertainty due to the remaining 9 parameters x_{B^c} in section 5.1. We use the following form for the emulator of each $f_i(x_B)$ similar to that of equation (6),

$$f_i(x_B) = \sum_j \beta_{ij}\, g_{ij}(x_{[A_i]}) + u_i(x_{[A_i]}) + w_i(x_B), \tag{18}$$

where the active variables $x_{[A_i]}$ are a subset of x_B, and as in section 3.4, u_i and w_i have zero prior expectation, $\mathrm{Cov}(u_i(x_{[A_i]}), u_i(x'_{[A_i]})) = \sigma_{u_i}^2 \exp(-||x_{[A_i]} - x'_{[A_i]}||^2/\theta_i^2)$, while $\mathrm{Var}(w_i(x_B)) = \sigma_{w_i}^2$ and $\mathrm{Cov}(w(x_B), w(x'_B)) = 0$ for $x \neq x'$.

The selection of the set of active variables $x_{[A_i]}$ for each output i is described in detail in appendix C.1, and the results are shown in table 2. It was found that 5 active variables could explain sufficient amounts of the variance of each $f_i(x)$. In appendix C.1 we also describe the remaining technical procedures involved in emulator construction: choosing the functions g_{ij}, assessing the regression coefficients β_{ij} and the Gaussian process parameters σ_{u_i}, σ_{w_i} and θ_i. Table 2 shows the adjusted R^2 corresponding to the polynomial part of the emulator which gives a good indication of the amount of variance of $f_i(x)$ that is explained. Note that at this stage, we only require a relatively simple emulator in order to make an initial reduction of the input space, while leaving the construction of more detailed emulators to subsequent waves of the analysis.

Output	bj 17	bj 21	bj 22.25	K 21	K 22.25	K 24.75	K 25.75
vhotdisk	x	x	x	x	x	x	x
aReheat	x	x	x	x	x	x	x
alphacool		x	x			x	x
vhotburst	x	x	x	x	x	x	x
epsilonStar	x	x		x			
stabledisk			x		x	x	x
alphahot	x			x	x		
yield							
Adj R^2	0.92	0.59	0.70	0.87	0.75	0.72	0.80

Table 2: Wave 1 Active variables and adjusted R^2 for the bj and K luminosity emulators.

Once the above emulator covariance specifications have been made, we can use the 993 wave 1 model runs to update our beliefs, in terms of expectations and variances, about the value of the Galform function $f_i(x)$ at a new input point x using the Bayes linear update equations (9), (10), as is described in section 3.4. This gives the adjusted

expectation and variance $\mathrm{E}_{D_i}[f_i(x)]$ and $\mathrm{Var}_{D_i}[f_i(x)]$ which are used in all subsequent implausibility measures.

Emulator construction should be performed in conjunction with physical considerations of the model in question. The emulator should reproduce, to a reasonable degree of accuracy, the outputs of the model, and should therefore share the physical features of the model. Careful expert assessment regarding the choice of the active variables and the form of the polynomial fit for each output was made to ensure that the emulators were consistent with insight into the physical interpretation of the model. For example, the polynomial for the first bj output has large (negative) contributions from terms involving vhotdisk and alphahot including a strong interaction between them. Both these parameters are used in the SNe feedback module of the Galform model (see table 1 and appendix B), and increasing either will decrease the luminosity function at the faint end. They are known to interact in the model, and therefore the form of the terms in the polynomial that they feature in makes physical sense.

4.4 Emulator Diagnostics

When constructing an emulator, it is essential to perform diagnostics to ascertain whether the emulator is sufficiently accurate for the desired task (Bastos and O'Hagan 2008). At each wave of the analysis, and for each emulator, we performed several types of diagnostic test including: examining the residuals from the polynomial fits; evaluating 200 diagnostic runs of the model (at each wave) and analysing the emulator's predictive diagnostics for these runs; and examining the implausibility measure diagnostics (as shown in figure 5 and discussed in section 6.1). At each wave the emulators were found to be sufficiently accurate to allow substantial reduction of the input space.

5 Quantification of Uncertainty

We now discuss the assessment of all of the remaining uncertainties relevant to linking the Galform Model to the real Universe. These uncertainties can be divided into two classes. The first corresponds to the Model Discrepancy which describes the possible deficiencies of the model, and the second to observational errors.

5.1 Model Discrepancy

As with most complex models of physical systems, modelling assumptions and approximate solutions to known physical equations imply that Galform's output will only be an approximation to what would occur in the real Universe. Further, Galform does not model specific galaxies that exist within our Universe: instead it simulates around a million galaxies from a 'possible' universe that should share statistical properties with our own. These statistical properties will also suffer from approximations inherent in the Galform modelling process. The model discrepancy ϵ_{md} links the system y to the model output evaluated at the actual system properties $f^* = f(x^*)$ via the equation

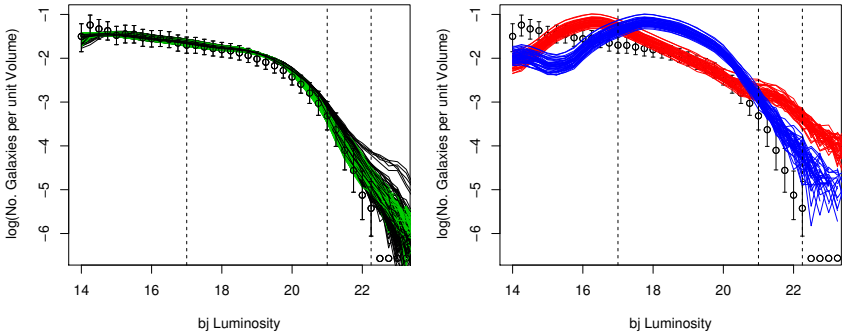

Figure 3: Left panel: the bj luminosity outputs from a sample of 500 runs of the model where only the 9 inactive parameters have been varied. Green and black lines represent the model output when tdisk is off or on respectively. It can be seen that varying the inactive parameters causes a small variance in the model output compared to the 8 active parameters (the effects of which are shown in figure 1). Right panel: The bj luminosity function output of the first 40 sub-volumes of the Dark Matter simulation, for two (blue and red) Wave 1 runs. This source of uncertainty was treated as a model discrepancy term, judged to have constant variance across all runs.

$y = f^* + \epsilon_{md}$. We decompose ϵ_{md} into three uncorrelated contributions:

$$\epsilon_{md} \;=\; \Phi_{IA} \;+\; \Phi_{DM} \;+\; \Phi_E. \tag{19}$$

where Φ_{IA} represents the discrepancy due to the nine inactive parameters, Φ_{DM} is the discrepancy due to the unknown Dark Matter configuration of the real Universe and Φ_E summarises the structural deficiencies of the full Galform model itself. The first two contributions can be assessed using additional runs of the model, while the third requires expert assessment as we describe in the next three sections.

Uncertainty Due to Inactive Variables: Φ_{IA}

As we were unable to run the Galform model while varying all 17 inputs simultaneously, we did not model the effect of the remaining 9 inactive variables in detail (a problem that was resolved before Wave 4 occurred). Therefore, we treat the effect of the 9 variables as initially contributing an extra term Φ_{IA} to the model discrepancy; a term which is dropped in the Wave 4 analysis. For the first three waves, we are essentially running a reduced model (using only 8 inputs), and therefore must use Φ_{IA} to account for the fact that the Galform model output may not match the observed data due to incorrect settings used for the remaining 9 inputs. For waves 1 to 3 these 9 inputs were set to their default values in the Galform code: values that were deemed physically reasonable by the cosmologists.

Quantification of Φ_{IA} was performed as follows. We judged that there was no overall a priori bias due to the extra 9 inputs and set $\mathrm{E}(\Phi_{IA}) = 0$. (These variables have already

been screened for main effects, in section 4.2.) As we had 1000 runs across the 9 inactive variables (with the original 8 inputs set at their default values) over the first 40 sub-volumes, we took the mean of the first 40 sub-volumes for each of these runs, and set the $\mathrm{Var}(\Phi_{IA})$ to be equal to the sample variance of the collection of 1000 means. We are treating as negligible any interactions between the 9 inactive variables and the choice of subvolume, and with the 8 original variables. In figure 3 (left panel), we show the first 500 out of the set of 1000 runs performed across these 9 inputs, with the 8 active variables set at the default value (which corresponds to the cosmologists' best match: a run which is borderline acceptable according to our matching criteria). Figure 4 (bottom panel) compares the standard deviation of all uncertainties discussed in this section, at every point on the bj luminosity function graph given in figure 1, and shows $\sqrt{\mathrm{Var}(\Phi_{IA})}$ as a light blue line (the K luminosity function has similar uncertainties which we do not show here). The three bj points that were chosen for emulation are given by the black dashed lines.

Note the similarity between the nugget term, $w_i(x_B)$, in equation (18), and the model discrepancy term given by Φ_{IA}. Both are treated as independent of x, have expectation zero and constant variance. This greatly simplifies subsequent calculations and allows a straightforward reduction of the input space in the first wave. In subsequent waves, we model these effects in more detail.

Dark Matter Uncertainty: Φ_{DM}

We now assess the uncertainty due to the unknown Dark Matter configuration of the real Universe. The Millennium Simulation provides 512 possible forcing functions, each representing a possible configuration of dark matter to be used by the Galform model. We perform runs using only the first 40 sub-volumes out of the 512, to facilitate comparison with previous studies. While using more sub-volumes would be more accurate, the extra run time would allow fewer model evaluations. We have therefore emulated the mean of the function output over these 40 sub-volumes given by $f_i(x)$ (equation 17). Figure 3 (right panel) shows the luminosity output from the first 40 sub-volumes for two runs of the model (given by the collection of red and blue lines).

The processing of the observational data and associated errors has effectively elevated the data to represent the density of galaxies as measured over a much larger volume of the Universe than is defined by the 512 sub-volumes of the Galform model. We take this volume to be effectively infinite and represent the uncertainty due to analysing the mean of only 40 sub-volumes as the model discrepancy term Φ_{DM}. We assessed Φ_{DM} by first judging that there was no overall bias a priori and setting $\mathrm{E}(\Phi_{DM}) = 0$. We then used the outputs $f_i^{(j)}(x)$ for each of the 40 sub-volumes for the 993 runs performed in Wave 1 to derive an approximate value for the variance of Φ_{DM} as follows. For each of the 993 runs we calculated the standard error of the mean output over 40 sub-volumes, and averaged this over all 993 runs. This was done for each of the 7 outputs. While this is a relatively straightforward assessment, given the important simplifying judgement that Φ_{DM} is independent of x, it was felt that this captured the main source of uncertainty without going into detail that would be unwarranted at this stage of the analysis. A

more careful treatment would model the outputs of the sub-volumes individually, as has been performed in House et al. (2009), using exchangeable computer model techniques. To check that the first 40 sub-volumes are representative of the full set of 512, we ran a design of 100 runs at the same x input locations as the first 100 runs of the original Wave 1 design, but choosing 40 random sub-volumes out of the set of 512 instead of the first 40. We found that the variance across the random 40 sub-volumes was not significantly different from the original 40 and so did not alter the assessment for the $\mathrm{Var}(\Phi_{DM})$ described above. The size of Φ_{DM} for all bj luminosity outputs is shown as the dark blue line in figure 4 (bottom panel). Note that the relative size of Φ_{DM} is small compared to other sources of uncertainty, so that it was considered unnecessary to model its effect in more detail at this stage.

Full Galform Model Discrepancy: Φ_E

The model discrepancy term Φ_E is a 7 vector, the components of which need to be assessed from expert judgements. In the first wave of our analysis we perform only a univariate analysis of each of the 7 outputs, hence we required a univariate assessment of each component. In waves 3 and 4, multivariate analyses were performed and hence a more detailed assessment was required. We describe here the full multivariate elicitation.

As we are employing a Bayes Linear analysis, we only require specification of expectations, variances and covariances over all quantities of interest. Subjective assessment of $\mathrm{E}(\Phi_E)$ and $\mathrm{Var}(\Phi_E)$ is still a difficult task. Expert assessment for beliefs regarding deficiencies of the model was that discrepancy judgements were symmetric in that $\mathrm{E}(\Phi_E) = 0$. For the multivariate case, assessment of $\mathrm{Var}(\Phi_E)$ was required which is now a 7x7 matrix. The structure of this matrix came from Richard's opinion as to the deficiencies of the model as follows.

For Galform, there are two major physical defects that can be identified. The first is the possibility that the model has too much (or too little) mass in the simulated universe, possibly due to incorrect choices for the cosmological parameters used in the Millennium simulation (see section 2.2). This would lead to the 7 luminosity outputs all being too high (or too low), and would lead to positive correlation between all outputs in the $\mathrm{Var}(\Phi_E)$ matrix. The second possible defect is that the model incorrectly calculates the colour of the galaxies, due to inaccurate modelling of stellar populations or dust. This would lead to an apparent increase/decrease in the number of red galaxies and decrease/increase in the number of blue galaxies. This is represented as contributing a smaller negative correlation between the bj and K luminosity outputs. To respect the symmetries of these possible defects, the multivariate Model Discrepancy was parame-

terised in the following (3+4)x(3+4) block form:

$$\mathrm{Var}(\Phi_E) \;=\; a^2 \begin{pmatrix} 1 & b & b & c & c & c & c \\ b & 1 & b & c & c & c & c \\ b & b & 1 & c & c & c & c \\ c & c & c & 1 & b & b & b \\ c & c & c & b & 1 & b & b \\ c & c & c & b & b & 1 & b \\ c & c & c & b & b & b & 1 \end{pmatrix} \tag{20}$$

where now a^2 is the univariate variance of the model discrepancy; b is the correlation between outputs of the same luminosity graph (either bj or K luminosity) and c is the cross graph correlation. While Richard was satisfied with the form of the parameterisation of $\mathrm{Var}(\Phi_E)$ as given by equation (20), he was cautious about specifying values for a, b and c. He was, however, willing to provide the following ranges:

$$3.76 \times 10^{-2} < a < 7.52 \times 10^{-2}, \;\; 0.4 < b < 0.8, \;\; 0.2 < c < b. \tag{21}$$

This assessment involved examining the difference between Galform and a competing model of similar complexity, consideration of the above possible physical defects to the model, and from his previous years of experience coding and running such galaxy formation models. The maximum value of $a = 7.52 \times 10^{-2}$ is shown as the black line in figure 4 (bottom panel), where $\sqrt{\mathrm{Var}(\epsilon_{md:i})} = a$, for each i.

After the initial assessment we constructed an elicitation tool in order for Richard to confirm that his specification agreed with his intuition regarding the outputs of the luminosity function. A picture of this elicitation tool is shown in figure 4 (top panel), and it possesses the following features. The top two panels of the tool show the bj and K luminosity functions, with observational data points in black, error bars representing all uncertainties, dotted lines giving the 11 outputs of interest (additional outputs were used in later waves), and constructed (or fictitious) luminosity model output given by the red lines. This elicitation tool allows the user to experiment with various possible luminosity functions and see the corresponding values for the two implausibility functions $I_M(x)$ and $I(x)$ (see section 3.5 for definitions of these measures). Most importantly, the values of the multivariate model discrepancy parameters a, b and c can be adjusted. This allowed Richard to experiment with different specifications of a, b and c and to see the response of the implausibility measures. This is useful for the expert to get a feel for the behaviour of a multivariate implausibility measure, understand the ramifications of the structure of $\mathrm{Var}(\Phi_E)$ and also to check that intuitively acceptable runs would not be ruled out by the current specification.

Obviously it is possible to build in far more structure into $\mathrm{Var}(\Phi_E)$ if required. The aim here was to account for the main sources of model discrepancy, while maintaining a relatively simple structure of the $\mathrm{Var}(\Phi_E)$, as the more detailed the structure, the more difficult eliciting expert information becomes. As we have ranges for a, b and c, we will incorporate this into our analysis by performing a sensitivity analysis, and rule out parts of the input space only if they fail certain implausibility cutoffs for all values of a, b and c within the above ranges.

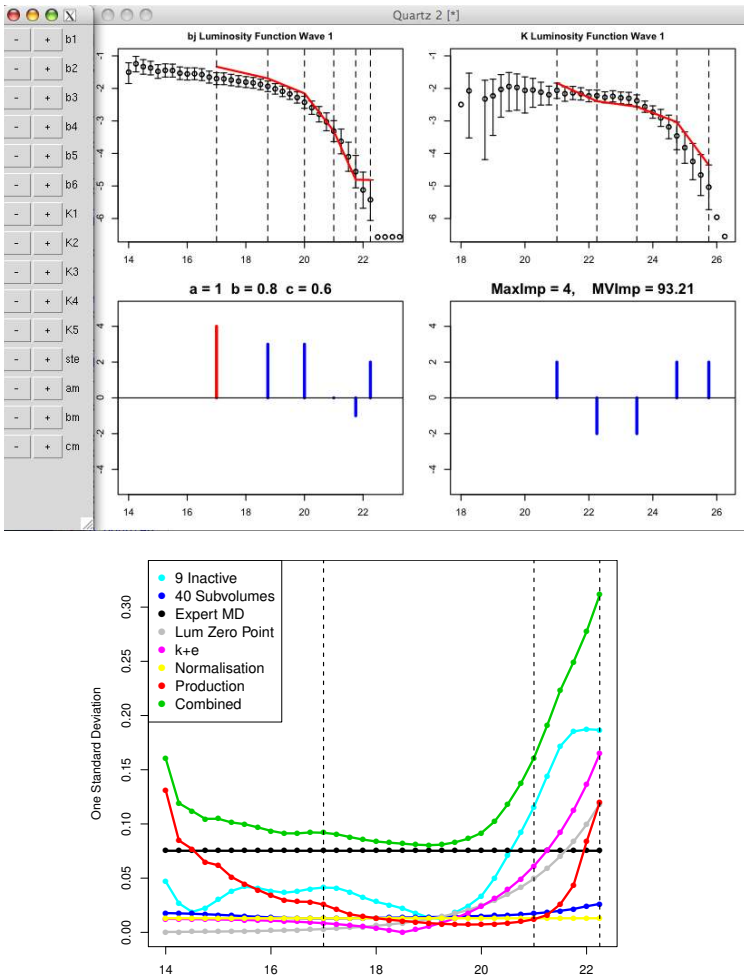

Figure 4: Top panel: the Elicitation Tool used to confirm the multivariate model discrepancy assessment represented by equations (20) and (21). It allows the expert to construct and adjust fictitious luminosity functions, and to explore the response of the implausibility measures to changes in a, b and c (see section 7.3). Bottom panel: the sd of each contribution from the various sources of uncertainty for the full range of the bj luminosity function (the x-axis is the same as figure 1). The vertical lines represent the three bj outputs chosen for emulation in Wave 1. The green line represents the total uncertainty due to all contributions, and it is this value that is used in all bj luminosity plots such as figure 1. The K luminosity results are similar.

5.2 Observational Errors

The generation of the observational data shown as the black points in figure 1, is an extremely intricate task. It involves data from several sky surveys, which is processed using

both information from various simulations and additional theoretical and experimental knowledge related to the evolution of the Universe. Due to this, the observational errors ϵ_{obs} defined in equation (1) are complex. Due to space limitations we only summarise the four contributions to $\text{Var}(\epsilon_{obs})$ here; see Cole et al. (2001) for more details.

The Luminosity Zero Point Error - this is derived from the difficulty of defining the Luminosity Zero Point: that is the point on the x-axis of the luminosity graph (see figure 1) corresponding to a galaxy of 'zero' brightness. This results in a correlated error on every output point (grey line in figure 4 (bottom panel)).

The k+e error - a perfectly correlated error on all output points due to necessary corrections for two effects (i) Galaxies being so far away it takes light billions of years to reach us and (ii) Galaxies moving away from us so quickly their light is redshifted (purple line in figure 4 (bottom panel)).

The Normalisation Error - The data on galaxies comes from measurements made in our local vicinity and it is possible that we live in a relatively under/over populated part of the Universe. This error attempts to account for this using theoretical knowledge about variation in mass density in the Universe on large scales (yellow line in figure 4 (bottom panel)).

Galaxy Production Error - Bright/faint galaxies can be measured up to relatively large/short distances from our Milky Way. This error represents the uncertainty due to this effect and uses assumptions as to the shape of the mean luminosity function (red line in figure 4 (bottom panel)).

It is clear that significant contributions to the observational errors come from uncertainties related to the processing of the data (i.e. the $k + e$, Normalisation and Production Errors). These are distinct from measurement errors and are derived from complex theoretical and modelling uncertainties, and hence could be referred to as model discrepancy terms as opposed to observational errors. However, the calculations involved in determining these errors are intricate and rely upon specialist knowledge of Astronomy. Although it would be desirable to disentangle some of these errors, due to time constraints it was felt that this was impractical at the current stage.

6 First Wave History Match

6.1 History Matching via Implausibility

History Matching is the process of identifying $\mathcal{X}^*$ by iteratively discarding values of x, by the application of cutoffs to the Implausibility Measures. In Wave 1, we used the measures $I_{2M}(x)$ and $I_{3M}(x)$ to discard values of x that do not satisfy both:

$$I_{2M}(x) \; < \; I_{cut2} \quad \text{and} \quad I_{3M}(x) \; < \; I_{cut3}, \tag{22}$$

where I_{cut2} and I_{cut3} are the corresponding implausibility cutoffs. The choices made for the individual cutoffs come from a combination of examination of diagnostics (such as shown in figure 5), consideration of the amount of space cut out, and unimodality

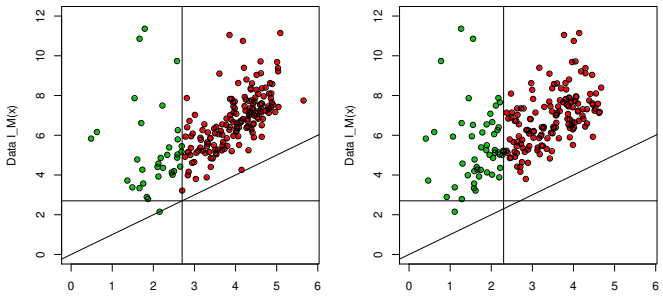

Figure 5: Implausibility diagnostics for the Wave 1 univariate emulators. Plots show the maximum data implausibility $I_M^{data}(x)$ calculated using actual runs, against the implausibility measures $I_{2M}(x)$ (left panel) and $I_{3M}(x)$ (right panel) which are calculated using the emulator. The vertical lines show the cutoffs imposed at this Wave, with the red points belonging to parts of the input space deemed implausible.

arguments based on Pukelsheim's 3-sigma rule as discussed in section 3.5. While the unimodal argument suggests using cutoffs of 3 or higher (depending on the correlation between outputs), consideration of figure 5 shows that this might be unnecessarily conservative. In response to this we choose cutoffs of $I_{cut2} = 2.7$ and $I_{cut3} = 2.3$ (shown as vertical lines in figure 5), recognising the fact that we want to balance a conservative cutoff with the amount of space that can be removed at Wave 1. These cutoffs resulted in approximately 85.1 percent of the input space being ruled out due to the Wave 1 analysis. Note that, as discussed at the end of section 5.1 and also in section 7.3, we effectively perform a sensitivity analysis on Φ_E by only ruling out inputs that do not satisfy the cutoffs *for all values of* a, b *and* c within the ranges given in equation (21). For the univariate cases discussed here, this is equivalent to setting a to its maximum value as $I_{(i)}(x)$ is monotonically decreasing with increasing a.

Figure 5 shows diagnostic plots regarding the choice of cutoffs I_{cut2} and I_{cut3}. It shows the maximum data implausibility $I_M^{data}(x)$ (that is the implausibility evaluated at a known run, given by equation (11)) across the 7 outputs for a latin hypercube of 200 diagnostic runs (y-axis), against $I_{2M}(x)$ (left panel) and $I_{3M}(x)$ (right panel). The vertical lines are the cutoffs that will be imposed, implying that the red points would be discarded. Note that most points are some distance above the diagonal $y = x$ line, suggesting that $I_M^{data}(x)$ will generally be higher than $I_{2M}(x)$ and $I_{3M}(x)$ as expected. Also note that the discarded points do indeed have high $I_M^{data}(x)$ (significantly higher than the 2.7 cutoff shown as a horizontal line), and hence suggest the space cutout in Wave 1 does not contain any inputs of interest. We test the sensitivity of such diagnostics and of the fraction of space cut out, to different values of cutoffs, before definite choices are made.

In figure 6 we show various 2-dimensional projections (top 3 panels) of values of the Implausibility Measures, with red areas representing high implausibility and green areas low, which were constructed as follows. For each plot we evaluated the emulator at a set

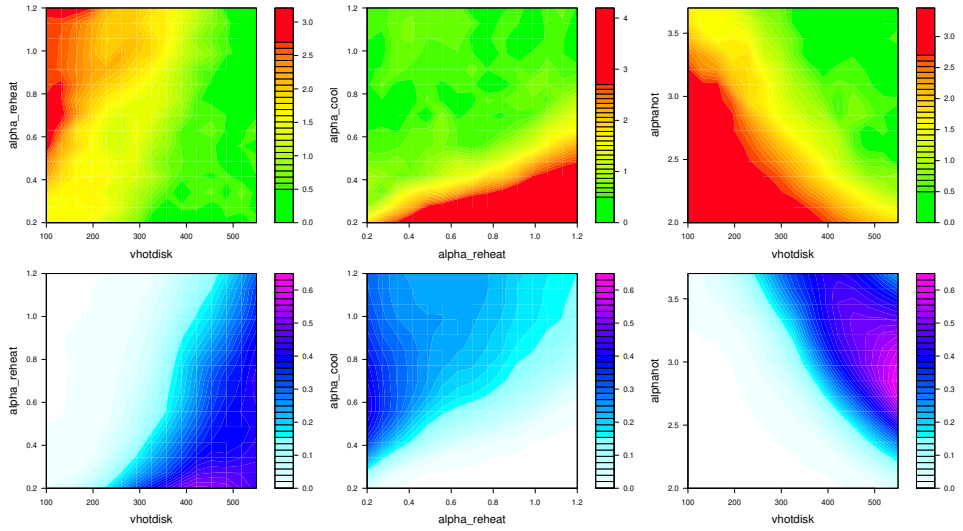

Figure 6: The top three panels give Wave 1 minimised implausibility projection plots: the red region indicates high implausibility for all values of the remaining inputs: here input points will be discarded. The bottom three panels give the 'optical depth' plots: these show the fraction of the hidden 5 dimensional volume (spanned by the remaining active variables) that satisfies the implausibility cutoff, at that grid-point.

of inputs specifically designed to produce a 2-dimensional projection in the appropriate input plane. For example, in the top left panel the projection is in the vhotdisk - alphareheat plane, and the emulator was evaluated on a (2d grid)x(5d latin hypercube) design, where the 2d grid was over the vhotdisk - alphareheat plane (and of size 15^2) while the latin hypercube was defined over the remaining 5 active inputs at Wave 1 (and was of size 1500). For each point on the grid, we then minimised the implausibility over the corresponding 1500 points at that grid location, the results of which provide the plots shown. This allows the following interpretation: a red area in one of these implausibility projection plots implies that even given all relevant uncertainties, and all possible choices for the other input parameters, it is highly unlikely that an acceptable match will be found at this point in the vhotdisk - alphareheat plane (for example). Such plots present serious computational complications as a large number of emulator evaluations are required for each projection. To generate these plots we have used novel Bayes Linear calculations that exploit both the cross-product symmetry of the *emulator design*, and similar symmetries that occur in the emulator update equations (9) and (10) due to the covariance structure of equation (24). These calculations greatly improve efficiency, and we will report on these techniques in detail elsewhere.

The bottom 3 panels of figure 6 show depth projection plots: these are constructed by calculating at each grid point, the fraction of the corresponding 1500 points of the latin hypercube that survive the implausibility cutoffs, given by equation (22). This gives

information as to the 'optical depth' of the 7 dimensional non-implausible volume when observed in a direction perpendicular to the vhotdisk - alphareheat plane (for example). They provide complimentary information to the implausibility projections. Consider the middle top and bottom panels of figure 6, where the implausibility projection (top panel) shows that non-implausible choices of alphareheat and alphacool exist over much of the alphareheat-alphacool plane. The depth plot demonstrates that the majority of the non-implausible volume is found at low values of alphareheat. These images give physical insights into the nature of the Galform model: in the top right panel of figure 6 we see that simultaneously low values of both vhotdisk and alphahot are ruled out, and that high values of both these parameters are possibly preferred. These parameters are involved in the same Galform module: that of Feedback from Supernovae (see equation (23) and appendix B), and increasing their size should increase the amount of material expelled from certain galaxies as opposed to being used to form stars. This will reduce the luminosity function at the faint end, and, as most of the Wave 1 runs are higher than the observed data, it makes physical sense that parameter choices that lower the luminosity function will be preferred. These physical features are also seen in the polynomial terms for the outputs bj 17 and K 21 (which are at the faint end of the luminosity function), specifically we find large and negative coefficients for the vhotdisk, alphahot and their interaction terms. The Wave 1 emulators are quite approximate, so there is a limit as to the physical insight they, and the corresponding implausibility measures, can provide.

Equation (22) defines a volume of input space $\mathcal{X}_1$ that we refer to as non-implausible after Wave 1, projections of which are shown in figure 5. We now *refocus* by running the Galform model within this volume, and repeat the above process of emulation, constructing implausibility measures and imposing cutoffs. We have gone through four iterations or waves, as described below.

7 Analysis of Waves 2 - 4

7.1 Wave 2 to 4: Design, New Outputs and Emulation

We apply the refocussing technique iteratively, and here we describe the designs and emulators used in waves 2 to 4. The design for the set of Wave 2 model evaluations was derived as follows. We first constructed a large maximin Latin Hypercube design containing 9500 points defined over the 8 dimensional input space corresponding to the 8 input variables explored in Wave 1. We then used the Wave 1 emulator and Implausibility measures to evaluate the implausibility of each proposed point in the design. Any points that did not satisfy the implausibility cutoffs, as given by equation (22), were discarded from further analysis. This left a design of 1414 points which were then evaluated using the Galform model, the results of which were used to construct the Wave 2 emulator. The Wave 3 design of 1620 points was constructed in a similar manner. Between Waves 3 and 4, the problems preventing simultaneous varying of all 17 parameters in the Galform model were resolved. Hence, the Wave 4 design came from a large latin hypercube defined over the full 17 dimensional input space. Again, only

Wave	Runs	Act	I_M	I_{2M}	I_{3M}	I_{MV}	% Space
1	993	5	-	2.7	2.3	-	14.9 %
2	1414	8	-	2.7	2.3	-	5.9 %
3	1620	8	-	2.7	2.3	26.75	1.6 %
4	2011	10	3.2	2.7	2.3	26.75	0.26 %

Table 3: The fraction of parameter space deemed non-implausible after each wave of emulation. Column 1: the wave; Column 2: the number of model runs used to construct the emulator; Column 3: the number of Active Variables; Column 4-7: the implausibility thresholds; Column 8: the fraction of the parameter space deemed non-implausible.

points that satisfied all of the previous 3 wave's implausibility cutoffs remained in the design, leaving a total of 2011 points. The number of design points was deliberately increased at each wave in anticipation of fitting more complex polynomials.

As the input space has been reduced after the Wave 1 analysis, it became easier to emulate model outputs. Therefore more outputs become informative regarding the input space, and warrant inclusion in the analysis. Consideration of the 1414 Wave 2 runs led to 4 additional outputs being included, the bj outputs with luminosity 18.75, 20 and 21.75, and the K output with luminosity 23.5. These are shown in figures 12 and 13 along with the original 7 outputs, as the dotted vertical lines. For each wave, emulation proceeded in a similar manner to Wave 1, with univariate emulators being used in all waves, and multivariate emulators used in waves 3 and 4. The details of the construction of these emulators are given in appendices C.2 and C.3. Table 3 summarises the number of runs used at each wave, number of active variables required and space remaining. At each wave, cluster analysis was performed to check that the non-implausible volume was simply connected (which was found to be the case), as separate emulators would have been required for unconnected volumes.

7.2 Comparing Emulators

At each wave, emulator accuracy increases. It is instructive to compare the emulators, to understand which features lead to this improvement. As the Wave 4 emulator involves all 17 input parameters, we leave discussion of it until section 8.1.

Figure 7 (left panel) shows the estimated value of the residual standard deviation σ_{u_i} for each of the first three waves, for all 11 emulated outputs (for completeness we show all 11 outputs for Wave 1 even though 4 of these were not considered at that stage). There are significant drops in σ_{u_i} from Wave 1 to 2 across all outputs, with even more substantial drops from Wave 2 to Wave 3, especially for the K luminosity outputs (outputs 7 to 11). The right panel of figure 7 shows the adjusted R^2 for each of the 11 emulators, for each of the 3 waves. It shows the improvement in percentage of output variance explained in Waves 2 and 3 compared to that of Wave 1. Note that although the Wave 3 adjusted R^2 is sometimes below that of Wave 2, this is to be expected:

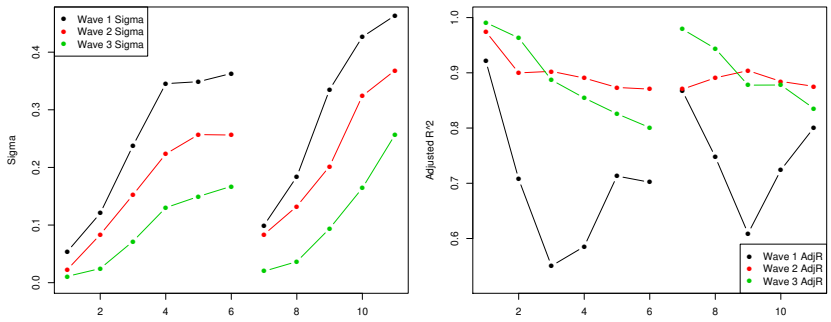

Figure 7: Plots showing the residual standard deviation σ for waves 1 to 3 (left panel) and the Adjusted R^2 for wave 1 to 3 (right panel). In each panel, the first 6 connected points correspond to the bj outputs chosen for emulation, the later 5 connected points are the K outputs (shown as vertical lines in figures 12 and 13 respectively).

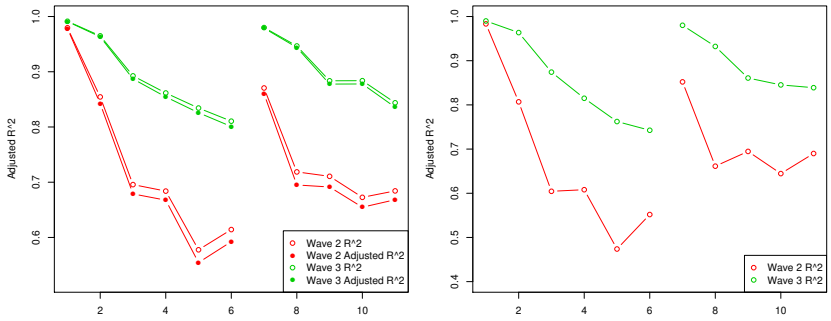

Figure 8: Left panel gives a plot showing the R^2 (open points) and adjusted R^2 (solid points) of the Wave 2 polynomial when used to predict the outputs of the Wave 3 runs (in red). Also shown are the corresponding Wave 3 polynomial R^2 (open points) and adjusted R^2 (solid points) in green. Note the large difference between red and green points. Right panel: shows the fairer comparison of the R^2 of the Wave 2 and 3 polynomials when used to predict 204 Wave 3 diagnostic runs. First 6 connected points: the bj outputs, last five: the K outputs, as in figure 7.

as the variance of the Wave 3 run outputs is less than that of the Wave 2 runs, the Wave 3 emulators may not be able to explain more of this variance than their Wave 2 counterparts, even though they are more accurate.

Further confirmation of the difference between the Wave 2 and 3 polynomials is given by figure 8. As the Wave 2 and 3 polynomials have been fitted using highly non-orthogonal designs of input points, it is not trivial to compare their polynomial coefficients directly, in order to determine any differences between them. In figure 8 (left panel) we show the R^2 and adjusted R^2 of the Wave 2 polynomial calculated using the Wave 3 runs (in red). Also shown are the R^2 and adjusted R^2 of the Wave 3 polynomial calculated with the same Wave 3 runs (in green). Note the dramatic

difference in variance explained between the red and green points. This demonstrates that the two sets of polynomials are substantially different. While this comparison is not strictly fair (as the Wave 3 points were used to fit the Wave 3 polynomial), equivalent polynomials would be expected to have much smaller differences in their R^2 values. To highlight this point, figure 8 (right panel) shows the R^2 of the Wave 2 and Wave 3 polynomials calculated using a set of 204 Wave 3 diagnostic runs. Again a clear difference between the explanatory power of the two polynomials can be seen. This suggests that the emulators are picking up new features of the model at each wave through improved polynomial fits: a natural feature as we build more structure into the mean function, as opposed to the Gaussian process residual.

7.3 Implausibility Measures and Space Reduction

Table 3 summarises which of the four implausibility measures $I_M(x)$, $I_{2M}(x)$, $I_{3M}(x)$ and $I(x)$ were used in each of the four Waves, along with the implausibility cutoffs that were imposed. Note that the multivariate cutoff I_{MV}, employed at Wave 3, was chosen to be equal to 26.75, the critical value of 0.995 from a chi squared distribution with 11 degrees of freedom. This cutoff was employed in a conservative manner as follows. The expert asserted ranges on a, b and c which parameterise $\mathrm{Var}(\Phi_E)$ ((20),(21)). Therefore, inputs x were only discarded due to the multivariate measure $I(x)$ if $I(x) > I_{MV}$ for all values of a, b and c within their specified ranges (see Vernon and Goldstein (2009)).

Figure 9 shows the progression of implausibility and optical depth plots, in the vhotdisk and alphacool plane, for Waves 1 to 3. Note that the size of the non-implausible region decreases with each wave as expected, occupying a volume of 14.9%, 5.9% and 1.6% respectively. Even though the non-implausible volume occupies a small part of the input space, it still covers a large part of the two dimensional projection.

8 Results of Wave 4 and 5

8.1 Wave 4

The Wave 4 emulator gives an accurate description of the non-implausible region of input parameter space $\mathcal{X}^*$. Visualising this region is a difficult task, as it is a complicated object in a ten-dimensional space. We here confine our analysis to useful two dimensional projections of the space. Figure 10 shows the minimised Implausibility projections (below the diagonal) and optical depth plots (above the diagonal) corresponding to all possible pairs of active variables. The plots above the diagonal have been transposed to have the same orientation as those below the diagonal for ease of comparison. Figure 10 highlights many features of the Galform model, which are of great interest to the cosmologists. It suggests that acceptable fits can be found over large ranges of the input parameters. It also demonstrates clear relationships between certain parameters, for example, the positive correlation between vhotdisk and alphareheat: if one input is increased, then the second should be increased to compensate. This make physical

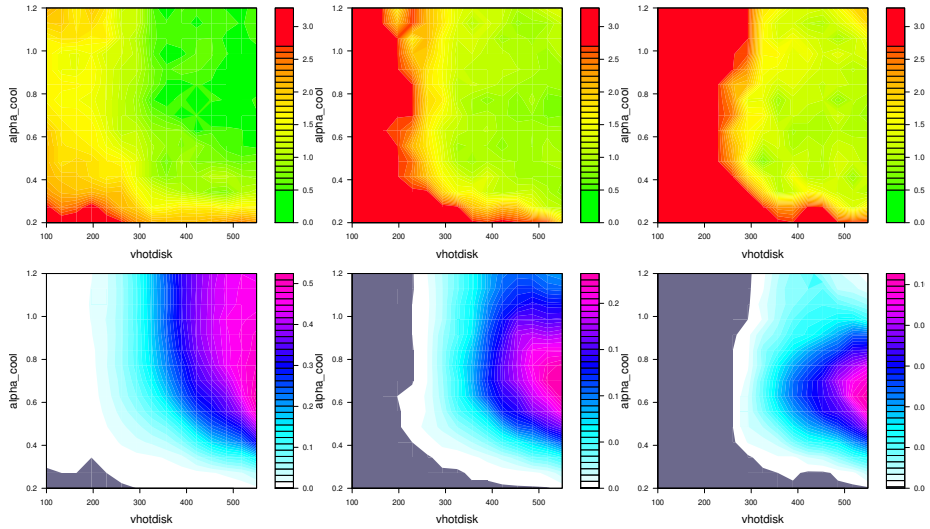

Figure 9: The top three panels give Wave 1, 2 and 3 implausibility projection plots: the red region indicates high implausibility where input points will be discarded. Note that the yellow and green regions occupy only 15%, 5.9% and 1.6% of the input space respectively (the non-implausible region), even though they take up much larger areas of the 2-dimensional projection. The bottom three panels give the depth plots, showing the fraction of the hidden 6-dimensional volume that satisfies the implausibility cutoff, at that grid-point.

sense as both these parameters are involved with feedback from supernovae: vhotdisk is related to the gas blown out of a galaxy due to supernovae while alphareheat regulates the time taken for this gas to return. Similarly, there exist a strong negative correlation between vhotdisk and alphahot: another input related to supernovae feedback. Figure 10 also shows which parameters influence the luminosity functions, and are therefore constrained, and which parameters do not. Inputs related to the Reionisation and Galaxy Mergers modules of the Galform function (see table 1 and appendix B) are all inactive save tau0mrg ($f_{\rm df}$), which only has a subtle impact. Therefore the physical processes represented by these modules can be concluded to have little impact on the luminosity function. There are many more physical interpretations that can be obtained from this analysis. For example, by applying principal component analysis to a set of points belonging to the non-implausible region, several approximate linear relationships between groups of variables can be obtained (see Bower et al. (2010)).

8.2 Wave 5

After the Wave 4 analysis, we ran a final batch of 2000 model evaluations within the non-implausible region defined by the Wave 4 emulator. We refer to these as Wave 5

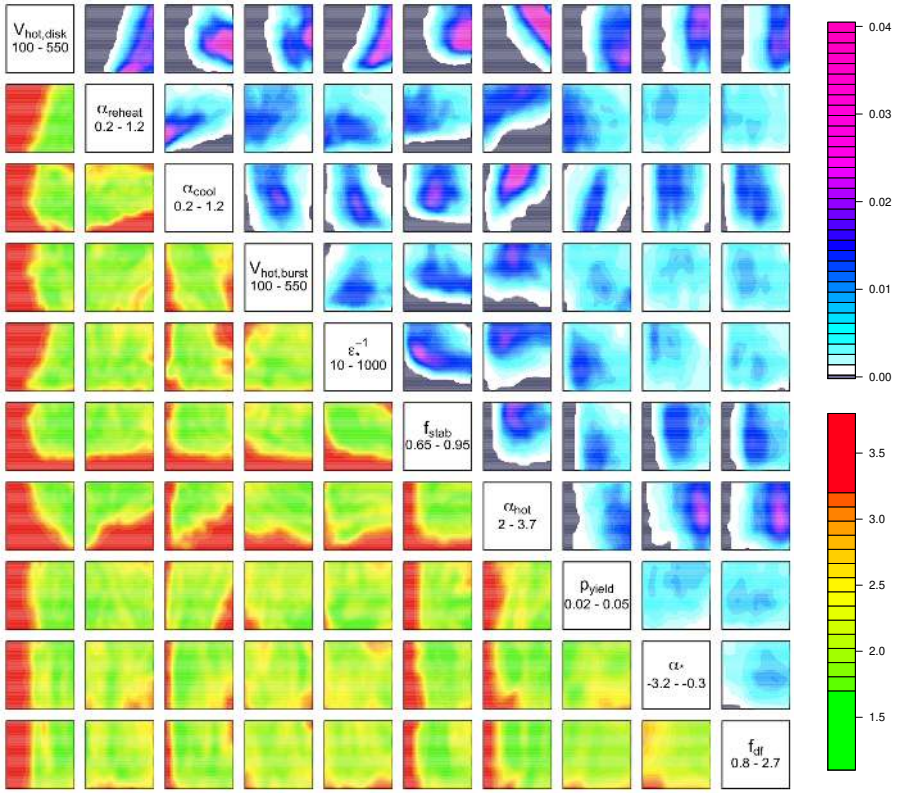

Figure 10: All Wave 4 Implausibility (below diagonal) and Optical Depth (above diagonal) projections. Compare the Implausibility plots with the Wave 5 runs of figure 11.

runs. These runs were evaluated to see if we could determine whether the set $\mathcal{X}^*$ was non-empty. and if so to check that a significant volume of the non-implausible region did indeed correspond to acceptable runs (and therefore that another wave of analysis is not required), and to generate a large set of realised acceptable runs for the cosmologists to use to perform provisional explorations of other output data sets.

Figure 11 shows the two-dimensional projections of these Wave 5 runs, coloured using the data implausibility (that is the implausibility without any emulator variance). The colour scale is the same as that of figure 10 to allow direct comparison. It can be seen that we do indeed find a large number of acceptable runs: 306 of the 2000 Wave 5 runs satisfied the implausibility cutoffs, with approximately 800 more runs within 10 percent of the cutoff boundary. This is expected as the surface area of a complex 10-dimensional object can be large compared to its volume. The acceptable runs do span a large range in several of the inputs, as was suggested by the Wave 4 analysis: a fact that was a surprise to the cosmologists. In general the Wave 5 runs are in good agreement with the Wave 4 analysis, suggesting that the Wave 4 emulator is of sufficient

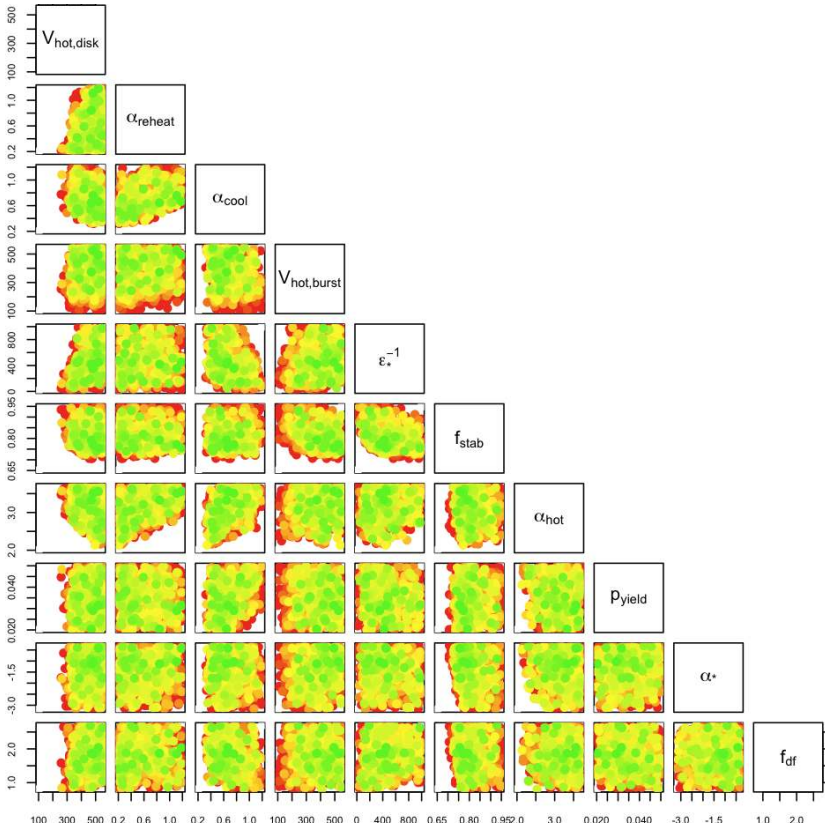

Figure 11: The Wave 5 runs coloured by the data implausibility, consistent with fig 10.

accuracy. For this reason, and due to the large number of acceptable runs obtained, we concluded that another wave of analysis was unnecessary. The acceptable runs were used to perform provisional explorations of additional outputs of the Galform model, as described in Bower et al. (2010).

To illustrate the improvement in the model runs from Wave 1 to Wave 5, figures 12 and 13 show the first 500 model runs bj and K outputs from Waves 1, 2, 3 and the 'good' runs from Wave 5, defined as those that satisfy $I_M(x) < 2.5$. It can be seen that a large number of acceptable runs have been found, which are acceptable across all outputs of interest, not just the 11 used for the emulation process.

9 Conclusion

In this Case Study we have presented the results of an uncertainty analysis of the galaxy formation model known as Galform. The main aim was to identify the set of inputs that would give rise to an acceptable match between model output and observed data,

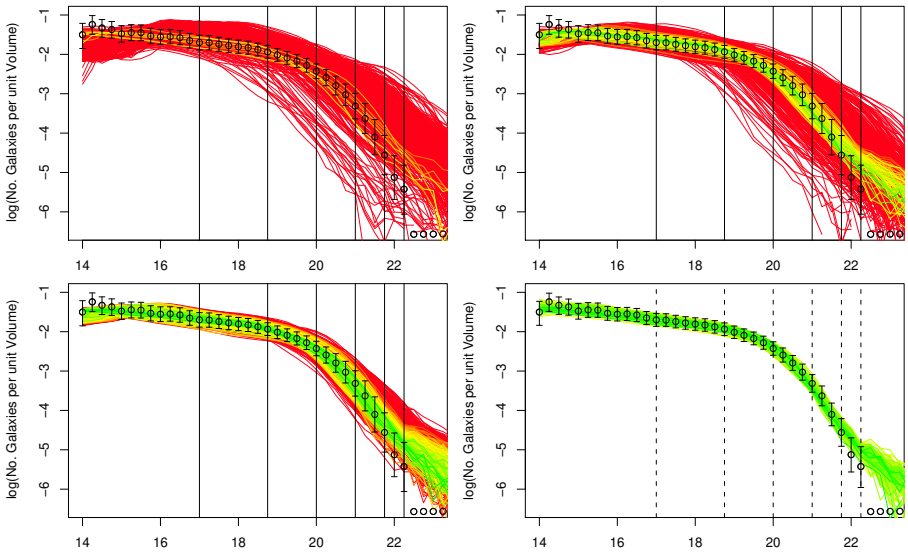

Figure 12: The bj Luminosity function output for the first 500 runs of Waves 1, 2 and 3 (top left, top right and bottom left panels respectively). The colours represent the maximum implausibility $I_M(x)$ and are consistent with the colour scale of figures 10 and 11. Bottom right panel: the Wave 5 runs that satisfy $I_M(x) < 2.5$. (Note the tighter error bars compared to previous waves as Φ_{IA} has been dropped).

taking into account all of the major uncertainties present in such a situation.

This analysis can be seen as a demonstration of the power of the iterative refocussing technique in addressing a difficult and important problem: difficult in the sense that Galform is a complex model with a significant run time, and with a large number of active parameters many of which exhibit intricate interactions; important in that Galform is a state-of-the-art model, and that the results we present provide insight into the physics of galaxy formation for the cosmology community. At each iteration, improved fits for the emulators are obtained, and new features of the model are seen (section 7.2). This iterative strategy leads to a collection of emulators that are increasingly accurate over regions of the input space of increasing interest. It is hard to see how such an accurate description of the non-implausible region of input space could be obtained in one step, without requiring an infeasibly large number of model evaluations. As the non-implausible region is so small (less that 0.26% of the initial space), it is clearly beneficial to perform a History Match before attempting any form of fully Bayesian calibration.

What improvements could have been made to this project? We have had the benefit of substantial computational resources, courtesy of the Galform group. This has allowed relatively large numbers of runs to be performed at each wave of the analysis, when it may have been possible to obtain broadly similar results using fewer evaluations. Also,

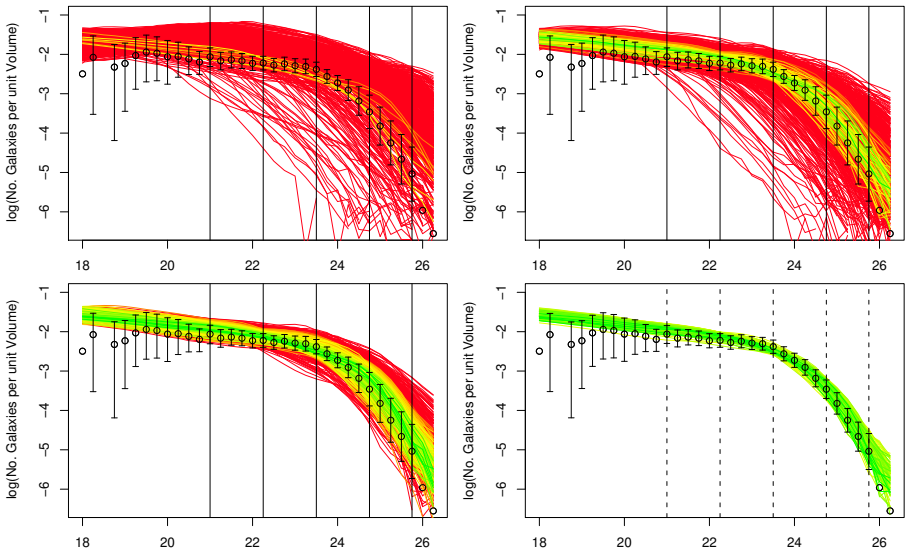

Figure 13: The K luminosity version of figure 12. Note the disparity at luminosity $\leq$ 19 for the Wave 5 runs (bottom right panel) is due to the limited resolution of the Dark Matter simulation (see Bower et al. (2006)) and so is not considered to be of interest.

certain simplifying assumptions used when assessing the Model Discrepancy could have been dropped. For example, the assumption that the effect of the Dark Matter forcing function Φ_{DM} was independent of x, has been addressed in House et al. (2009), where Galform models with different Dark Matter configurations are treated as exchangeable computer models. This is a particular aspect of a more general treatment of model discrepancy known as Reification (Goldstein and Rougier (2009)).

The identification of the non-implausible region shown in figure 10 provides several immediate physical insights into the Galform model, e.g. the relations between certain inputs, the ranges of feasible values for the inputs, as well as identifying which inputs are not restricted by the luminosity function, all of which are of significant scientific interest. However, there may be several physical features that are hard to obtain from simple 2- or even 3-dimensional projections, or from linear analyses such as PCA (Bower et al. 2010). Visualising the complexities of the full 10-dimensional volume efficiently is a difficult task (even using packages such as Ggobi (www.ggobi.org)), but must be addressed in order to extract the full information provided by the emulators. This is made even more difficult by the fact that although the emulators are very fast to evaluate, they are still not fast enough to completely cover a (possibly complex) 10-dimensional object. We have developed efficient emulator designs and calculation routines for high-dimensional visualisation purposes and will report on these elsewhere.

The set of Wave 5 evaluations provided a large number of realised acceptable runs for use by the cosmologists in provisionally exploring further Galform outputs. Several

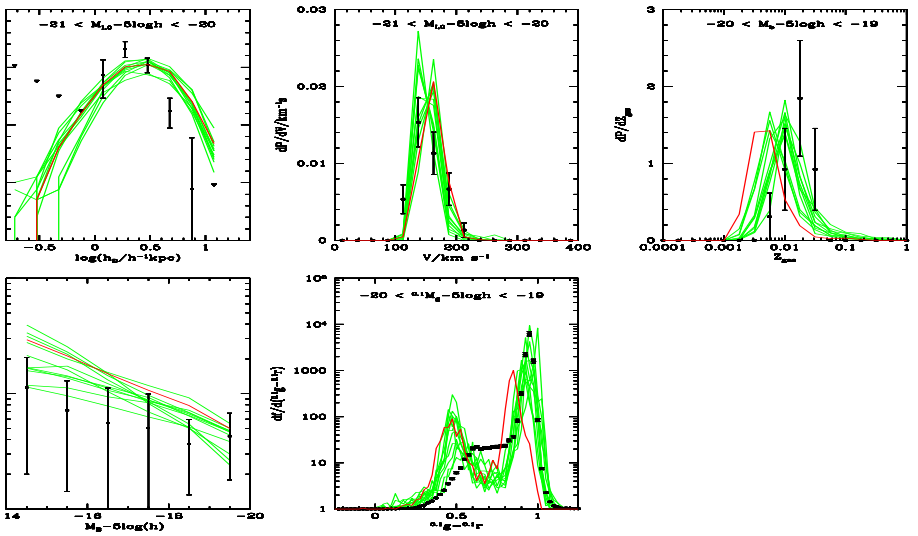

Figure 14: 5 new outputs of the Galform model describing galaxy disk sizes, TF relation, gas metallicity, gas mass to L_B and BH mass. The cosmologists best fit is in red, with a group of the best Wave 5 runs in green. Already we have found better simultaneous fits to these additional data sets.

examples of such output datasets describing various galaxy properties (disk sizes, TF relation, gas metallicity, gas mass to L_B and BH mass), along with corresponding observed data (the black points) are shown in figure 14 (see Bower et al. (2010)). The single red line represents the cosmologists' single best run prior to this analysis, and the green lines are ten of the best Wave 5 runs. We found many runs that were substantially better fits to the luminosity functions than had ever been seen previously by the cosmologists, and as figure 14 shows, have already found several runs that are an improved match to these other output data sets. The next step in this ongoing collaboration is to apply the emulation and History Matching procedures outlined in this report to these new output data sets, in order to understand their impact on the input space, and to determine which regions of input space will provide acceptable matches to all possible outputs.

Acknowledgments

IRV and MG acknowledge the support of the Basic Technology initiative as part of the Managing Uncertainty for Complex Models project. IRV acknowledges the support of an EPSRC mobility fellowship. RGB acknowledges the support of a Durham-University Christopherson-Knott Fellowship. We would like to thank the Durham Semi-analytical group based at the Institute for Computational Cosmology, Physics Department, Durham University for access to the Galform model and to their computational resources.

References

Bastos, T. S. and O'Hagan, A. (2008). "Diagnostics for Gaussian process emulators." *Technometrics*, 51: 425–438. 640

Baugh, C. M. (2006). "A primer on hierarchical galaxy formation: the semi- analytical approach." *Rept. Prog. Phys.*, 69: 3101–3156. 622, 663, 664

Bower, R. G., Benson, A. J., et al. (2006). "The broken hierarchy of galaxy formation." *Mon.Not.Roy.Astron.Soc.*, 370: 645–655. 619, 622, 657, 664

Bower, R. G., Vernon, I., Goldstein, M., et al. (2010). "The Parameter Space of Galaxy Formation." *Mon.Not.Roy.Astron.Soc.*, 407: 2017–2045. 653, 655, 657, 658

Cole, S. et al. (2001). "The 2dF Galaxy Redshift Survey: Near Infrared Galaxy Luminosity Functions." *Mon. Not. Roy. Astron. Soc.*, 326: 255–273. 625, 646, 664, 665

Colless, M. et al. (2001). "The 2dF Galaxy Redshift Survey: Spectra and redshifts." *Mon.Not.Roy.Astron.Soc.*, 328: 1039–1066. 624

Conti, S., Gosling, J. P., Oakley, J. E., and O'Hagan, A. (2009). "Gaussian process emulation of dynamic computer codes." *Biometrika*, 96: 663–676. 630

Craig, P. S., Goldstein, M., Seheult, A. H., and Smith, J. A. (1996). "Bayes linear strategies for history matching of hydrocarbon reservoirs." In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (eds.), *Bayesian Statistics 5*, 69–95. Oxford, UK: Clarendon Press. 630, 633, 634, 667

— (1997). "Pressure matching for hydrocarbon reservoirs: a case study in the use of Bayes linear strategies for large computer experiments." In Gatsonis, C., Hodges, J. S., Kass, R. E., McCulloch, R., Rossi, P., and Singpurwalla, N. D. (eds.), *Case Studies in Bayesian Statistics*, volume 3, 36–93. New York: Springer-Verlag. 630, 633, 634

Cressie, N. (1991). *Statistics for Spatial Data*. New York: Wiley. 667

Cumming, J. A. and Goldstein, M. (2009). "Bayes linear uncertainty analysis for oil reservoirs based on multiscale computer experiments." In O'Hagan, A. and West, M. (eds.), *Handbook of Bayesian Analysis*. Oxford, UK: Oxford University Press. 633

Currin, C., Mitchell, T., Morris, M., and Ylvisaker, D. (1991). "Bayesian prediction of deterministic functions with applications to the design and analysis of computer experiments." *Journal of the American Statistical Association*, 86(416): 953–963. 626, 636

De Finetti, B. (1974). *Theory of Probability*, volume 1. London: Wiley. 628

— (1975). *Theory of Probability*, volume 2. London: Wiley. 628

Goldstein, M. (1999). "Bayes linear analysis." In Kotz, S. et al. (eds.), *Encyclopaedia of Statistical Sciences*, 29–34. Chichester: Wiley. 628

— (2006). "Subjective Bayesian Analysis: Principles and Practice." *Bayesian Analysis*, 1(3): 403–420. 630

— (2010). "External Bayesian analysis for computer simulators." In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M. (eds.), *To appear in Bayesian Statistics 9*. Oxford University Press. 630

Goldstein, M. and Rougier, J. C. (2006). "Bayes linear calibrated prediction for complex systems." *Journal of the American Statistical Association*, 101(475): 1132–1143. 633

— (2009). "Reified Bayesian modelling and inference for physical systems (with Discussion)." *Journal of Statistical Planning and Inference*, 139(3): 1221–1239. 628, 657

Goldstein, M. and Wooff, D. A. (2007). *Bayes Linear Statistics: Theory and Methods*. Chichester: Wiley. 628

Heitmann, K., Higdon, D., et al. (2009). "The Coyote Universe II: Cosmological Models and Precision Emulation of the Nonlinear Matter Power Spectrum." *Astrophys. J.*, 705(1): 156–174. 630

Higdon, D., Gattiker, J., Williams, B., and Rightley, M. (2008). "Computer Model Calibration Using High-Dimensional Output." *Journal of the American Statistical Association*, 103(482): 570–583. 633

Higdon, D., Kennedy, M., Cavendish, J. C., Cafeo, J. A., and Ryne, R. D. (2004). "Combining field data and computer simulations for calibration and prediction." *SIAM Journal on Scientific Computing*, 26(2): 448–466. 630

House, L., Goldstein, M., and Vernon, I. (2009). "Exchangeable Computer Models." *MUCM Technical Report 10/01, submitted to Journal of the Royal Statistical Society, Series B*. 626, 643, 657

Kennedy, M. C. and O'Hagan, A. (2001). "Bayesian calibration of computer models." *Journal of the Royal Statistical Society, Series B*, 63(3): 425–464. 633

Norberg, P., Cole, S., et al. (2002). "The 2dF Galaxy Redshift Survey: The b_J-band galaxy luminosity function and survey selection function." *Mon.Not.Roy.Astron.Soc.*, 336: 907–934. 625

Oakley, J. and O'Hagan, A. (2002). "Bayesian inference for the uncertainty distribution of computer model outputs." *Biometrika*, 89(4): 769–784. 630

O'Hagan, A. (2006). "Bayesian analysis of computer code outputs: A tutorial." *Reliability Engineering and System Safety*, 91: 1290–1300. 630

Pukelsheim, F. (1994). "The three sigma rule." *The American Statistician*, 48: 88–91. 634

Raftery, A. E., Givens, G. H., and Zeh, J. E. (1995). "Inference from a deterministic population dynamics model for bowhead whales (with Discussion)." *Journal of the American Statistical Association*, 90: 402–430. 633

Rougier, J. C. (2008). "Efficient emulators for multivariate deterministic functions." *Journal of Computational and Graphical Statistics*, 17(4): 827–843. 668, 669

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). "Design and analysis of computer experiments." *Statistical Science*, 4(4): 409–435. 626, 636

Santner, T. J., Williams, B. J., and Notz, W. I. (2003). *The Design and Analysis of Computer Experiments*. New York: Springer-Verlag. 626, 636, 667

Spergel, D. N. et al. (2003). "First Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Determination of Cosmological Parameters." *Astrophys. J. Suppl.*, 148: 175–194. 664

Springel, V. et al. (2005). "Simulating the joint evolution of quasars, galaxies and their large-scale distribution." *Nature*, 435: 629–636. 663

Vernon, I. and Goldstein, M. (2009). "Bayes Linear Analysis of Imprecision in Computer Models, with Application to Understanding Galaxy Formation." In Augustin, T., Coolen, F. P. A., Moral, S., and Troffaes, M. C. M. (eds.), *ISIPTA'09: Proceedings of the Sixth International Symposium on Imprecise Probability: Theories and Applications*, 441–450. Durham, UK: SIPTA. 652

Appendix A: A Guide to Galaxy Formation

The aim of galaxy formation studies is to understand why the universe appears as it does. We wish to explain the characteristic properties of galaxies, such as their distribution of luminosities, colours and ages. As we will describe below, the present problem is not so much to understand why galaxies form, but to understand why they are relatively few and far between. By understanding this, we hope also to explain why galaxy formation appears to proceed very differently to that expected in the simplest theories. The basic ingredients have been in place for some time (the force of gravity and radiative cooling of baryonic matter), but we are only now beginning to understand how the formation of galaxies is regulated. The surprising result is that the black holes (the densest known objects in the universe) appear to play a key role in this.

Galaxy Formation - a Beginners Guide

So how do galaxies form? Why is the universe filled with such objects? In principle, it is a straightforward consequence of the dominance of the gravitational force. Since all matter makes a positive contribution to the gravitational force, the clumping of the universe's mass is a runaway process. As the condensations of matter become denser, they become more effective as attractors. These matter concentrations are referred to as haloes. The observational evidence shows that most of this mass, however, is not normal, "baryonic", matter (that you and I are made from) and that the universe is dominated by "Cold Dark Matter" (CDM): massive particles that interact very weakly. The CDM particles may be associated with super-symmetric extensions of the standard model of particle physics. Recent observations have also shown that a vacuum energy contribution is required.

The CDM particles explain the collapse and growth of the gravitating dark matter haloes, but to describe the formation of the luminous galaxies, we must turn to the astrophysics of the baryonic matter. As the baryons are pulled together by the collapse of

the dark matter halo, they heat up and start to resist further compression. The baryonic gas (but not the collisionless dark matter) radiate this energy and cool leading to a run-away contraction that is only stopped by the conservation of angular momentum. The baryons form thin, cold spinning disks of gas. Further condensing leads to the formation of stars, and empirical measurements show that the rate of formation of stars is proportional to the surface density of gas (for current theoretical models, this empirical calibration is entirely sufficient).

In this scenario, small haloes are able to convert almost all their baryonic component into stars, but this does not accurately reflect the universe we live in. In contrast to our initial model, the fraction of the baryonic material that is observed to form into stars is rather small, only about 10% of the total baryonic content of the universe. The origin of this discrepancy is a key cosmological puzzle, and astronomers appeal to "feedback" to resolve the discrepancy: somehow the formation of stars must inject energy that prevents further gas cooling. One of the key aims of the GALFORM project is to identify the feedback schemes that are needed to account for the observed universe. In small galaxies, we believe that the primary regulation mechanism is supernovae: the energetic explosions that massive stars undergo at the end of their life. In weak gravitational potentials, these are capable of driving gas out of the galaxy.

The strength and importance of feedback is best assessed by comparing the observed galaxy mass function (the numbers of galaxies in a given mass per unit volume) with the halo mass function. If star formation were uniformly efficient, there would be a constant offset between the two. However, a comparison shows that they differ dramatically in shape: the dark matter mass function has far more small haloes than are observed to host dwarf galaxies in the universe and lack a sharp cut-off at high masses. While supernovae may solve the problem with faint galaxies, it cannot explain the sharp cutoff at high masses. Of the solutions proposed, the current front runner is a form of feedback associated with the accretion of gas on to black holes.

Black holes are tiny compared to galaxies, their size (measured as their Schwarzschild radius or radius of their event horizon) is only 1.5×10^8 km. It is surprising that an object so small can heat a volume with radius 10^{11} times larger. Yet this is just what is observed in clusters of galaxies. Clusters are gravitationally bound systems containing 1000s of galaxies and 10^{15} solar masses of (largely) dark matter. Gas at the centres of these systems is dense enough that it should cool, promoting the formation of stars in the central object. Yet, little cooling is observed. Instead these systems host a powerful radio galaxy — a galaxy with a central black hole (or AGN) that is the source of a jet of magnetised high energy plasma. Although the details are not yet clear, relativistic particle jets from the black hole are capable of replacing the energy that is lost as cooling, keeping the central gas hot and starving the central galaxy of fuel for star formation. The frequency of the discovery of such objects is also remarkable - they seem to occur everywhere the runaway cooling process would generate a problem. It is now widely accepted that it provides an essential ingredient for models that explain the formation of galaxies.

Modelling Galaxy Formation

There are essentially two approaches to modelling the formation of galaxies. These are usually referred to as "numerical simulation" and "semi-analytic modelling".

The idea of "numerical simulation" is simple and direct. A powerful computer is programmed with the fundamental physical equations that describe the growth of fluctuations of dark matter, the hydrodynamical response of the intergalactic gas and its loss of energy through key atomic cooling processes. However, the equations are missing some key components of galaxy formation physics and massively over-produce the abundance of stars. Unfortunately, such codes have no hope of directly following the formation of stars or the winds they may generate at their death, and are many more orders of magnitude from being able to track the formation of black holes or the processes that generate the jets that regulate the formation of bright galaxies.

"Semi-analytic modelling" represents the alternative approach. Rather than tackling the whole problem in a single numerical integration, we break it down into its separate components. Of course, we must make some level of approximation by doing this, but we hope to create a model that encompasses the main physical processes with a minimum of complexity. For example, one component of the model is the growth and merging of dark matter haloes. This can be computed through an analytic approximation or by running a numerical calculation that only includes the force of gravity. In terms of the behaviour of the dark matter, this approximation is extremely good. We must then add components to describe such features as the collapse and cooling of gas; the formation of stars; the growth of black holes; merging of galaxies; the feedback effect of supernova explosions and jets from black holes, and then link them together through a network of interactions. Adding further components complicates the model but may improve its physical realism and ability to match the data. Each component is based on the results of a targeted set of simulations - or, failing this, on physically plausible scaling relations. In many cases, however, the physical process is not completely understood or characterised: to cope with this we introduce a number of parameters to account for this uncertainty. The result is a network of equations (or algorithms) whose behaviour is driven by the underlying growth and merging of the dark matter haloes, and whose response is governed by a number of adjustable input parameters. Because of the intrinsic complexity of the galaxy formation problem, "semi-analytic models" currently offer the best avenue for progress.

Appendix B: Galform - Physical Details

We now outline some relevant technical details of the GALFORM code. For an extended description and discussion of the Galform implementation see Baugh (2006). In essence, the model consists of a set of modules, each having associated input parameters.

1. Dark matter merger trees. These are extracted from the "Millennium" dark matter simulation (Springel et al. (2005)). This is a full numerical simulation of the growth of dark matter structures in the universe from cosmological initial conditions.

The initial spectrum of density fluctuations is set to be consistent with the WMAP satellite observations of the cosmic microwave background (Spergel et al. (2003)). The subsequent evolution involves solving the gravitational N-body problem for a collection of 10^{10} particles. The computations took several months on state of the art super-computers at the Max Planck Society's Rechenzentrum in Munich, Germany. Fortunately, this part of the model need only be solved once, and the main part of the GALFORM code can then be applied to populate the dark matter haloes with galaxies. This approach improves accuracy over previous analytic approximations to gravitational structure growth, but means that we must fix the cosmological parameters for our model. In future, improved analytic modelling of the merger trees will allow us to include the uncertainty in the cosmological parameters. For now, cosmological parameters are fixed to the canonical year 3 observations of WMAP in which $\Omega_b = 0.045$, $\Omega_M = 0.25$, $\Lambda = 0.75$ and $\sigma_8 = 0.9$ at the present day. The model assumes $H_0 = 0.73$, although we quote luminosities and space densities in term of $h = H_0/100\text{kms}^{-1}$ so that this dependence is explicit.

2. Gas Accretion and Cooling. As dark matter haloes grow, the gas that they contain cools and flows to the centre. This occurs at different rates depending on the mass of the halo, and the rate at which the halo mass grows. The supply of gas is determined by computing the mass of gas for which the cooling timescale is less than the halo, and the mass of gas which has had sufficient time to cool and fall to the centre (Cole et al. 2001; Baugh 2006). The newer version of the code (referred to as B06), which is considered in this case study, made several important advances (Bower et al. (2006)). One of these is to emphasise the distinction between haloes for which the gas supply is limited by the rate of cooling (henceforth "hydrostatic" haloes) and those haloes for which the free-fall timescale is the limiting factor (henceforth "rapid cooling" haloes). In the B06 model, it is assumed that energy from the central black hole can only offset the cooling in hydrostatic haloes. The parameter α_{cool} determines the exact ratio of timescales at which this distinction is made.

3. Star Formation. As the hot gas cools or is accreted by a halo, it builds up a reservoir of cold gas in the central galaxy. This gas provides the fuel for the formation of further stars. The code assumes that the star formation rate is related to the dynamical timescale of the galaxy, and its mass of gas, giving

$$\dot{m}_* = \epsilon_\star \left(\frac{m_{\text{cold}}}{\tau_{\text{disk}}} \right) \left(\frac{v_{\text{disk}}}{200\text{kms}^{-1}} \right)^{\alpha_\star}$$

where $\dot{m}_*$ is the star formation rate, m_{cold} is the mass of cold gas, τ_{disk} is the disk dynamical time and v_{disk} is the disk rotation speed. $\alpha_\star$ and $\epsilon_\star$ are parameters that control the rate of star formation and its dependence on galaxy mass. In B06, an additional mode of star formation is also considered. If the disk becomes too massive, it becomes susceptible to warps that grow, funnelling gas to the centre of the galaxy. Such secular evolution may generate many of the bulges that are observed. In the model it is assumed that instabilities occur if the disk's gravity exceeds the stabilising gravity of the halo. The threshold at which this occurs is set by the parameter f_{stab}, at which point the disk stars are added to the galaxy's bulge and the disk gas is consumed in a burst of star formation.

4. Feedback - from supernovae. Soon after the most massive stars form, they explode in powerful supernova explosions. These are thought to be responsible for preventing the efficient formation of stars in small galaxies - as the stars form, gas is driven out of the system by the supernovae. We model feedback from supernovae by assuming that the ratio of material expelled from the galaxy into the halo to that formed into stars is given by the ratio β, where

$$\beta = (v_{\rm disk}/v_{\rm hot})^{-\alpha_{\rm hot}} \tag{23}$$

where $v_{\rm hot}$ and $\alpha_{\rm hot}$ are poorly constrained parameters. We allow $v_{\rm hot}$ to take different values for quiescent and burst star formation which we denote as $V_{\rm hot,burst}$ and $V_{\rm hot,disk}$.

The gas that is driven out of galaxies flows into the halo, but does not immediately become available for cooling. The timescale on which the gas becomes available is determined by the parameter $\alpha_{\rm reheat}$. If this is unity, and cooling is efficient, ejected gas will be allowed to fall back into the galaxy on the dynamical timescale.

5. Galaxy mergers. When dark haloes collide, the galaxies at their centres do not immediately merge. Rather their relative motion slowly decays due to dynamical friction. This process is discussed extensively in Cole et al. (2001). The merging time is set by an overall normalisation parameter $f_{\rm df}$. If the time since the halo was accreted is less than the merging time, the galaxy from the "satellite" galaxy orbits inside the larger one. Such satellite galaxies do not collect any gas from the halo, and so star formation quickly subsides as the cold gas reservoir is exhausted. If the time since accretion exceeds the merging timescale, the galaxy mergers with the central galaxy in the parent halo. If the mass ratio of the galaxies exceeds $f_{\rm ellip}$, this can cause disturbance to the underlying galaxy, transforming it from a spiral type galaxy to an elliptical one. This morphological transformation may be associated with a burst of star formation. If the mass ratio exceeds $f_{\rm burst}$, there is no morphological transformation, but a burst of star formation still occurs.

6. Black holes and their feedback. The model assumes that black holes grow through three distinct channels: (i) by black hole - black hole mergers when the parent galaxies merge; (ii) by accretion of gas that is funnelled to the galaxy centre during bursts of star formation (these being driven either by mergers or disk instabilities); (iii) by diffuse gas accretion from hydrostatic haloes (i.e., as a result of "radio mode" feedback). The star burst driven accretion results in luminous quasars, but the current model assumes that these events do not contribute to the feedback. The parameter $F_{\rm bh}$ controls the amount of gas that is accreted by the black hole in these events. The feedback from "radio mode" accretion is, however, of key importance. The mass growth of the black hole is determined from the energy output required to counter-balance cooling of the halo, i.e. we implicitly assume that the mass accretion rate increases until the net cooling rate decreases to zero. However, accretion onto black holes, although an abundant source of energy has limits. We limit the maximum energy output to be less than $\epsilon_{\rm Edd} L_{\rm Edd}$ where $L_{\rm Edd}$ is the Eddington luminosity of the black hole and $\epsilon_{\rm Edd}$ is an adjustable parameter. Current models for black hole accretion suggest that $\epsilon_{\rm Edd}$ is of order 1%.

7. Reionisation At very early times, the majority of gas in the universe is neutral (and the universe is opaque to ultra-violet light). As stars and quasars form in abundance, the universe quickly ionizes. This creates an additional form of heating that may be extremely important in very low-mass galaxies. The details of this process are very important for understanding the paucity of dwarf galaxies that orbit in the milky-way halo. However, we are here concentrating on the properties of much more massive systems where these effects are less significant and it is sufficient to parameterise this process by two parameters, $z_{\rm cut}$ and $v_{\rm cut}$. Here, $z_{\rm cut}$ defines the redshift at which re-ionisation occurs: at lower redshifts, gas cooling is prevented in haloes with circular velocity below $v_{\rm cut}$.

Appendix C: Construction of the Wave 1-4 Emulators.

Univariate Emulation: Wave 1

In section 4.3 we discuss the construction of the wave 1 emulator (see equation 6), and here we describe in detail the procedures involved in this process, namely, active variable selection, choice of g_{ij} functions, assessment of the regression coefficients β_{ij} and the Gaussian process parameters σ_{u_i}, σ_{w_i} and θ_i.

In choosing the set of active variables $x_{[A_i]}$ for each output i the aim is to explain a large amount of the variance of $f_i(x)$ using as few variables as possible. For each of the 7 outputs, we used the 993 wave 1 runs to initially reduce the set x_B by backwards stepwise elimination, starting with a model containing the 8 linear terms. At this stage individual inputs were discarded in turn based upon the size of their main effect. Before an input would be discarded, a third order polynomial was fitted to see the extent of variance explained with the current set of active variables. It was found that 5 active variables could explain satisfactory amounts of the variance of $f_i(x)$ for each output i (see table 2), based on the adjusted R^2 of the polynomial fits. In each case, more than 5 variables yielded little extra benefit (compared to the increase in the size of the input space), while less than 5 led to substantially worse fits.

Once the set of active variables $x_{[A_i]}$ has been determined, the full set of regression terms $g_{ij}(x_{[A_i]})$ can be chosen. This was done by forward stepwise selection starting with a model containing the linear terms in the active variables, and adding possible terms from the full 3rd order polynomial in the active variables, using standard stepwise routines in R, based on criteria such as AIC. When the regression terms have been chosen for each output $f_i(x)$, estimates for the $B = \{\beta_{ij}\}$ coefficients can be obtained using Ordinary Least Squares, assuming uncorrelated errors. We have a sufficiently large collection of model evaluations that such data analytic techniques will result in small variances on the regression coefficients and generally acceptable results from OLS fitting. Therefore, we would expect such results to overwhelm prior judgements. However, any substantial contradictions between the data and the qualitative form of such judgements requires further investigation.

As the $u_i(x_{[A_i]})$ represent local deviations from the regression surface, there will

be a correlation between u_i at neighbouring values of $x_{[A_i]}$, which we must specify. Various choices are available, each of which involves parameters related to the width and shape of the correlation function. Estimation of these parameters can be a difficult task. However, these parameters are representations of our subjective assessment of the smoothness of the function and precise assessment of them is not necessarily meaningful, and nor is it required in order to construct an emulator of sufficient accuracy for our needs. Here we choose the following Gaussian covariance structure:

$$\mathrm{Cov}(u_i(x_{[A_i]}), u_i(x'_{[A_i]})) \quad = \quad \sigma^2_{u_i} \exp(-||x_{[A_i]} - x'_{[A_i]}||^2/\theta_i^2), \tag{24}$$

where $\sigma^2_{u_i}$ is the point variance at any given $x_{[A_i]}$, θ_i is the correlation length parameter that controls the strength of correlation between two separated points in the input space (for points a distance θ apart, the correlation will be exactly $\exp(-1)$), and $||\cdot||$ is the Euclidean norm. As $w_i(x_B)$ represents all the remaining variation in the inactive variables, it is often small and we treat it as uncorrelated random noise with $\mathrm{Var}(w_i(x_B)) = \sigma^2_{w_i}$. We consider the point variances of these two processes to be proportions of the overall residual variance of the computer model given the emulator trend, σ_i^2, and write $\sigma^2_{u_i} = (1 - w_i)\sigma_i^2$ and $\sigma^2_{w_i} = w_i\sigma_i^2$ for some small w_i. Various techniques for estimating the correlation length and parameters θ_i and w_i from the data are available (for example variograms (Cressie (1991)), REML (Santner et al. (2003)); however, these estimation procedures can often be non-robust as the output from a computer model rarely behaves exactly like an actual Gaussian Process. An alternative is to specify the θ_i parameters a priori (Craig et al. 1996) followed by an approximate assessment of the nugget term w_i, which is the approach we adopt here.

We may provide approximate order of magnitude values for the correlation length parameters θ_i, by appealing to the heuristic that the regression residuals may be viewed as deriving from a polynomial of order one higher than the fitted polynomial, as they correspond to the first order of terms which are neglected by the regression fit. Here this implies that values of θ_i should be chosen corresponding to the shape of a 4th order polynomial. In such a case, we would not want the correlation length to be greater than the average distance between roots of a 4th order polynomial: approximately 0.25 of the range of the input. Alternatively it can be argued that there should be positive correlation between outputs at the turning points and the adjacent roots of the polynomial, and that the correlation length must therefore be greater than this distance: approximately 0.125 of the range of the input. This argument tends to give more conservative (i.e. smaller) specifications for the correlation length compared to maximum likelihood or variogram methods. As we have scaled all inputs to the range $[-1, 1]$, this argument suggests that a working estimate of θ_i might lie between 0.25 and 0.5, and therefore we selected the same value for all θ_i of 0.35, checked by emulator diagnostics discussed in section 4.4.

The value of the nugget parameter w_i represents the proportion of residual variance due to the inactive variables. We obtained a working assessment of w_i by examining the variance explained by the inactive variables for each of the seven outputs, and comparing this to the residual variance from the active variable polynomial fit. These considerations led to a conservative value of 0.2 for all w_i acknowledging a reasonable

contribution from the inactive variables at each output. Provided conservative choices are made and are combined with analysis of the emulator diagnostics, such specifications lead to emulators of sufficient accuracy for the task of providing a first stage reduction of the input space, while avoiding the complex and often misleading problem of estimating such parameters from the data alone. At this stage, we only require a relatively simple emulator in order to make an initial reduction of the input space, while leaving the construction of more detailed emulators to subsequent waves of the analysis.

Univariate Emulation: Waves 2 to 4

The Wave 2 to 4 univariate emulators were constructed using similar methods as were used in Wave 1, as described in detail in section 4.3. Here we give a summary of their construction, highlighting the differences with the Wave 1 case.

Recall that for Waves 1-3 we only explored 8 of the input parameters, which were the set of proposed active variables described in section 4.2 and shown in table 1, with the effect of the remaining 9 inputs being described by the model discrepancy term Φ_{IA} (see section 5.1). The selection of Wave 2 and 3 Active Variables proceeded as for Wave 1, and it was found that all 8 input parameters were required as active in these cases. Therefore, the only difference to the form of the Wave 1 emulator given by equation (6), is that now there is no nugget term $w_i(x_B)$. The selection and fitting of the polynomial terms was performed as in section 4.2 and appendix C.1, and a similar Gaussian covariance function to equation (24) was assumed. In Wave 4, it was found that improved polynomial fits could be obtained using 10 active variables, composed of the 8 variables used in Wave 1-3 (and given in table 1) with the addition of the inputs alphastar and tau0mrg. The remaining 7 inputs were found to have little impact on the 11 luminosity function outputs considered. As the effect of all 17 inputs are represented by the Wave 4 emulator, the Φ_{IA} model discrepancy term (representing the 9 previously inactive variables) was dropped at this stage. Table 3 summarises the number of runs used at each wave, along with the number of active variables required.

Multivariate Emulation: Waves 3 and 4

In Waves 1 and 2 univariate emulators were used, which allow only the use of univariate implausibility measures to reduce the input space. Therefore, at Wave 3 we constructed a multivariate emulator in order to develop the corresponding multivariate implausibility measure $I(x)$ introduced in section 3.5. $I(x)$ will be of use as it measures different aspects of the model output compared to the univariate implausibility measures, namely it is sensitive to the shape of the luminosity functions.

Constructing a tractable multivariate emulator can be a challenging task. An emulator that utilizes a weakly stationary process (such as $u_i(x)$ in equation (24)) suffers from what is referred to as the $(nq)^3$ problem (Rougier (2008)), where n is the number of model evaluations and q is the number of outputs to be emulated. The process of updating our beliefs about the emulator given the n model evaluations generally requires

the inverting of a matrix of size $nq \times nq$, a computation that scales as $(nq)^3$. At Wave 4 say we have $n = 2011$ and $q = 11$, leading to a problematic matrix inversion of size 22121. However, by specifying covariance structures of suitably symmetric form this problem can be avoided.

The wave 3 emulator has the same form as that of wave 2, where again we use all 8 inputs as Active Variables (that is $x_{[A_i]} = x_B$), and we consider the same set of 11 outputs. Again the $g_{ij}(x_B)$ and β_{ij} terms were chosen by model selection techniques and OLS fitting respectively: we compare these polynomials to those of previous waves in section 7.2. We then assume the following separable multivariate covariance structure for the process $u_i(x_B)$:

$$\mathrm{Cov}(u_i(x_B), u_j(x'_B)) \quad = \quad \Sigma_{ij} \exp(-||x_{[B]} - x'_{[B]}||^2/\theta^2), \tag{25}$$

where the i and j indices denote each of the 11 outputs, Σ is an 11×11 covariance matrix and note we have removed the i index on θ as we have assumed the same correlation length for each output. We assess the matrix Σ by taking the covariance matrix of the 11 sets of residuals from each of the polynomials. The separable form of equation (25) allows the above problematic matrix to be written as a direct product, which greatly simplifies the calculation of its inverse, as we can invert each component of the direct product individually. See Rougier (2008) for further discussions regarding calculations for multivariate emulators.

The construction of a multivariate emulator allows the use of a Multivariate Implausibility measure which can be defined as (using equation (16)):

$$I^2(x) = (\mathrm{E}(f(x)) - z)^T (\mathrm{Var}(f(x)) + \mathrm{Var}(\epsilon_{md}) + \mathrm{Var}(\epsilon_{obs}))^{-1} (\mathrm{E}(f(x)) - z). \tag{26}$$

$I(x)$ is a useful measure to consider as it captures the shape of the luminosity function output. It will allow the discarding of inputs corresponding to runs that satisfy the univariate matching criteria and hence are close to the data points, but that have an unphysical shape in either the bj or K luminosity function.