Samuel E. Jackson*, Ian Vernon, Junli Liu and Keith Lindsey

Understanding hormonal crosstalk in Arabidopsis root development via emulation and history matching

https://doi.org/10.1515/sagmb-2018-0053
Received September 7, 2018; accepted May 12, 2020

Abstract: A major challenge in plant developmental biology is to understand how plant growth is coordinated by interacting hormones and genes. To meet this challenge, it is important to not only use experimental data, but also formulate a mathematical model. For the mathematical model to best describe the true biological system, it is necessary to understand the parameter space of the model, along with the links between the model, the parameter space and experimental observations. We develop sequential history matching methodology, using Bayesian emulation, to gain substantial insight into biological model parameter spaces. This is achieved by finding sets of acceptable parameters in accordance with successive sets of physical observations. These methods are then applied to a complex hormonal crosstalk model for Arabidopsis root growth. In this application, we demonstrate how an initial set of 22 observed trends reduce the volume of the set of acceptable inputs to a proportion of 6.1×10^{-7} of the original space. Additional sets of biologically relevant experimental data, each of size 5, reduce the size of this space by a further three and two orders of magnitude respectively. Hence, we provide insight into the constraints placed upon the model structure by, and the biological consequences of, measuring subsets of observations.

Keywords: Arabidopsis; Bayesian uncertainty analysis; emulation; history matching; parameter search.

1 Background

1.1 Use and understanding of scientific models in systems biology

One of the major challenges in biology is to understand how functions in cells emerge from molecular components. Computational and mathematical modelling is a key element in systems biology which enables the analysis of biological functions resulting from non-linear interactions of molecular components. The kinetics of each biological reaction can be systematically represented using a set of differential equations (Alves et al. 2006; Boogerd et al. 2007; Jamshidi and Palsson 2008; Liu et al. 2010; Smallbone et al. 2010). Due to the multitude of cell components and the complexity of molecular interactions, the kinetic models often involve large numbers of reaction rate parameters, that is parameters representing the rates at which reactions encapsulated by the model are occurring (Mobius and Laan 2015; Moore et al. 2015a, 2015c). Quantitative experimental measurements can be used to formulate the kinetic equations and learn about the associated rate parameters (Boogerd et al. 2007; Liu et al. 2010; Liu et al. 2013; Mobius and Laan 2015; Torres and Santos 2015). This in turn provides insight into the functions of the actual biological system.

***Corresponding author: Samuel E. Jackson**, Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, UK, e-mail: s.e.jackson@soton.ac.uk. https://orcid.org/0000-0003-3695-5362
Ian Vernon: Department of Mathematical Sciences, Durham University, Durham, UK
Junli Liu and Keith Lindsey: School of Biological and Biomedical Sciences, Durham University, Durham, UK

An important question is therefore how much information about the kinetic equations and parameters can be obtained from an experimental measurement. Since a key aspect of experimental measurements in modern biological science is the study of the functions of specific genes, the answer to the above question is also important for understanding the role of each gene within the components of a biological system.

In plant developmental biology, a major challenge is to understand how plant growth is coordinated by interacting hormones and genes. Previously, a hormonal crosstalk network model – which describes how three hormones (auxin, ethylene and cytokinin) and the associated genes coordinate to regulate Arabidopsis root development – was constructed by iteratively combining regulatory relationships derived from experimental data and mathematical modelling (Liu et al. 2010, 2013; Moore et al. 2015a, 2015b, 2015c, 2017). However, for the mathematical model to best link with Arabidopsis root development, it is necessary to understand the parameter space of the model and identify all acceptable parameter combinations. Little is known about how acceptable parameter combinations of a model can be identified in light of specific experimental data. Therefore, this work explores how the acceptable parameter space of a complex model of hormonal crosstalk (Liu et al. 2013; Moore et al. 2015a, 2015b, 2015c, 2017) is assessed given a combination of quantitative experimental measurements and qualitative experimental trends by employing Bayesian history matching techniques (Craig et al. 1997; Cumming and Goldstein 2010; Gong and DiazDelaO 2017; Zhang et al. 2012). Additionally, we utilise these techniques to analyse how learning about the functions of a gene through particular relevant experiments can inform us about acceptable model parameter space.

1.2 Efficient analysis of scientific models

Complex systems biology models are frequently high dimensional and can take a substantial amount of time to evaluate (Moore et al. 2015c), thus comprehensive analysis of the entire input space, requiring vast numbers of model evaluations, may be unfeasible (Vernon et al. 2018). We are frequently interested, as is the case in this paper, in comparing the scientific model to observed data (usually measured with uncertainty), necessitating a possibly high dimensional parameter search. Our history matching approach aims to find the set of all possible combinations of input rate parameters which could have plausibly led to the observed data, given all sources of uncertainty involved with the model and the experimental data (Craig et al. 1997; Vernon et al. 2010a, 2018). This biologically relevant aim requires comprehensive exploration of the model's behaviour over the whole input space, and therefore efficient techniques, such as emulation (Castelletti et al. 2012; Craig et al. 1997; Kennedy and O'Hagan 2001; Williamson et al. 2013), are required. An emulator mimics the biological model, but is substantially faster to evaluate, hence facilitating the large numbers of evaluations that are needed.

We are often keen to understand the contribution of particular sets of observations towards being able to answer critical scientific questions. Sequential incorporation of datasets into a history matching procedure, as presented in this article, is very natural and can allow us to attain such understanding. Comprehensive understanding and parameter searching of the hormonal crosstalk model for Arabidopsis root development (Liu et al. 2013), by sequentially history matching specific groups of experimental observations, is the focus of this paper.

2 Methods

2.1 Bayes linear emulation

In this section we review the process of constructing an emulator for a complex systems biology model. For more detail see Vernon et al. (2018). We represent the set of input rate parameters of the model as a vector x of length d, and the outputs of the model as vector $f(x)$ of length q.

A Bayes linear emulator is a fast statistical approximation of the systems biology model built using a set of model runs, providing an expected value for the model output at a particular point x, along with a corresponding uncertainty estimate reflecting our beliefs about the uncertainty in the approximation (Goldstein 1999; Goldstein and Wooff 2007). The main advantage of emulation is its computational efficiency: often an emulator is several orders of magnitude faster to evaluate than the model it is mimicking. Emulation has been successfully applied across a variety of scientific disciplines such as climate science (Castelletti et al. 2012; Castruccio et al. 2014), cosmology (Bower et al. 2010; Heitmann et al. 2010), epidemiology (Farah et al. 2014), humanitarian relief (Overstall and Woods 2016), as well as systems biology (Vernon et al. 2018).

We index by $i = 1,\ldots,q$ the output components of the model. Each output component of the model $f_i(x)$ can be represented in emulator form as presented by Vernon et al. (2010a):

$$f_i(x) = \sum_{j=1}^{J_i} \beta_{ij} g_{ij}(x_{A_i}) + u_i(x_{A_i}) + w_i(x) \tag{1}$$

where x_{A_i} represents the subset of active variables, that is the input components of x which are most influential for output $f_i(x)$, g_{ij} are J_i known functions of x_{A_i}, and β_{ij} are the corresponding coefficients to the regression functions g_{ij}. $u_i(x_{A_i})$ is a second-order weakly stationary stochastic process which captures residual variation in x_{A_i}. *A priori*, we assume that $\mathrm{E}[u_i(x_{A_i})] = 0$, along with the following covariance structure:

$$\mathrm{Cov}[u_i(x_{A_i}), u_i(x'_{A_i})] = \sigma^2_{u_i} \exp\left(- \sum_{k \in S_{A_i}} \left\{ \frac{x_k - x'_k}{\theta_{ik}} \right\}^2 \right) \tag{2}$$

where S_{A_i} is the set of indices of the active inputs for output i. $w_i(x)$ is a zero-mean "nugget", or residual error, term with constant variance $\sigma^2_{w_i}$ over x and $\mathrm{Cov}[w_i(x), w_i(x')] = 0$ for $x \neq x'$. The nugget represents the effects of the remaining inactive input variables (Vernon et al. 2010a). We also make the assumption that:

$$\mathrm{Cov}\left[\beta_{ij}, u_i(x_{A_i})\right] = \mathrm{Cov}\left[\beta_{ij}, w_i(x)\right] = \mathrm{Cov}[u_i(x_{A_i}), w_i(x)] = 0$$

for all i, j. There are several methods available for eliciting the parameters in Equation (2). We may have *a priori* beliefs as to sensible values of these parameters, perhaps as a result of expert knowledge of the general behaviour of the model, or from past computer experiments involving the same or a similar model. Alternatively, we may estimate the values of the parameters from the data, for example, using maximum likelihood estimates if we are happy with any implication of the specified prior distributions, or via predictive diagnostics such as Leave-One-Out-Cross-Validation (Andrianakis and Challenor 2011). Crucially, the resulting emulator should be assessed for adequacy using diagnostic measures, such as those discussed at the end of this section.

Suppose $D_i = (f_i(x^{(1)}), \ldots, f_i(x^{(n)}))$ represents model output component i evaluated at n model runs performed at locations $x^{(1)},\ldots, x^{(n)}$. The Bayes linear emulator output for simulator output component i at a new x is given by the Bayes linear update formulae (Goldstein 1999; Goldstein and Wooff 2007):

$$\mathrm{E}_{D_i}[f_i(x)] = \mathrm{E}[f_i(x)] + \mathrm{Cov}[f_i(x), D_i]\mathrm{Var}[D_i]^{-1}(D_i - \mathrm{E}[D_i]) \tag{3}$$

$$\mathrm{Var}_{D_i}[f_i(x)] = \mathrm{Var}[f_i(x)] - \mathrm{Cov}[f_i(x), D_i]\mathrm{Var}[D_i]^{-1}\mathrm{Cov}[D_i, f_i(x)] \tag{4}$$

where the notation $\mathrm{E}_{D_i}[f_i(x)]$ and $\mathrm{Var}_{D_i}[f_i(x)]$ reflects the fact that we have adjusted our prior beliefs about $f_i(x)$ by model runs D_i, and can be obtained for any point x using Equations (1) and (2). We note that in the literature it is common to assume normal and Gaussian process priors for β and $u(x)$ in Equation (1) (Conti et al. 2009; Johnson et al. 2011; Kennedy and O'Hagan 2001), thus resulting in Gaussian process emulation (see Rasmussen and Williams (2006) for a general discussion of Gaussian processes). The resulting Bayesian update equations using Gaussian processes are practically similar to Equations (3) and (4) presented above, however, methodologically involve additional distributional assumptions. We would rather go as far as possible without making such distributional assumptions, as they may not always be valid, whilst still affecting the resulting inference. In this spirit, the Bayes linear framework is more similar to traditional kriging (Journel and Huijbregts 1978), noting that this term is now sometimes used to mean several different related approaches. Having said that, we note that kriging is derived from classical unbiased estimator arguments, whereas the Bayes linear paradigm follows from a foundational position, following DeFinetti (de Finetti 1974, 1975), that treats expectation as primitive and does not invoke concepts such as unbiasedness. The Bayes linear paradigm has been applied in a wide range of scenarios (Gosling et al. 2013; O'Hagan 1987; Whittle 1958): for example, in the context of the emulation of computer models, it has allowed tractable multilevel emulation due to multi-fidelity models, thus going beyond standard universal kriging (Craig et al. 1997; Cumming and Goldstein 2007).

Emulator design is the process of selecting the points in the input space $x^{(1)},\ldots, x^{(n)}$ at which the simulator will be run in order to construct an emulator (Santner et al. 2003). A popular design choice in the computer model literature is the Maximin Latin Hypercube design (Currin et al. 1991; McKay et al. 1979), however, other options are also available (see, for example, Fisher (1937); Montgomery (2009)).

The quality of any emulator should be assessed using diagnostics to judge whether it is fit for purpose (Bastos and O'Hagan 2008). For example, we can calculate standardised prediction errors:

$$U_i(x) = \frac{f_i(x) - \mathrm{E}_{D_i}[f_i(x)]}{\sqrt{\mathrm{Var}_{D_i}[f_i(x)]}} \tag{5}$$

for a set of validation data. Large errors $U_i(x)$ indicate conflict between simulator and emulator as a result of the emulator being overconfident in its predictions, and hence the emulator is not valid for inference. Systematically small errors indicate that the emulator is underconfident. As a rule of thumb, we expect most of the errors to lie between −3 and 3, appealing to Pukelsheim's 3σ rule, which states that at least 95% of the probability mass of any unimodal continuous distribution will lie within ±3 standard deviations of the mean, regardless of asymmetry or skew (Pukelsheim 1994). To contrast with a Gaussian process emulator, these errors should theoretically follow a (standard) normal or a t-distribution, depending on the precise method of emulator update and inference (for example, approach to selecting the correlation length parameters). Poor diagnostics suggest that the emulator prior beliefs were misspecified, for example, as a result of incorrect prior specifications for the parameters β, σ_u^2 and θ. Alternatively, it could be indication of an erratically behaved model that would require substantially more model runs in order to be emulated well.

2.1.1 Dimensional example: In this section we demonstrate emulation techniques on a simple one-dimensional example. We will suppose that we wish to emulate the simple function $f(x) = 0.1x + \cos(x)$ in the range $[0, \frac{11\pi}{3}]$, where we treat x as a rate parameter that we wish to learn about, and $f(x)$ as a chemical concentration that we could measure. We assume an emulator of the form given by Equation (1) with covariance structure given by Equation (2). We assume a zero mean function, that is, $g^T(x)\beta = 0$, and that $\sigma_u^2 = 0.5$, $\theta = 1.5$ and $\sigma_w^2 = 0$. We also specify a prior expectation $\mathrm{E}[f(x)] = 0$. Having specified our prior beliefs, we then use the update rules given by Equations (3) and (4) to obtain the adjusted expectation $\mathrm{E}_D[f(x)]$ and variance $\mathrm{Var}_D[f(x)]$ for $f(x)$. The results of this emulation process are shown in the left panel of Figure 1. The blue lines represent the emulator expectation $\mathrm{E}_D[f(x)]$ of the simulator output for the test points. The red lines represent the emulator mean ±3 emulator standard deviations, given as $\mathrm{E}_D[f(x)] \pm 3\sqrt{\mathrm{Var}_D[f(x)]}$, these being bounds for a 95% credible interval, following Pukelsheim's 3σ rule (Pukelsheim 1994). By comparison with the right panel of Figure 1, we can see that the emulator estimates the simulator output well, with some uncertainty. We note that we would not expect such large emulator uncertainty on such a smooth function as this, but have deliberately ensured that there is a large uncertainty for illustrative purposes, and in particular to highlight the effects of additional runs on reducing emulator uncertainty in the continuation of this example in Section 2.2.

2.2 History matching

History matching concerns the problem of finding the set of inputs to a model for which the corresponding model outputs give acceptable matches to observed historical data, given our state of uncertainty about the model itself and the measurements. History matching has been successfully applied across many scientific disciplines including oil reservoir modelling (Craig et al. 1996, 1997; Cumming and Goldstein 2009, 2010; Oliver and Chen 2011), engineering (Gardner et al. 2018; Gong and DiazDelaO 2017), epidemiology (Andrianakis et al. 2015, 2017a, 2017b; McCreesh et al. 2017), climate modelling (Williamson et al. 2013) and systems biology (Vernon et al. 2018). Here we provide a brief summary of the history matching procedure (see Vernon et al. (2010a, 2018) for more details).

We need a general structure to describe the link between a complex model and the corresponding physical system. We use the direct simulator approach, otherwise known as the best input approach (Goldstein and Rougier 2006), where we posit that there exists a value x^* such that $f(x^*)$ best represents the real biological system, which we denote as y (Goldstein and Rougier 2006, 2009). We then formally link the ith output of the model to the ith real system value y_i via

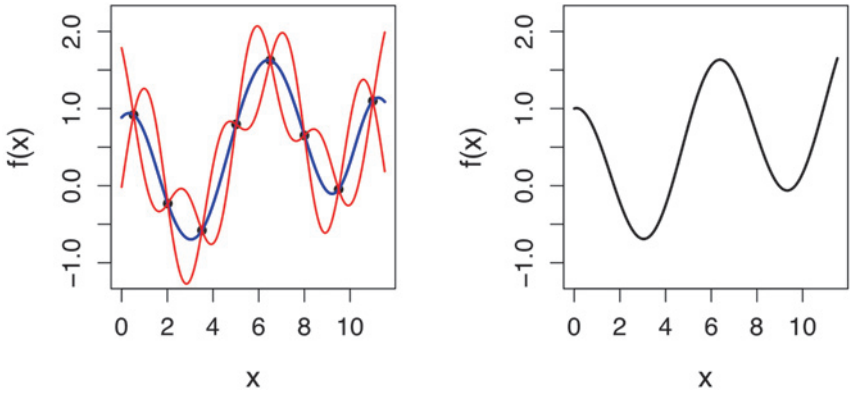

Figure 1: Left: Emulator expectation $\mathrm{E}_D[f(x)]$ (blue) with emulator credible intervals $\mathrm{E}_D[f(x)] \pm 3\sqrt{\mathrm{Var}_D[f(x)]}$ (red) for an emulator of $f(x) = 0.1x + \cos(x)$ constructed using 8 training points. Right: Simulator function $f(x) = 0.1x + \cos(x)$.

$$y_i = f_i(x^*) + \epsilon_i \tag{6}$$

and link the experimental observation z_i corresponding to output i to the real system via

$$z_i = y_i + e_i \tag{7}$$

where we assume $f_i(x^*) \perp\!\!\!\perp \epsilon_i \perp\!\!\!\perp e_i$, with $a \perp\!\!\!\perp b$ indicating that random variables a and b are uncorrelated (Goldstein and Wooff 2007). Here, $f_i(x^*)$ is the model run at best input x^*, ϵ_i is a random variable which reflects our uncertainty due to discrepancy between the model run at the best possible input combination setting and the real world (Arendt et al. 2012; Brynjarsdottir and O'Hagan 2014; Goldstein et al. 2013; Kennedy and O'Hagan 2001), and e_i is a random variable which incorporates our beliefs about the error between each desired real world quantity and its corresponding observed measurement. We assume $\mathrm{E}[\epsilon_i] = \mathrm{E}[e_i] = 0$, $\mathrm{Var}[\epsilon_i] = \sigma^2_{\epsilon_i}$ and $\mathrm{Var}[e_i] = \sigma^2_{e_i}$. The connection between system, observation and model given by (6) and (7) is simple but well-used (Andrianakis et al. 2015; Craig et al. 1997; Goldstein and Rougier 2006), and judged sufficient for our purposes. For discussion of more advanced approaches see Goldstein and Rougier (2009).

We then aim to find the set $\mathscr{X}^*$ of all input combinations x that are consistent with Equations (6) and (7), that is those that will provide acceptable matches between model output and data. Note that classifying points in this way can lead to $\mathscr{X}^*$ being empty, an informative conclusion which would contradict the posited existence of x^* in Equation (6), and imply that the model may not be fit for purpose. To analyse whether a point $x \in \mathscr{X}^*$ it is practical to use implausibility measures for each output i, as given, for example, in Craig et al. (1996, 1997); Vernon et al. (2010a):

$$I_i^2(x) = \frac{\left(\mathrm{E}_{D_i}[f_i(x)] - z_i\right)^2}{\mathrm{Var}_{D_i}[f_i(x)] + \sigma^2_{\epsilon_i} + \sigma^2_{e_i}} \tag{8}$$

If $I_i(x)$ is large this suggests that we would be unlikely to obtain an acceptable match between model output and observed data were we to run the model at x. This is after taking into account all the uncertainties associated with the model and the measurements. We develop a combined implausibility measure over multiple outputs such as $I_M(x) = \max_i I_i(x)$, $I_{2M}(x) = \max_i(\{I_i(x)\} \backslash I_M(x))$ and $I_{3M}(x) = \max_i(\{I_i(x)\} \backslash \{I_M(x), I_{2M}(x)\})$ (Vernon et al. 2010a), where $\mathrm{S}_1 \backslash \mathrm{S}_2$ is general set notation for in set S_1 but not set S_2. We class x as implausible if the values of these measures lie above suitable cutoff thresholds (Craig et al. 1997; Vernon et al. 2010a).

History matching using emulators proceeds as a series of iterations, called waves, discarding regions of the input parameter space at each wave. At the kth wave emulators are constructed for a selection of well-behaved outputs Q_k over the non-implausible space remaining after wave $k-1$. These emulators are used to assess implausibility over this space where points with sufficiently large values are discarded to leave a smaller set $\mathscr{X}_k$ remaining (Vernon et al. 2010a, 2018).

The history matching algorithm is as follows:

(1) Let $\mathscr{X}_0$ be the initial domain space of interest and set $k = 1$.
(2) Generate a design for a set of runs over the non-implausible space $\mathscr{X}_{k-1}$, for example using a maximin Latin hypercube with rejection (Vernon et al. 2010a).
(3) Check to see if there are new, informative outputs that can now be emulated accurately and add them to the previous set Q_{k-1} to define Q_k.
(4) Use the design of runs to construct new, more accurate emulators defined only over $\mathscr{X}_{k-1}$ for each output in Q_k.
(5) Calculate implausibility measures over $\mathscr{X}_{k-1}$ for each of the outputs in Q_k.
(6) Discard points in $\mathscr{X}_{k-1}$ with $I(x) > c$ to define a smaller non-implausible region $\mathscr{X}_k$.
(7) If the current non-implausible space $\mathscr{X}_k$ is sufficiently small, go on to step 8. Otherwise repeat the algorithm from step 2 for wave $k+1$. The non-implausible space is sufficiently small if it is empty or if the emulator variances are small in comparison to the other sources of uncertainty (σ^2_ϵ and σ^2_e), since in this case more accurate emulators would do little to reduce the non-implausible space further.
(8) Generate a large number of acceptable runs from $\mathscr{X}_k$, sampled according to scientific goal.

It should be the case that $\mathscr{X}^* \subseteq \mathscr{X}_k \subseteq \mathscr{X}_{k-1}$ for all k, where $\mathscr{X}^* = \{x : \max_i I_i(x) < c\}$ for some threshold c, where each $I_i(x)$ is calculated using expression (8) with $\mathrm{E}[f_i(x)] = f_i(x)$ and $\mathrm{Var}[f_i(x)] = 0$, that is were we to know the simulator output everywhere. The choice of cutoff $c = 3$ is frequently chosen, motivated by Pukelsheim's 3-sigma rule (Pukelsheim 1994), which in this case implies that $P(I_i(x) < 3 | x = x^*) > 0.95$ for any unimodal continuous distribution for the combined error term $\epsilon_i + e_i + (f_i(x) - \mathrm{E}[f_i(x)])$. This iterative procedure is powerful as it quickly discards large regions of the input space as implausible based on a small number of well behaved (and hence easy to emulate) outputs. In later waves, outputs that were initially hard to emulate, possibly due to their erratic behaviour in uninteresting parts of the input space, become easier to emulate over the much reduced space $\mathscr{X}_k$. Careful consideration of the initial non-implausible space $\mathscr{X}_0$ is important. It should be large enough such that no potentially scientifically interesting input combinations are excluded. A more in-depth discussion of the benefits of this history matching approach, especially in problems requiring the use of emulators, may be found in Vernon et al. (2018).

Note that whilst in this article we take $\mathrm{E}_{D_i}[f_i(x)]$ and $\mathrm{Var}_{D_i}[f_i(x)]$ to be adjusted beliefs resulting from a Bayes linear emulator, such as those discussed in Section 2.1, these quantities can be substituted with the mean and variance resulting from the corresponding estimates of a Gaussian process emulator if preferred (see, for example, Hamdi et al. (2017) and Gardner et al. (2018)). In this case, history matching can proceed practically similarly. The main difference lies in the choice of implausibility cutoff threshold c, which will be lower in comparison to that used in a Bayes linear framework whilst ensuring the same upper bound on the probabilities of a false rejection (say around 2 as opposed to 3 above). This lower threshold can be utilised by appealing to the features of the spread of probability mass of the predictions assumed by making use of Gaussian processes, as opposed to general rules such as Pukelsheim's 3σ rule discussed above. A brief comparison of Bayes linear emulators and Gaussian process emulators was presented in Section 2.1. For further discussion and a comparison of using Bayes linear emulators and Gaussian process emulators within a history match, see Vernon et al. (2010c).

2.2.1 Dimensional example continued: Figure 2 (left panel) shows the emulator expectation and conservative bounds for the 95% credible intervals, as given by Figure 1 (left panel), however, now an observation $z = -0.3$ along with observed error is included as solid and dashed lines respectively. In this example, we let model discrepancy be 0, and set the measurement standard error $\sigma_e = 0.05$. Along the bottom of the figure is the implausibility $I(x)$ for each x value represented by colour: red for large implausibility values, orange and yellow for borderline implausibility, and green for low implausibility ($I(x) < 3$) (Pukelsheim 1994).

$\mathscr{X}_0$ is the full initial range of x, that is $0 \le x \le \frac{11\pi}{3}$. $\mathscr{X}_1$ is as shown by the green regions in Figure 2 (left panel). Wave 2, shown in Figure 2 (right panel), involves designing a set of three more runs over $\mathscr{X}_1$, constructing another emulator over this region and calculating implausibility measures for each $x \in \mathscr{X}_1$. This second emulator is considerably more accurate than the observed error, thus $\mathscr{X}_2 \approx \mathscr{X}^\star$, so the analysis can be stopped at this point as extra runs would do little to further reduce the non-implausible space.

2.3 Sequential history matching of observations

Part of the novelty of our history matching approach involves dividing experimental observations into subsets and sequentially performing a history match on the model using each group of observations. Much scientific insight can be gained from performing a history match, however, using all output components simultaneously can mask which experiments are informative for certain aspects of the scientific system.

Breaking the data down into subsets and sequentially adding them to the history match is a novel approach which allows for further scientific insight. Most prominently, it not only allows inferences to be made about the system quantities associated with the model input parameters, but also provides insight into the links between quantities associated with both the input and output. Note that adding model outputs sequentially in this way is different from bringing outputs in sequentially due to emulator capability (step 3 of the algorithm) (Vernon et al. 2010a). We will explore this in detail for the Arabidopsis model.

2.4 History matching and Bayesian inference

In this section we discuss how history matching is informative for aiding the understanding of physical systems described by a computer model, and can also be used in conjunction with alternative methods of analysis, such as a general Bayesian analysis,

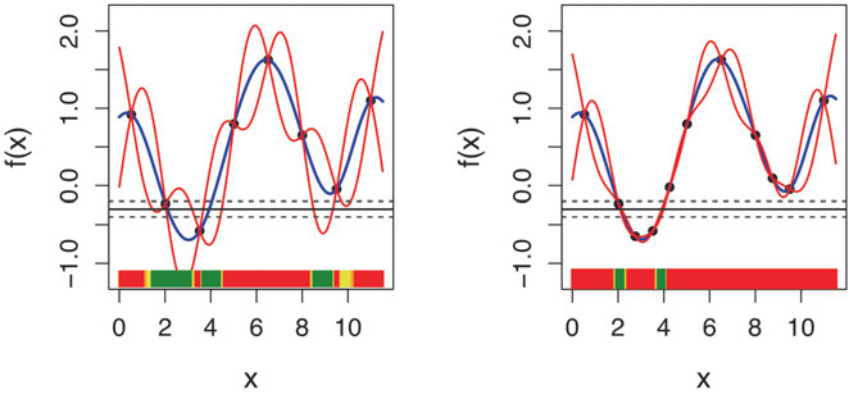

Figure 2: Left panel: Emulators for the simple 1D example $f(x) = 0.1x + \cos(x)$ as given by the left panel of Figure 1. The blue line represents the emulator's updated expectation $\mathrm{E}_D[f(x)]$ and the pair of red lines give the credible interval $\mathrm{E}_D[f(x)] \pm 3\sqrt{\mathrm{Var}_D[f(x)]}$. Observation z along with observed error are shown as horizontal black solid and dashed lines respectively. The implausibilities $I(x)$ are represented by the colours on the x-axis, with red representing high implausibility, orange and yellow representing borderline implausibility, and green representing low implausibility ($I(x) < 3$). Right panel: The wave 2 emulator for the same function, now including three additional wave 2 runs.

Bayesian optimisation and Approximate Bayesian Computation (ABC). In doing so we will compare some of the similar and differing features between history matching and the standard form of a general Bayesian analysis.

History matching is a computationally efficient and practical approach to identifying if a model is consistent with observed data, and, if so, utilising the key uncertainties within the problem to identify where in the input space acceptable matches lie (Craig et al. 1997). History matching is applicable for situations where the observed data is provided as either qualitative experimental trends or quantitative experimental measurements (or a combination of the two such as in the Arabidopsis model application in the following sections), attempting to answer some of the main questions that a modeller may have. In contrast, a general Bayesian framework typically requires full probabilistic specification of all uncertain quantities, providing a theoretically coherent method to obtain probabilistic answers to scientific questions. For example, in the context of a direct simulator, as given by Equations (6) and (7), a general Bayesian analysis will provide a posterior distribution for the location of the true best input x^*, whereas a history match provides a set, which may be empty, of points that could not implausibly be x^* under the posited assumption that x^* exists. History matching therefore has the benefit of avoiding the challenging task of making a full joint distributional specification over all uncertain quantities, to which resulting analyses may be sensitive.

Regardless of how prior distributions have been specified, performing the necessary calculations for a full Bayesian analysis is hard and typically computationally intensive, for example by making use of time-consuming numerical schemes such as Markov Chain Monte Carlo (MCMC) (Brooks et al. 2011). Many model evaluations are required to thoroughly explore the multi-modal likelihoods over the entire input space. Emulators can facilitate these large numbers of model evaluations at the cost of uncertainty (Higdon et al. 2008; Kennedy and O'Hagan 2001). However, since the likelihood function is constructed from all outputs of interest, we need to be able to emulate with sufficient accuracy all such outputs, including their possibly complex joint behaviour. Erratically behaved outputs may lead to emulators with large uncertainty or emulators which fail diagnostics. The consequential likelihood, and hence posterior, may be sensitive to the misspecification of these emulators.

Another related issue is that the posterior distribution of a general Bayesian analysis may be concentrated over a very small subspace of the initial input domain of interest $\mathscr{X}_0$. This is particularly true for models with relatively high-dimensional parameter spaces, and for which the initial ranges of interest for the parameters (that is, those defining $\mathscr{X}_0$) are large, such as for the Arabidopsis model discussed in the following sections. In such cases, a sufficiently accurate emulator over the whole input space will still require far too many model evaluations, hence alternative approaches to performing analysis, often sequential in nature and also involving the use of emulators, must be used. A variety of approaches are proposed in the literature. One group of such approaches is to use more sophisticated iterative MCMC algorithms, such as Population MCMC (Mohamed et al. 2012). Another aim in such a setting is Bayesian optimisation, using measures such as Expected Improvement and Expected Quantile Improvement to find points in the input space which result in minimised distance between model output and observed data (Forrester 2010; Picheny et al. 2013). Alternatively, there are several approaches drawing on the popular field of ABC (Smith and Gelfand 1992; Wilkinson 2013), such as ABC-SMC (Sequential Monte Carlo, Toni et al. (2009)), Bayes linear adjustment based ABC (Nott et al. 2014) and Rare Event ABC (Prangle et al. 2018). Whilst such approaches can effectively handle intractable likelihoods, finding very small regions in relatively high-dimensional spaces can still be challenging.

History matching is designed to efficiently cut out the uninteresting parts of the input space, thus allowing more accurate emulators to be constructed over the region of interest $\mathscr{X}^*$, where the vast majority of the mass of the posterior distribution should lie. A more detailed analysis, whether this be general Bayesian or an alternative such as ABC, can then be performed within this much smaller volume of input space, where the probabilistic specifications can now be considered more carefully. For example, Wilkinson (2014) performs a history match as a precursor to ABC. Although in that paper the likelihood function is the subject of the history match, rather than the model outputs themselves, it nevertheless highlights the benefits of carrying out a history match prior to performing a more detailed analysis. Directly emulating the likelihood may have computational advantages by reducing the emulated output dimension, which in some cases may reduce the number of required emulators substantially (although crucially probably not the number of model runs). However, the likelihood may exhibit complex behaviour which is difficult to emulate, particularly if there are several erratically behaved outputs. For discussions comparing History Matching and ABC, see, for example, Holden et al. (2018) and McKinley et al. (2018). For further information about how history matching can fit into a Bayesian paradigm, we refer the interested reader to Vernon et al. (2018), and also Wang et al. (2018), where history matching is used to help elicit reasonable prior choices from expert-elicited information.

To conclude this section, we suggest that history matching can be seen as both a useful precursor (rather than a competitor) to a more detailed analysis, and also as a form of analysis in its own right, particularly for modellers who do not wish to make the more detailed specifications required for a general Bayesian analysis. This latter situation is the motivation for this work, and we will discuss at length the insights that can be gained from having sequentially history matched the Arabidopsis model. Whilst the discussion around the results relates to the Arabidopsis model in particular, the history matching measures used within the discussion are generic and can be utilised in the context of history matching in many different fields of application.

3 Application of methods to Arabidopsis model

We now describe the relevant features of the hormonal crosstalk model as constructed by Liu et al. (2013) for applying the methods discussed in Section 2.

3.1 Model Structure

3.1.1 Description and network

The model represents the hormonal crosstalk of auxin, ethylene and cytokinin of Arabidopsis root development as a set of 18 differential equations, given in Table 1, which must be solved numerically. The model takes

Table 1: Arabidopsis model differential equations.

$$\frac{d[Auxin]}{dt} = \frac{k_{1a}}{1+\frac{[X]}{k_1}} + k_2 + k_{2a}\frac{[ET]}{1+\frac{[CK]}{k_{2b}}}\frac{[PLSp]}{k_{2c}+[PLSp]} + \frac{V_{IAA}[IAA]}{Km_{IAA}+[IAA]} - \left(k_3 + \frac{k_{3a}[PIN1pm]}{k3auxin+[Auxin]}\right)[Auxin]$$

$$\frac{d[X]}{dt} = k_{16} - k_{16a}[CTR1^*] - k_{17}[X]$$

$$\frac{d[PLSp]}{dt} = k_8[PLSm] - k_9[PLSp]$$

$$\frac{d[Ra]}{dt} = -k_4[Auxin][Ra] + k_5[Ra^*]$$

$$\frac{d[Ra^*]}{dt} = k_4[Auxin][Ra] - k_5[Ra^*]$$

$$\frac{d[CK]}{dt} = \frac{k_{18a}}{1+\frac{[Auxin]}{k_{18}}} - k_{19}[CK] + \frac{V_{CK}[cytokinin]}{Km_{CK}+[cytokinin]}$$

$$\frac{d[ET]}{dt} = k_{12} + k_{12a}[Auxin][CK] - k_{13}[ET] + \frac{V_{ACC}[ACC]}{Km_{ACC}+[ACC]}$$

$$\frac{d[PLSm]}{dt} = \frac{k_6[Ra^*]}{1+\frac{[ET]}{k_{6a}}} - k_7[PLSm]$$

$$\frac{d[Re]}{dt} = k_{11}[Re^*][ET] - (k_{10} + k_{10a}[PLSp])[Re]$$

$$\frac{d[Re^*]}{dt} = -k_{11}[Re^*][ET] + (k_{10} + k_{10a}[PLSp])[Re]$$

$$\frac{d[CTR1]}{dt} = -k_{14}[Re^*][CTR1] + k_{15}[CTR1^*]$$

$$\frac{d[CTR1^*]}{dt} = k_{14}[Re^*][CTR1] - k_{15}[CTR1^*]$$

$$\frac{d[PIN1m]}{dt} = \frac{k_{20a}}{k_{20b}+[CK]}[X]\frac{[Auxin]}{k_{20c}+[Auxin]} - k_{1_v21}[PIN1m]$$

$$\frac{d[PIN1pi]}{dt} = k_{22a}[PIN1m] - k_{1_v23}[PIN1pi] - k_{1_v24}[PIN1pi] + \frac{k_{25a}[PIN1pm]}{1+\frac{[Auxin]}{k_{25b}}}$$

$$\frac{d[PIN1pm]}{dt} = k_{1_v24}[PIN1pi] - \frac{k_{25a}[PIN1pm]}{1+\frac{[Auxin]}{k_{25b}}}$$

$$\frac{d[IAA]}{dt} = 0$$

$$\frac{d[cytokinin]}{dt} = 0$$

$$\frac{d[ACC]}{dt} = 0$$

Table 2: The list of 18 original model outputs, along with their initial conditions. The values of 0 or 1 for IAA, cytokinin and ACC correspond to no feeding or feeding of Auxin, Cytokinin or Ethylene respectively. See Liu et al. (2010) and Liu et al. (2013) for details.

Output	Initial concentration	Output	Initial concentration
Auxin	0.1	*Re∗*	0.3
X	0.1	*CTR*1	0
PLSp	0.1	*CTR*1∗	0.3
Ra	0	*PIN*1*m*	0
Ra∗	1	*PIN*1*pi*	0
CK	0.1	*PIN*1*pm*	0
ET	0.1	*IAA*	0 or 1
PLSm	0.1	*Cytokinin*	0 or 1
Re	0	*ACC*	0 or 1

an input vector of 45 rate parameters (k_1, k_{1a}, k_2,…) and produces an output vector of 18 chemical concentrations ([*Auxin*], [*X*], [*PLSp*],…). Note that, for simplicity, we refer to all components of the model, including hormones, proteins and mRNA, as "chemicals". Experiments accumulated over many years have established certain relationships between some of the 18 concentrations. For example, either manipulation of the PLS gene or exogenous application of IAA (a form of auxin), cytokinin or ACC (ethylene precursor) affects model outputs [*Auxin*], [*CK*], [*ET*] and [*PIN*]. We use initial conditions for the model, given in Table 2, that are consistent with Liu et al. (2010, 2013).

The hormonal crosstalk network of auxin, cytokinin and ethylene for Arabidopsis root development is shown in Figure 3. The auxin, cytokinin and ethylene signalling modules correspond to the model of Liu et al. (2010). The PIN functioning module is the additional interaction of the PIN proteins introduced in Liu et al. (2013). Solid arrows represent conversions whereas dotted arrows represent regulations. The v_i represent reactions in the biological system and link to the rate parameters k_i on the right hand side of the equations in Table 1. For full details of the model see Liu et al. (2013).

3.1.2 Mutants and feeding

We will be interested in comparing the differences in chemical concentrations (corresponding to model outputs [*Auxin*], [*CK*], [*ET*], [*PLSm*] and [*PIN*]) for different mutants (wild type (*WT*), *pls* mutant, PLS overexpressed transgenic (*PLSox*), ethylene insensitive *etr1*, double mutant *plsetr1*) and feeding regimes (no feeding f_0, feeding auxin f_a, feeding cytokinin f_c, feeding ethylene f_e, feeding any combination of these hormones f_af_c, f_af_e, f_cf_e and $f_af_cf_e$) of Arabidopsis (Liu et al. 2013). Note that *WT* refers to the typical plant occurring naturally in the wild that has not been mutated, however, we include this unmutated option in the list of mutants for notational convenience. Note that for simplicity of terminology, exogenous application of IAA, cytokinin or ACC is referred to as "feeding auxin, cytokinin or ethylene" respectively.

In the model, mutant type is controlled by altering the parameters representing the expression of the two genes *PLS* and *ETR1*. Input rate parameter k_6 controls the amount to which PLS is suppressed, hence *pls* is represented by setting $k_6 = 0$ and *PLSox* is represented by increasing the size of k_6 to a value greater than that of the WT plant. Input rate parameter k_{11} represents the rate of conversion of the active form of the ethylene receptor to its inactive form. The ethylene insensitive *etr1* mutant is represented by decreasing the size of k_{11} to a much smaller value than that of WT. *plsetr1* is represented by both setting $k_6 = 0$ and k_{11} to its much decreased value. Feeding regime is represented by the initial conditions of certain outputs. [*IAA*], [*cytokinin*] and [*ACC*] take initial condition values 0 or 1, as indicated in Table 2, depending on whether or not the respective chemical auxin, cytokinin or ethylene has been fed.

3.1.3 Model structure and the inputs

Model structure sometimes restricts what we are able to learn about certain parameter relationships. For example, in this case, there is a constraint that $k_{16}/k_{16a} = 0.3$, which ensures that the term $k_{16} - k_{16a}[CTR1^*]$ in

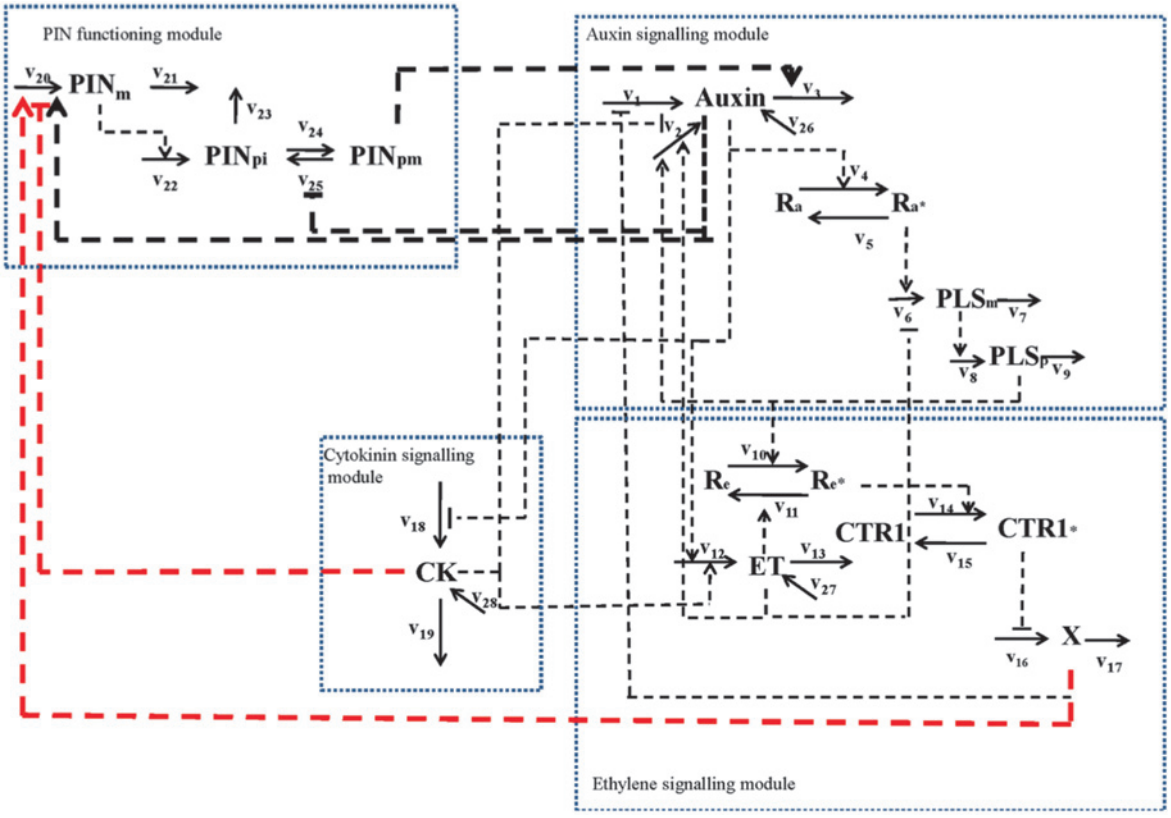

Figure 3: The Arabidopsis model network for the interaction of PIN, PLS and hormonal crosstalk. The auxin, cytokinin and ethylene signalling modules correspond to the model of Liu et al. (2010). The PIN functioning module is the additional interaction of the PIN proteins introduced in Liu et al. (2013). Solid arrows represent conversions whereas dotted arrows represent regulations. The v_i represent reactions in the biological system and link to the rate parameters k_i on the right hand side of the differential equations in Table 1.

the $d[X]/dt$ equation is non-negative, thus effectively removing an input from the set of rate parameters in the equations in Table 1. In principle, given sufficient runs, history matching should discover such restrictions in the model (for example, this restriction was identified for a simpler model via history matching in Vernon et al. (2018)), but the ability to identify these restrictions before we start will make the process more efficient.

In addition to this restriction, we are only interested in comparing the model output to data at equilibrium, thus allowing a substantial dimensional reduction of the input space. At equilibrium, the derivatives on the left hand side of the model equations given in Table 1 will equal zero, and hence the right hand side can be rearranged in terms of one less parameter (Vernon et al. 2018). For this reason, measurements of outputs of this system will only allow us to learn about certain ratios of the input rate parameters to one another. For example, the equation for $d[PLSp]/dt$ becomes

$$\frac{d[PLSp]}{dt} = 0 = k_8[PLSm] - k_9[PLSp] \tag{9}$$

$$\Rightarrow \quad 0 = \frac{k_8}{k_9}[PLSm] - [PLSp] \tag{10}$$

which only depends on the ratio k_8/k_9.

Another restriction arises from the fact that the initial conditions for the feeding chemicals $[IAA]$, $[cytokinin]$ and $[ACC]$ can only take the values 0 or 1 and then remain constant. This is because, although the

expressions $\frac{V_{IAA}[IAA]}{Km_{IAA}+[IAA]}$, $\frac{V_{CK}[cytokinin]}{Km_{CK}+[cytokinin]}$ and $\frac{V_{ACC}[ACC]}{Km_{ACC}+[ACC]}$ respectively in the equations for [*Auxin*], [*CK*] and [*ET*] take the specific form following the biological mechanism, they can only be learnt about as a whole, essentially comparing the case of a constant reservoir of chemical being available for uptake into the plant with the case of no feeding at all. Feeding of IAA, cytokinin and ACC with any concentration can be rescaled to $[IAA] = 1$, $[cytokinin] = 1$, and $[ACC] = 1$ by adjusting the parameters V_{IAA}, V_{CK}, and V_{ACC} in each equation respectively. Note that specific equations for the rate of change of the feeding chemicals may allow more insight into the effects of feeding if deemed biologically relevant.

Following the previous section, we let k_{6w} and k_{11w} represent the values that k_6 and k_{11} respectively should take for WT. We let the two additional parameters $k_{6m} > 1$ and $k_{11m} \ll 1$ represent the values these parameters should be multiplied by in order to obtain the corresponding model run for the *PLSox* and *etr1* mutants respectively, that is with $k_6 = k_{6m}k_{6w}$ and $k_{11} = k_{11m}k_{11w}$. Doing this allows exploration of a reasonable class of representations of these mutants using independent parameters.

Finally, since we consider ranges of rate parameters and rate parameter ratios which are always positive, many spanning many orders of magnitude, we choose to convert them to a log scale. We therefore define the reduced 31 dimensional vector of input parameters for the model to be:

$$x = \log(k_1,\ k_{1a}/k_2\,, k_{2a}/k_2\,, \ldots,\ k_{11m}) \tag{11}$$

as given in the left hand column of Table 3.

Table 3: A table of parameter ranges (which were converted to [−1, 1] for the analysis). These define the initial search region $\mathscr{X}_0$.

Input rate Parameter (Ratio)	Initial value	Minimum	Maximum
k_1	1	0.1	10
k_{1a}/k_2	5	0.5	50
k_{2a}/k_2	14	1.4	140
k_{2b}	1	0.1	10
k_{2c}	0.01	0.000001	0.1
k_3/k_2	10	1	100
k_{3a}/k_2	2.25	0.225	22.5
k_{3auxin}	10	1	100
k_5/k_4	1	0.1	10
k_{6a}	0.2	0.002	2000
k_{6w}/k_7	0.3	0.03	3
k_9/k_8	1	0.1	10
k_{10a}/k_{10}	16600	166	16600
k_{11}/k_{10}	16600	16.6	166000
k_{12a}/k_{12}	1	0.1	10
k_{13}/k_{12}	10	1	1000
k_{15}/k_{14}	0.0283	0.000283	0.283
k_{17}/k_{16a}	0.1	0.01	1
k_{19}/k_{18a}	1	0.1	10
k_{18}	0.1	0.01	10
k_{20a}/k_{1_v21}	0.8	0.08	8
k_{20b}	1	0.1	10
k_{20c}	0.3	0.03	3
k_{22a}/k_{1_v23}	1.35	0.135	13.5
k_{25a}/k_{1_v24}	0.1	0.01	1
k_{25b}	1	0.1	10
$V_{IAA}/k_2(Km_{IAA}+1)$	2.27	0.05	50
$V_{CK}/k_{18a}(Km_{CK}+1)$	0.45	0.01	1
$V_{ACC}/k_{12}(Km_{ACC}+1)$	4.55	0.1	100
k_{6m}	1.5	1	4
k_{11m}	0.006	0.001	0.1

In order to perform a full analysis on the model, we introduce a parameter $\lambda = V_i/V_m$ to represent the ratio of the cytosolic volume V_i to the volume of the cell wall V_m. Full details of why we introduce this parameter are included in Appendix A. For a typical cell, we fixed $\lambda = 6$ and considered that a reasonable range of possible values for λ was [2, 16] for a plant root cell.

3.2 Eliciting the necessary information for history matching

To perform a history match, we need to understand how real-world observations relate to model outputs, thus aiding the specification of observed values z_i, model discrepancy terms σ_{e_i} and measurement error terms σ_{e_i}. History matching is a versatile technique which can deal with observations of varying quality, such as we have for the Arabidopsis model.

3.2.1 Relating observations to model outputs

Each Arabidopsis model output relating to a biological experiment can be represented by:

$$h_{j,m,a}(x, t)$$

where:

$$j \in \{[Auxin], [PLSm], [CK], [ET], [PIN]\}$$
$$m \in \{WT, pls, PLSox, etr1, plsetr1\}$$
$$a \in \{f_0, f_a, f_c, f_e, f_af_c, f_af_e, f_cf_e, f_af_cf_e\}$$

Here, the subscript j indexes the measurable chemical, m indexes the plant type and a indexes the feeding action, where f_0 indicates no feeding and f_a, f_c and f_e indicate the feeding of auxin, cytokinin and/or ethylene respectively, for a particular set-up of the general model h (the Arabidopsis model equations given in Table 1). The vector x represents the vector of rate parameter ratios and t represents time. There are 200 possible experiments given by the possible combinations of j, m and a.

The average PIN concentration in both the cytosol and the cell wall is calculated as follows:

$$[PIN] = \frac{[PIN1pm] + \lambda[PIN1pi]}{1 + \lambda} \tag{12}$$

We collected the results of a subset of 32 of the possible experiments from a variety of experiments in the literature (see Liu et al. (2010, 2013) and references therein for details). 30 of these observations are measures of the trend of the concentration of a chemical for one experimental condition relative to another experimental condition (usually chosen to be wild type). We therefore need our outputs of interest to be ratios of the outputs of our model h with different experimental subscript settings. We choose to work with log model outputs since these will be more robust and allow multiplicative error statements. Since we only consider model outputs to be meaningful at equilibrium, that is as $t \to \infty$, we therefore, following Vernon et al. (2018), define the main outputs of interest to be:

$$f_i(x) = \lim_{t\to\infty} \log\left\{\frac{h_{j,m_2,a_2}(x, t)}{h_{j,m_1,a_1}(x, t)}\right\} \tag{13}$$

where the subscript i indexes the combinations of $\{j, m_1, a_1, m_2, a_2\}$ that were actually measured. This function $f_i(x)$ will be directly compared to the observed trends. All but one of the trends were relative to WT with no feeding, with the exception being the ratio of auxin concentration in the *pls* mutant fed ethylene to the *pls* mutant without feeding. The remaining two observations are non-ratio WT measurements of the chemicals [*Auxin*] and [*CK*]. The outputs of interest for these observations are given as $\lim_{t\to\infty} \log\{h_{[Auxin],WT,f_o}(x, t)\}$ and

$\lim_{t\to\infty} \log\{h_{[CK],WT,f_o}(x, t)\}$ respectively. Including these experiments within the history match ensures that acceptable matches will not have unrealistic concentrations of auxin and cytokinin.

The full list of 32 outputs is given in the left hand column of Table 4. These are notated in the form:

$$mutant\,(\text{if not wild type})\_feeding\,(\text{if any})\_chemical \tag{14}$$

and are assumed to be ratios relative to WT with no feeding unless otherwise specified. NR indicates that an output is not a ratio. For example, *fe_CK* indicates the cytokinin concentration ratio of WT fed ethylene relative to WT no feeding, and *PLSox_ET* represents the ethylene concentration ratio of the POLARIS overexpressed mutant relative to WT.

We sequentially history match the Arabidopsis model to these experimental observations in 3 phases *A*, *B* and *C*, with the group to which each experiment belongs presented in Table 4. We will history match the Dataset *A* observations to obtain a non-implausible set $\mathscr{X}_A$. Additional insight will be gained by further history matching to Dataset *B* to obtain $\mathscr{X}_B$, and then finally history matching to Dataset *C*. Dataset *B* contains the outputs involving the feeding of ethylene. History matching this group separately provides insight into how the

Table 4: A table showing the natural ranges and logarithmic ranges of simulator output values that would be accepted at implausibility cutoff 3. Column 2 shows which of the three Datasets each output belongs to. These outputs are notated in the form *mutant*(if not wild type)*_feeding*(if any)*_chemical* and are assumed to be ratios of the output for the specified mutant relative to that for wild type with no feeding unless otherwise specified. NR indicates that an output is not a ratio, and * indicates that the data for that experiment was a general trend.

Experiment	Dataset	Minimum log ratio value	Maximum log ratio value	Minimum ratio value	Maximum ratio value
WT_Auxin (NR)	A	−3.772	0.833	0.023	2.3
pls_Auxin	A	−1.531	0.366	0.216	1.442
PLSox_Auxin	A	−0.576	0.708	0.562	2.031
*etr1_Auxin**	A	0.182	2.303	1.2	10
plsetr1_Auxin	A	−0.792	0.600	0.453	1.823
f_a*_Auxin**	A	0.182	2.303	1.2	10
f_c*_Auxin*	A	−2.303	1.099	0.1	3
f_e*_Auxin**	B	0.182	2.303	1.2	10
$pls\_f_e\_Auxin/pls\_Auxin$	B	−1.204	−0.010	0.3	0.99
WT_CK (NR)	A	−3.730	0.875	0.024	2.4
pls_CK	A	0.049	1.253	1.05	3.5
*PLSox_CK**	A	−2.303	−0.182	0.1	0.834
f_a*_CK**	A	−2.303	−0.182	0.1	0.834
f_c*_CK**	A	0.182	2.303	1.2	10
f_e*_CK**	B	−2.303	−0.182	0.1	0.834
*pls_ET**	A	−0.342	0.336	0.71	1.4
*PLSox_ET**	A	−0.342	0.336	0.71	1.4
f_a*_ET**	A	0.182	2.303	1.2	10
f_c*_ET**	A	0.182	2.303	1.2	10
f_e*_ET**	B	0.182	2.303	1.2	10
f_a*_PLSm**	C	0.182	2.303	1.2	10
f_c*_PLSm**	C	−2.303	−0.182	0.1	0.834
f_e*_PLSm**	C	−2.303	−0.182	0.1	0.834
f_af_c*_PLSm*	C	−0.554	3.449	0.575	31.482
f_af_e*_PLSm*	C	0.207	3.315	1.23	27.528
pls_PIN	A	−0.650	1.007	0.522	2.738
PLSox_PIN	A	−1.629	0.456	0.196	1.578
etr1_PIN	A	−1.892	0.182	0.151	1.199
plsetr1_PIN	A	−1.175	0.613	0.309	1.846
f_a*_PIN**	A	0.182	2.303	1.2	10
f_c*_PIN**	A	−2.303	−0.182	0.1	0.834
f_e*_PIN*	B	−0.730	0.893	0.482	2.443

inputs of the model are constrained based on physical observations of a plant having been fed ethylene relative to its WT counterpart. Dataset *C* contains the outputs involving the measurement of *PLSm*, thus demonstrating how useful observing the effects of the POLARIS gene function were for gaining increased understanding about the model and its rate parameters.

3.2.2 Observed value, model discrepancy and measurement error

Although some of our collected measurement values were estimated values of a trend or ratio, many of the measurements were only general trend directions or estimated ranges for the ratio value, given with various degrees of accuracy (Liu et al. 2013). We therefore use a level of modelling appropriate to the nature of the data to propose order of magnitude estimators for z_i, σ_{ϵ_i} and σ_{e_i} that are consistent with the observed trends and expert judgement concerning the accuracy of the model and the relevant experiments. Doing this demonstrates that we can apply our history matching approach to vague, qualitative data, whilst demonstrating the increased power of this analysis were we to have more accurate quantitative data for all the experiments.

A general trend of "Up", "Down" or "No Change" was collected for 17 of the experiments, these being indicated by an asterix in Table 4. Following the conservative procedure given in (Vernon et al. 2018), we specify $z_i = 1.24$, -1.24 and 0 and $\sigma_{c_i} = 0.35$, 0.35 and 0.061 for the "Up", "Down" and "No Change" trends respectively, where σ_{c_i} represents the combined model discrepancy and measurement error, that is $\sigma_{c_i} = \sqrt{\sigma^2_{\epsilon_i} + \sigma^2_{e_i}}$. These combined specifications have been chosen such that $z_i \pm 3\sigma_{c_i}$ represents a 20% to 10 fold increase for the "Up" trends, a 20% to 10 fold decrease for the "Down" trends, and a 40% decrease to 40% increase for the "No Change" trends. To avoid confusion, we here define a 20% decrease to imply that a 20% increase on the decreased value returns the original value. This specification conservatively captures the main features of the trend data, although more in-depth specification could be made if quantitative measurements were available across these outputs. We specify z_i to be in the middle of the logged ratio range. In this work we considered that the deficiencies in the model would be of a similar order of magnitude to the observed errors on the data. We therefore specify both model discrepancy and measurement error to be of equal size and satisfy the ratio intervals above.

For the remaining cases, the observed values z_i, model discrepancies σ_{ϵ_i} and measurement errors σ_{e_i} were chosen using a more in-depth expert assessment of the accuracy of the relevant trend measurements and their links to the model (see Liu et al. (2010, 2013) and references therein for details). Since we will use a maximum implausibility threshold of $c = 3$ by appealing to Pukelsheim's 3 sigma rule (Pukelsheim 1994) when working with the simulator runs, it is most appropriate to specify the logged ranges of $z_i \pm 3\sigma_{c_i}$, as these are the ranges which if a simulator run falls outside it will be classed as implausible. These ranges are specified in Table 4 in both logged and not logged form.

3.3 Input ratio ranges

The initial ranges of values for the 31 input parameters were chosen based on those in the literature (Liu et al. 2010) and further analysis of the model (Liu et al. 2013), and are shown in Table 3. Many of the input ranges were chosen to cover an order of magnitude either side of the single satisfactory input parameter setting found in Liu et al. (2010). Some parameters of particular interest were subsequently increased to allow a wider exploration of the input parameter space. This gave us a large initial input space $\mathscr{X}_0$ which was thought to be suitable for our purposes. The logged ratio ranges were all converted to the range [–1, 1] prior to analysis.

We now apply the technique of sequential history matching using Bayes linear emulation to the Arabidopsis model (Liu et al. 2013). Analysis of the results, after history matching to each of Datasets *A*, *B* and *C*, will involve consideration of the following:

- The volume reduction of the non-implausible input space (Vernon et al. 2018).
- Input plots of the non-implausible space (Vernon et al. 2018).
- The variance resolution of individual inputs and groups of inputs.

- Output plots of the non-implausible space (Vernon et al. 2018).
- The degree to which each output was informative for learning about each input.

3.4 Insights from initial simulator runs

A wave 1 set of 2000 training runs were designed using a maximin Latin hypercube design over the initial input space $\mathscr{X}_0$. Figure 4 shows the wave 1 output runs $f_i(x)$ for all 32 outputs considered. The targets for the history match, as given by the intervals $z_i \pm 3\sigma_{c_i}$ and the ranges in Table 4, are shown as vertical error bars. Black error bars represent Dataset *A* outputs, blue error bars represent Dataset *B* outputs and red error bars represent Dataset *C* outputs. Note that the measurements for Datasets *B* and *C* are shown for illustrative purposes, and in general may have been obtained at a later point in time to when the wave 1 model runs were performed. In this case, however, the model would still produce a value for all output components (both those with and without corresponding physical measurements).

Figure 4 gives substantial insight into the general behaviour of the model over the initial input space $\mathscr{X}_0$, for example informing us about model outputs that can take extreme values, for example, f_c_*Auxin*, f_c_*ET* and $f_a f_c$_*PLSm*. More importantly, the runs also inform us as to the class of possible observed data sets that the model could have matched, and hence gives insight into the model's flexibility. There exist outputs with constrained ranges. In particular, many outputs seem to be constrained to being either positive or negative, for example, the logged trend for *pls_CK* must be positive and that of *PLSox_CK* must be negative. If such constrained outputs, which are consequences of the biological structure of the model, are found to be consistent with observations, this provides (partial) evidence for the model's validity. Conversely, we may be concerned about an overly flexible model that was capable of reproducing any combination of positive or negative observed data values for outputs in this dataset. Specifically, we should doubt claims that such a model has been validated by comparison to this data, as it would have inevitably matched any possible data values and hence arguably may not contain much inherent biological structure at all.

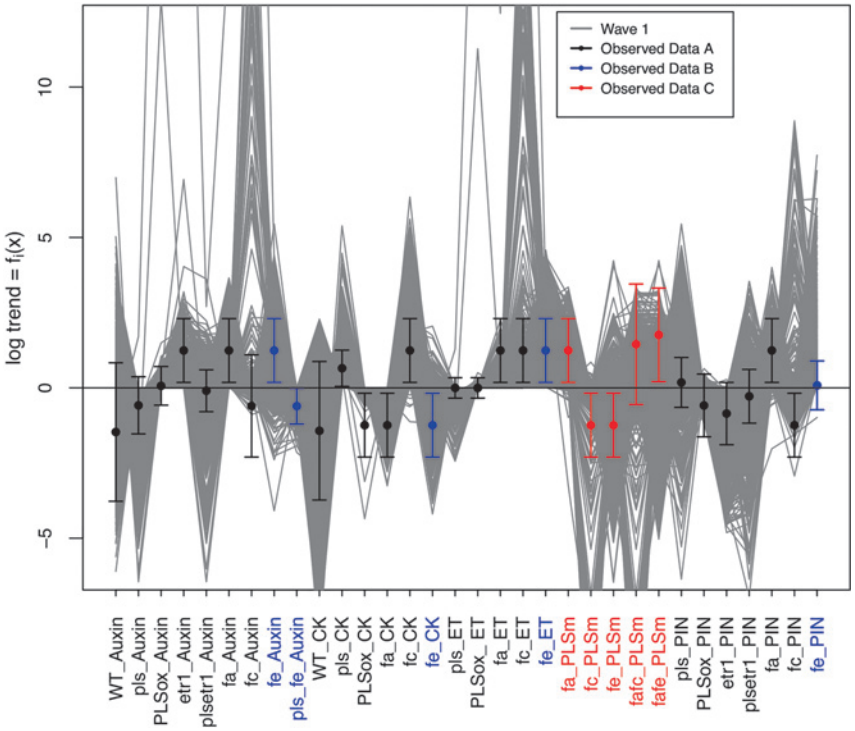

Figure 4: 2000 wave 1 output runs $f_i(x)$ for all 32 outputs considered. The targets for the history match, as given by the intervals $z_i \pm 3\sigma_{c_i}$ and the ranges in Table 4, are shown as vertical error bars. Black error bars represent Dataset *A* outputs, blue error bars represent Dataset *B* outputs and red error bars represent Dataset *C* outputs. The horizontal black line at zero represents zero trend.

There are some outputs for which the majority of the wave 1 runs already go through the corresponding error bars, for example *PLSox_Auxin* and *PLSox_ET*. This is an indication that these outputs did not help much to constrain the input space. Despite this, none of the wave 1 runs pass through all of the target intervals of the outputs in Dataset *A* simultaneously, thus already suggesting that the volume of the final non-implausible space would be small or indeed zero.

3.5 History matching the model

We outline the general decisions required to perform the history match. Several packages are available that perform standard Gaussian Process emulation, possibly with Automatic Relevance Determination (Neal 1997; Williams and Rasmussen 1996), for example the BACCO (Hankin 2005) and GPfit (MacDonald et al. 2015) packages in R (R Core Team) or GPy (since 2012) for Python, which may be used as an alternative to the emulators we describe here. Emulators accurate enough to reduce the size of the non-implausible space at each wave to some degree are sufficient for our purposes. When constructing emulators, we decided to put more detail into the mean function, but incorporate more complicated structures for the residual process at each wave, thus sequentially increasing the complexity of the emulators at each wave. We provide a summary of the choices made in the history match at each wave in Table 5, including the dataset history matched to (column 2), the number of design runs (column 3), the implausibility cut-off thresholds (columns 4–6) and the emulation strategy (column 7), each of which is discussed in more detail below.

The amount of space that was cut out after each wave is shown in Table 6. We let $V(\mathscr{X}_i)$ represent the volume of the non-implausible space after wave *i*, as judged by the emulators, and $V(\mathscr{X}_G)$ represent the volume of the space with acceptable matches to the observed data in Dataset *G*, as judged using actual model runs (hence without emulator error). Then columns 2 and 3 give the proportion of the previous wave and initial non-implausible spaces respectively still classed as non-implausible, and columns 5 and 6 give the proportion of the wave *i* and initial non-implausible spaces giving rise to actual acceptable matches to the data in Dataset *G*. The proportion of space cut out at each wave is influential for deciding the number of waves and emulator technique at each wave. In addition, Table 6 presents the radical space reduction obtained by performing the history match. A proportion of 6.1×10^{-7} of the original space was still considered non-implausible after history matching to Dataset *A*. A proportion of only 8.5×10^{-10} of the original space was still considered non-implausible after history matching to Datasets *A* and *B*, thus the 5 trends in Dataset *B*, for exogenous application of ACC, facilitated an additional reduction of 3 orders of magnitude. After all experimental observations had been matched to, the non-implausible space had been reduced to a proportion of 7.2×10^{-12} of the original space, thus the 5 trends in Dataset *C*, for measurement of POLARIS gene expression, refocused the set by another 2 orders of magnitude. Such small proportions of the original space being classed as non-implausible

Table 5: A summary of the wave-by-wave emulation strategy. Column 1: wave number. Column 2: Datasets history matched at wave *i*. Column 3: Number of model runs used to construct the emulator. Columns 4–6: Cutoff thresholds used at each wave for each of the implausibility critieria. Column 7: Emulation strategy for wave *i*.

Wave (*i*)	Dataset (*D*)	Runs	I_M^{cut}	I_{2M}^{cut}	I_{3M}^{cut}	Emulation Strategy
1	*A*	2000	–	3	2.9	Linear models
2	*A*	2000	3	2.9	–	Linear models
3, 4	*A*	2000	3	2.8	–	Linear models
5	*A*	2000	3	2.8	–	Single fixed correlation length
6, 7	*A*	2000	3	2.8	–	Several correlation lengths per output
8, 9	*A*, *B*	2000	3	2.9	–	Linear models for Dataset B outputs only
10	*A*, *B*	2000	3	2.9	–	Single fixed correlation length
11	*A*, *B*	3500	3	2.9	–	Several correlation lengths per output
12	*A*, *B*, *C*	2000	3	2.9	–	Single fixed correlation length
13	*A*, *B*, *C*	3500	3	2.9	–	Several correlation lengths per output

Table 6: A summary of the space cut out by the 13 waves of emulation and additional space cut out by the simulators for each dataset. Column 2: proportion of previous wave's non-implausible space still classed as non-implausible. Column 3: proportion of original space still classed as non-implausible. Column 5: proportion of wave *i* non-implausible space giving rise to acceptable matches to the data in Dataset *G* using simulations. Column 6: proportion of original space giving rise to acceptable matches to the data in Dataset *G* using simulations.

Wave (*i*)	$\frac{V(\mathscr{X}_i)}{V(\mathscr{X}_{i-1})}$	$\frac{V(\mathscr{X}_i)}{V(\mathscr{X}_0)}$	Dataset (*G*)	$\frac{V(\mathscr{X}_G)}{V(\mathscr{X}_i)}$	$\frac{V(\mathscr{X}_G)}{V(\mathscr{X}_0)}$
1	0.45	4.5×10^{-1}			
2	0.12	5.4×10^{-2}			
3	0.035	1.9×10^{-3}			
4	0.25	4.7×10^{-4}			
5	0.12	5.7×10^{-5}			
6	0.15	8.5×10^{-6}			
7	0.55	4.7×10^{-6}	*A*	0.13	6.1×10^{-7}
8	0.25	1.2×10^{-6}			
9	0.11	1.3×10^{-7}			
10	0.55	7.1×10^{-8}			
11	0.15	1.1×10^{-8}	*A*, *B*	0.08	8.5×10^{-10}
12	0.1	1.1×10^{-9}			
13	0.45	4.8×10^{-10}	*A*, *B*, *C*	0.015	7.2×10^{-12}

means that acceptable runs within these spaces would likely be missed by more ad-hoc parameter searching methods of analysis.

Linear model emulators with uncorrelated residual processes were used in the initial waves since they are very cheap to evaluate, substantially more so even than emulators involving a correlated residual process, which may only be slightly more accurate (Andrianakis et al. 2017a). For these emulators, we estimated the value of $\sigma^2_{u_i}$ to be the estimated variance parameter from the linear model fit. As the amount of space being classed as implausible at each wave started to drop, we introduced emulators with a Gaussian correlation residual process. There are various methods in the literature for assessing correlation length parameters, as explained in Section 2.1. Some of the methods in the literature for picking the correlation lengths θ and variance parameter $\sigma^2_{u_i}$ tend to be computationally intensive and their result highly sensitive to the sample of simulator runs (Andrianakis and Challenor 2009, 2011). The choice was therefore made, at wave 5, to use a single correlation length parameter value of $\theta = 2$ for all input-output combinations, and to fit $\sigma^2_{u_i}$ using the corresponding linear model fit, these choices being checked using emulator diagnostics. The motivation for this choice of correlation length parameter was made by appealing to the heuristic argument made in Vernon et al. (2010a) that the regression residuals may be derived from a polynomial of order one higher than the fitted polynomials, the alteration in the chosen value taking into account the higher dimensionality of the input space.

At wave 6 the complexity of the residual process was increased still further by splitting the active inputs x_{A_i} for each output emulator into five groups based on similar strength of effect, and using maximum likelihood to fit the same correlation length to all inputs in each group, along with the variance parameter $\sigma^2_{u_i}$. This extension to the literature of fitting several different correlation length parameter values strikes a balance between the stability of the maximum likelihood process (which can become very challenging were we to include 31 separate correlation lengths for each of the 31 input components) and the overall complexity of the residual process. It should be noted that maximum likelihood makes use of probabilistic distributions to make an assessment of the correlation length parameters. Whilst it can be argued that such an approach lies outside of the Bayes linear paradigm in which we presented the construction of our emulators, it provides a useful tool for calculating adequate emulators which satisfy diagnostics. At wave 8 we introduced the Dataset *B* outputs by first using linear model emulators for the new outputs only, and then using emulators with residual correlation processes for all outputs. In waves 12 and 13 we incorporated emulators with residual processes for the Dataset *C* outputs.

The number of design points per wave was largely kept constant at 2000. 2000 was deemed a suitable number of runs per wave as it meant that the matrix calculations involved in the emulator were reasonable, whilst allowing adequate coverage of the non-implausible input space with simulator runs. At waves 11 and 13, 3500 design runs were used to build more accurate emulators.

In terms of design, a maximin Latin hypercube (Currin et al. 1991; McKay et al. 1979) was deemed sufficient for our needs as we required a simple and efficient space-filling design. The speed of the simulator meant that more structured and tactical designs were unnecessary for our requirements. At wave 1 we constructed a Latin hypercube of size 2000 to build emulators for each of the outputs. At waves 2–7 we first built a large maximin Latin hypercube design containing a large number of points over the smallest hyper-rectangle enclosing the non-implausible set. We then used all previous wave emulators and implausibility measures to evaluate the implausibility of all the proposed points in the design (Vernon et al. 2010a). Any points that did not satisfy the implausibility cut-offs were discarded from further analysis. If a single Latin hypercube was not sufficient to generate enough design points, multiple Latin hypercubes were taken in turn and the remaining points in each were taken to be the design. From wave 8 onwards an alternative sampling scheme was necessary to generate approximately uniform points from the non-implausible sets, since generating points using Latin hypercubes became infeasible due to the size of the non-implausible space. There are several alternative ways presented in the literature to approximately sample uniformly distributed points over the non-implausible space (Andrianakis et al. 2015, 2017a; Gong and DiazDelaO 2017; Williamson and Vernon 2013). We used a simple Metropolis-Hastings MCMC algorithm (Brooks et al. 2011), which provided adequate coverage of the non-implausible space.

At each wave we performed diagnostics on the mean function linear model, the emulator and the implausibility criteria using 200 points in the non-implausible space. The diagnostic test for implausibility which we used was the one described in Vernon et al. (2018). It compares the data implausibility cut-off criteria $I_M^{data}(x)$, that is the implausibility evaluated at a known diagnostic run, against the chosen implausibility cut-off criteria for the emulator outputs.

Many waves were necessary to complete this history matching procedure due to the complex structure of the Arabidopsis model. We see from the first few waves that linear model emulators are sufficient for learning a great deal about the input parameter space, but that including full emulators with correlated residuals at later waves can be useful. In addition, emulators have greatly increased the efficiency of our analysis over simply running the model. To make a comparison, the model itself takes 30 s to run on a standard laptop computer (not very slow, but slow enough to cause problems) in comparison to the tens of millions of runs per second for the linear model emulators and tens of thousands of runs per second for the more complex later-wave emulators. At the end of the procedure we obtained many runs satisfying each of the datasets A, B and C. We now go on to describe the results of the parameter search using various graphical representation techniques and discuss their biological implications.

4 Visual representation of history matching results

A major aim of this work is to evaluate how acceptable parameter combinations to a model can be assessed as a result of experimental measurements.

Figure 5 shows, below the diagonal, a "pairs" plot for a subset of the inputs x. A "pairs" plot shows the location of various points in the 31-dimensional input space projected down into 2-dimensional spaces corresponding to two of the inputs. For example, the bottom left panel shows the points projected onto the k_{1a}/k_2 vs k_{11m} plane. Inputs to wave 1 runs are given by grey points. Inputs to runs of the simulator with acceptable matches to the observed data in Datasets A, B and C are given as yellow, pink and green points respectively. Above the diagonal are shown 2-dimensional optical depth plots of inputs to runs with acceptable matches to all of the observed data for the same subset of the inputs. Optical depth plots show the depth or thickness of the non-implausible space in the remaining 29 dimensions not shown in the 2d projection (Vernon et al. 2010a, 2018). More formally, suppose we partition input x as $x = (x', x'')$, where x' is the two-dimensional vector

representing the parameters we wish to project onto, and x'' represents the remaining 29 parameters, then the optical depth plot is given by:

$$\rho(x') \propto V(x \in \mathscr{X}_C | x' \text{ fixed}) \tag{15}$$

where V here represents volume in the remaining 29 dimensions. The orientation of these plots has been flipped to be consistent with the plots below the diagonal. Along the diagonal are shown 1-dimensional optical depth plots.

Figure 5 provides much insight into the structure of the model and the constraints placed upon the input rate parameters by the data. Some of the inputs, such as k_{6a}, k_{18}, k_{19}/k_{18a} and $V_{ACC}/k_{12}(Km_{ACC}+1)$ are constrained even in terms of 1-dimensional range. Some inputs only appear constrained when considered in combination with other inputs, for example k_{11}/k_{10} and k_{13}/k_{12} exhibit a positive correlation. This is reasonable, since an increase in k_{11}, the rate constant for converting the activated form of ethylene receptor into its inactivated form, can be compensated by an increase in k_{13}, the rate constant for removing ethylene, since ethylene promotes the conversion of the activated form of ethylene receptor into its inactivated form. More complex constraints involving three or more inputs are more difficult to visualise. Below the diagonal, the pairs plot gives insight into which input parameters were learnt about by which set of outputs. For example, the parameter $V_{ACC}/k_{12}(Km_{ACC}+1)$ is largely learnt about by Dataset B, as is clear from the difference between the area of the yellow points and pink points in plots involving this input. This is not surprising, since this term corresponds to the feeding and biosynthesis (k_{12}) of ethylene, which we would expect to be learnt from the feeding ethylene experiments. We can see that input combinations with large values of k_{6a} are classed as implausible, thus constraining this input to be relatively low.

Figure 6 shows the output runs $f_i(x)$ corresponding to the input combinations shown in Figure 5 for all 32 outputs considered. The colour scheme is directly consistent with Figure 5, with wave 1 runs given as grey lines, and simulator non-implausible runs after history matching Datasets A, B and C given as yellow, pink and green lines respectively. Runs which pass within the error bar of a particular output i satisfy the constraint of being within $z_i \pm 3\sigma_{c_i}$, thus being in alignment with the results of the corresponding experimental observation, given our beliefs about model discrepancy and measurement error. Black error bars represent Dataset A outputs, blue error bars represent Dataset B outputs and red error bars represent Dataset C outputs.

Figure 6 gives much insight into joint constraints on possible model output values that are in alignment with all of the observed data (and so would pass through all of the error bars). Some model outputs have been

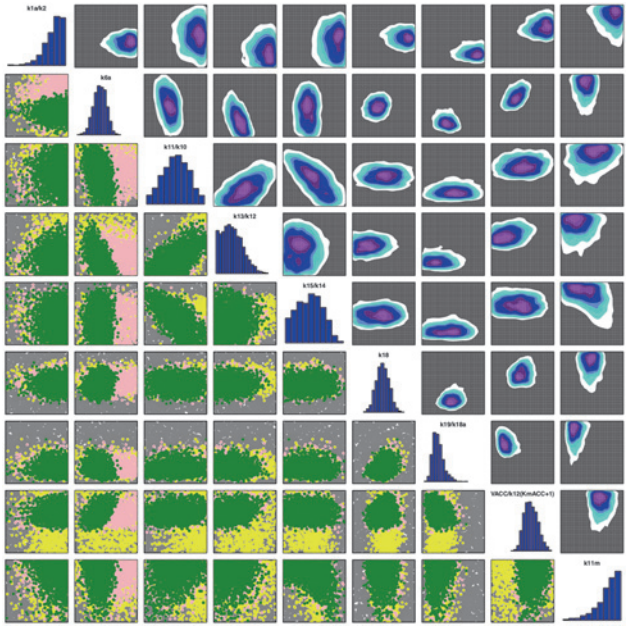

Figure 5: Below diagonal: A pairs plot for a subset of the inputs x. Inputs to wave 1 runs are given by grey points. Inputs to runs of the simulator with acceptable matches to the observed data in Datasets A, B and C are given as yellow, pink and green points respectively. Above diagonal: 2-dimensional optical depth plots of inputs to runs with acceptable matches to all of the observed data for the same subset of the inputs. The orientation of these plots has been flipped to be consistent with the plots below the diagonal. Along diagonal: 1-dimensional optical depth plots.

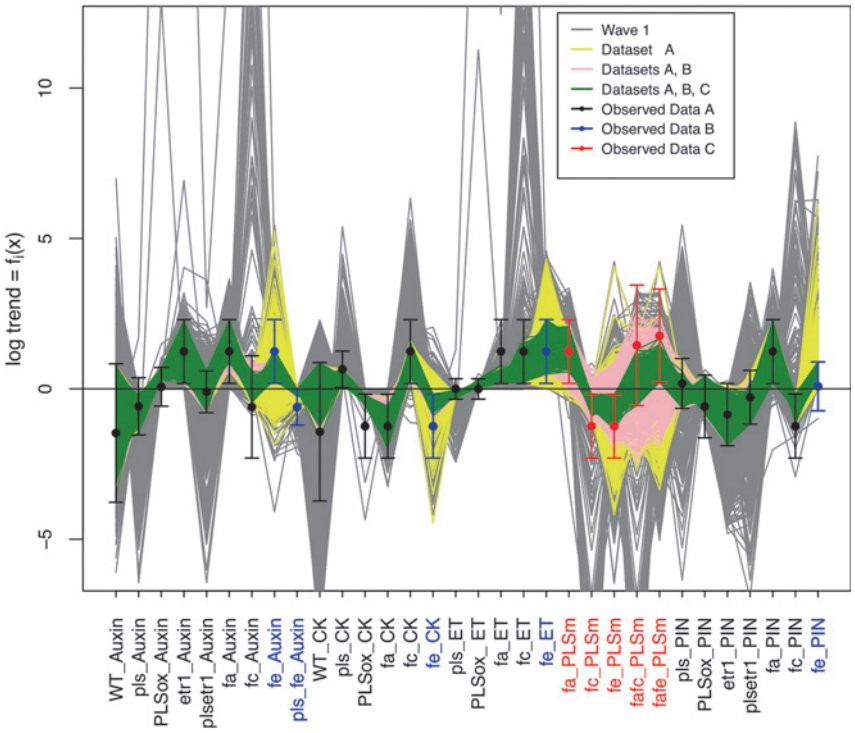

Figure 6: Output runs $f_i(x)$ for all 32 outputs considered. Wave 1 runs are given as grey lines. Simulator non-implausible runs after history matching Datasets *A*, *B* and *C* are given as yellow, pink and green lines respectively. The targets for the history match, as given by the intervals $z_i \pm 3\sigma_{c_i}$ and the ranges in Table 4, are shown as vertical error bars. Black error bars represent Dataset *A* outputs, blue error bars represent Dataset *B* outputs and red error bars represent Dataset *C* outputs. The horizontal black line at zero represents zero trend.

constrained much more than the range of their error bars, for example, *PLSox_Auxin* is constrained to the upper half of its error bar while f_c*_CK* is constrained to take smaller values. It is interesting that many of the yellow runs already go through the error bars of some of the outputs in Datasets *B* and *C*, for example *pls_*f_e*_Auxin* and f_a*_PLSm*. This indicates that the additional experimental observations corresponding to such outputs did not help to further constrain the input space.

Figure 7 presents the proportion of simulator runs at each wave which pass through the error bar of each output. Lower numbers for a particular output at a particular wave indicate that the output could be informative for learning more about the input parameter space. The two vertical black lines represent the waves

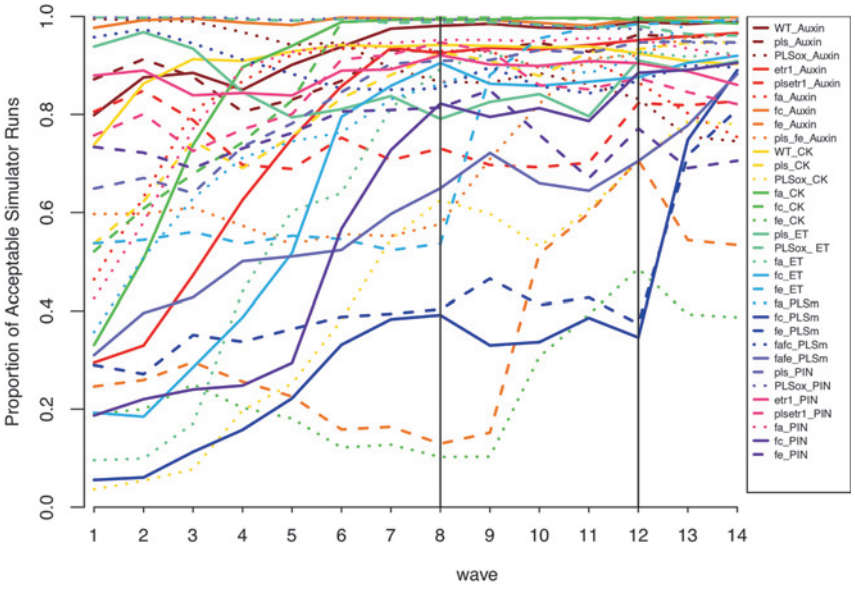

Figure 7: The proportion of simulator runs at each wave which pass through the error bar of each output. The two vertical black lines represent the waves where datasets B and C were incorporated.

where datasets B and C were incorporated. Some outputs, for example *PLSox_ET* and *PLSox_PIN*, had a high proportion (close to 1) of runs passing through their error bars at wave 1, in accordance with Figure 4. These outputs were not very informative for the history matching process. Some outputs, for example, *etr*1_*Auxin* and f_c_*PLSm*, had a low proportion (0.29 and 0.08 respectively) of runs passing through their error bars at wave 1, but a high proportion (over 0.8) after 13 waves of history matching. Space that would be classed as implausible by these simulator outputs became classed as implausible by the emulators during the waves of history matching. Some outputs, for example f_e_*Auxin* and f_e_*CK*, had relatively low proportions (less than 0.6) of runs passing through their error bars even at the end of the history matching procedure. This is indication that these outputs may have been difficult to emulate throughout.

As expected, we notice that the outputs in Datasets *B* and *C* start to have higher numbers of runs passing through their error bars once those outputs have been history matched to observations. Interestingly, as can also be detected in Figure 6, some of the outputs in Datasets *B* and *C*, for example $f_a f_e$_*PLSm*, get a surprisingly increased proportion (from 0.32 to 0.63) of runs passing through their error bars even before the output is incorporated into the history match. This is an indication that information from this output has already been learnt from observing some combination of the previously included outputs. There are a few components, most notably *PLSox_Auxin*, which had a high proportion of runs passing through their error bars before wave 1, but a much smaller proportion by the end. This is possible due to the joint constraints between the output components which involves non-implausible runs for this output component being classed as implausible by the constraints related to other output components. In addition, such behaviour is much less surprising if a particular output component was not included in the history match at early waves.

Although the overall proportion of space cut out is a very useful measure of the dependence of the model input parameter space on observed measurements, one may be interested in the degree to which specific parameters of particular interest have been constrained due to the observations. Sample variances of particular inputs in the non-implausible sets are a very informative and appropriate measure for this purpose as they take account of the density of the non-implausible space projected down onto the input dimensions of specific interest. Such measures are simple to calculate, and in many cases sufficient for our purposes. However, if we wanted to perform a full Bayesian analysis, we could appropriately re-weight the non-implausible points and recalculate these sample variances to obtain estimates of posterior (marginal) variances, provided we were confident enough to make all the additional assumptions that a full Bayesian analysis requires, as outlined in Section 2.4.

In Figure 8, sample variances (as a proportion of the original wave 1 sample variance) for each input of a sample of 2000 points with acceptable matches to the observed data in Datasets *A*, *B* and *C* are given by yellow, pink and green points respectively. Again, there is much insight to be gained from such a plot. We can see that different input ratios have been learnt about to different degrees by the observations of outputs in Datasets *A*, *B* and *C*. Some inputs are resolved well by Dataset *A* but then not really any further once Datasets *B* and *C* are additionally introduced. For example, k_1, representing inhibition of auxin transport by the ethylene downstream, *X*, is reduced by 0.43 by Dataset *A*, and then by less than 0.1 after both *B* and *C* have been additionally measured. This implies that experiments related to feeding ACC and measuring the PLS gene expression play a limited role in determining the parameter about inhibition of auxin transport by ethylene downstream. Some inputs are resolved slightly by Datasets *A* and *B*, and then substantially by Dataset *C*. For example, k_5/k_4, which governs the rate of conversion of auxin receptor from its active form *Ra* to its inactive form Ra^* and vice-versa, is reduced by less than 0.25 by Datasets *A* and *B*, and then by more than an additional 0.5 once Dataset *C* is measured. By analysing the model equations we see that [*Ra*] and [Ra^*] feature prominently in the [*PLSm*] equation, which is the output being measured in Dataset C. This indicates that measuring the *PLS* gene expression is important for determining the parameter relating to activation and inactivation of auxin receptor. Some inputs, for example k_{6a}, are learnt partially about by each dataset in turn, with overall high resolution. Some inputs have very little variance resolution at all. For example, $k_{22a}/k_{1,23}$, representing *PIN*1*m* translation to produce *PIN*1*pi*, has an approximate resolution of 0.1. Some information contained in Figure 8 may be quite intuitive, for example the fact that most of the variance resolution of $V_{ACC}/k_{12}(Km_{ACC}+1)$, the input corresponding to the feeding of ethylene, is obtained after measuring Dataset *B*. Checking that our results coincide

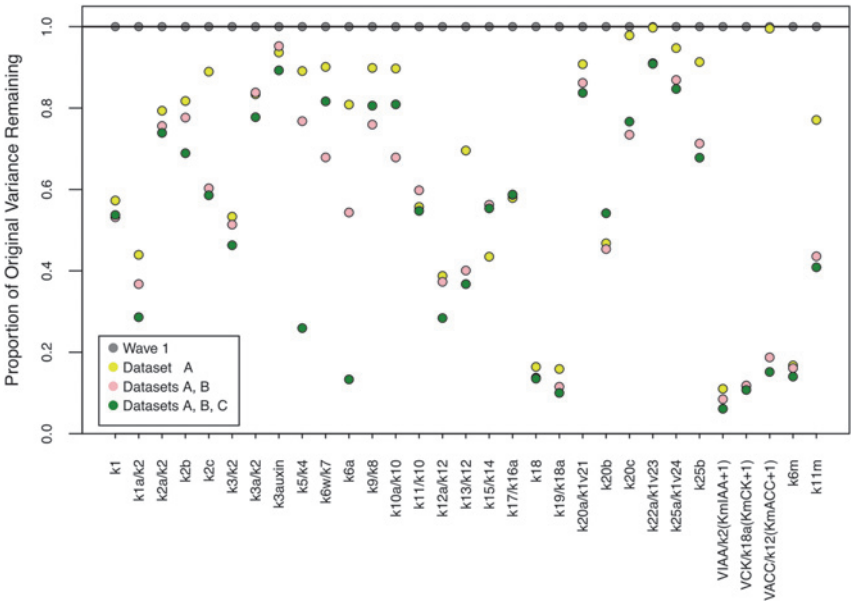

Figure 8: Sample variances (as a proportion of the original wave 1 sample variance) for each input of a sample of 2000 points with acceptable matches to the observed data in Datasets *A*, *B* and *C*, given by yellow, pink and green points respectively. The difference between the grey (wave 1) and yellow points shows the proportion of sample variance resolved by the Dataset A outputs. The differences between the yellow and pink, and pink and green points show the amount of additional space resolved (as a proportion of the original sample variance) by the Datasets B and C outputs respectively.

with this intuitive biological knowledge is an important diagnostic step, and provides evidence that our method has analysed the parameter space appropriately. Other information contained in Figure 8 is less intuitive and offers insight into the complex structure of the Arabidopsis model.

An analogous measure to space cut out in lower dimensions is range, area or volume reduction of the non-implausible space projected down onto the relevant input dimensions. These measures are far less informative than variance measures as they are very sensitive to extreme values, and it is not uncommon for the initial range of an input to be non-implausible in high dimensions. To get an idea of this, we compare Figures 8 and 9,

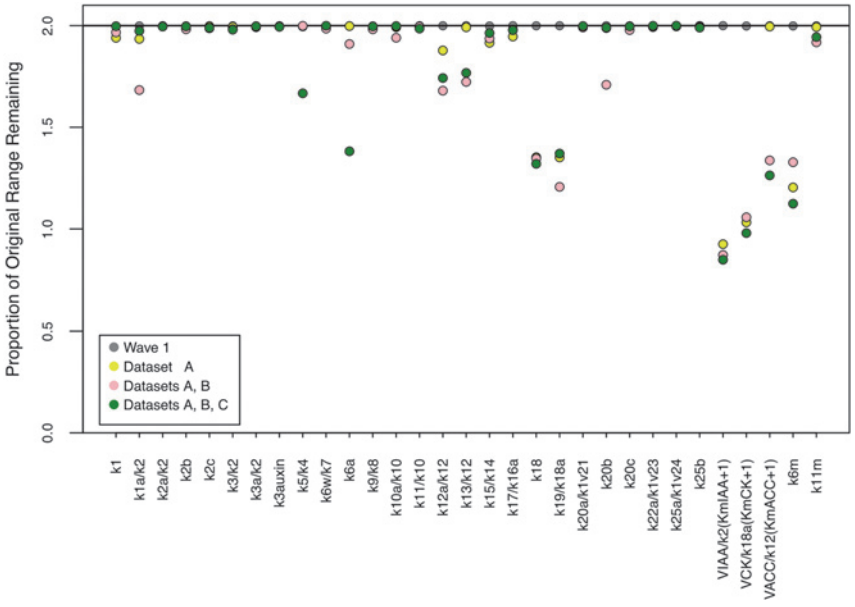

Figure 9: Sample ranges for each input of the runs used to build the wave 1 emulator are shown as grey points. Ranges for each input of a sample of 2000 points with acceptable matches to the observed data in Datasets *A*, *B* and *C* are given by yellow, pink and green points respectively.

which presents ranges for each input of a sample of runs used to build the wave 1 emulator as grey points, and ranges for each input of a sample of 2000 points with acceptable matches to the observed data in Datasets A, B and C as yellow, pink and green points respectively. We can see that certain inputs, for example k_{19}/k_{18a}, k_{6m} and in particular the feeding inputs $\frac{V_{IAA}[IAA]}{Km_{IAA}+[IAA]}$, $\frac{V_{CK}[cytokinin]}{Km_{CK}+[cytokinin]}$ and $\frac{V_{ACC}[ACC]}{Km_{ACC}+[ACC]}$, have their ranges significantly reduced. Many of the other inputs don't have their sample ranges reduced much at all. This does not necessarily mean that we don't learn about these inputs, just that for any specified value of one of these inputs there exists some combination of the remaining 30 inputs which can compensate, hence leading to a model output with an acceptable match to the observed data.

Simple measures involving the analysis of variance reduction or resolution can also be used to quickly describe joint constraints that alert us to strong relationships between inputs. Suppose we treat the vector of inputs as a multi-dimensional random variable W^u uniformly distributed over a non-implausible region $\mathscr{X}_u$, that is:

$$f_{W^u}(w^u) \propto \begin{cases} 1, & w^u \in \mathscr{X}_u \\ 0, & w^u \notin \mathscr{X}_u \end{cases}$$

Note that the uniform distribution is chosen here as we wish to treat all parts of the non-implausible set equally, as we may currently doubt the existence of a "true" best input x^*, and hence not want to perform a posterior re-weighting of the region $\mathscr{X}$. Given $f_{W^u}(w^u)$, we can calculate $\mathrm{Var}[W^u]$. Let us define the marginal variance for inputs $J = j_1,\ldots,j_m$ of random variable W^u corresponding to non-implausible space $\mathscr{X}_u$ as $\mathrm{Var}[W^u_J]$. We introduce the variance resolution measure for inputs J between non-implausible spaces $\mathscr{X}_u$ and $\mathscr{X}_v$ to be:

$$R_{uv}(\mathscr{X}_J) = 1 - \frac{\det\left(\mathrm{Var}\left[W^v_J\right]\right)}{\det\left(\mathrm{Var}\left[W^u_J\right]\right)} \tag{16}$$

We choose this measure as it relates to the product of the eigenvalues of the variance matrix and hence to a density-weighted volume of the projected non-implausible space, which is relevant for what we are interested in. We do not have exact distributions for $f_{W^u}(w^u)$ owing to not having an exact specification for $\mathscr{X}_u$. We therefore estimate $\mathrm{Var}[W^u]$ corresponding to $\mathscr{X}_u$ as $\mathrm{Var}[\mathscr{X}^s_u]$, where $\mathscr{X}^s_u$ is an (approximately) uniform sample of points from the non-implausible set $\mathscr{X}_u$.

Figure 10 shows sample variance resolutions $R_{0C}(\mathscr{X}^s_{j_1,j_2})$ between initial (0) and final (C) non-implausible spaces for each pair of inputs $J = j_1, j_2$, represented by colour, with red indicating high resolution and blue representing low resolution. Individual input variance resolutions, namely the difference between the initial grey and final green points in Figure 8, are represented along the diagonal. Note that an individual input resolution will never be greater than the joint variance resolution of that input with another one. We can see that learning jointly about k_{1a}/k_2 and k_{18}, namely those rate parameters representing auxin transport and biosynthesis, and regulation of cytokinin biosynthesis by auxin, is seemingly more informative than learning about either of the two parameters separately, in terms of variance resolution. As a converse example, we can see that little is learnt jointly between k_{3auxin} and $k_{22a}/k_{1,v23}$.

Although Figure 10 is informative, we really wish to determine the cases where the joint constraint on two input parameters is more severe than we would expect if we just assumed they were independently constrained. Assuming independence, the determinant of the sample variance matrix for inputs j_1, j_2 is the product of the sample variance for each input, that is:

$$\det\left(\mathrm{Var}\left[\mathscr{X}^s_{j_1,j_2}\right]\right) = \mathrm{Var}\left[\mathscr{X}^s_{j_1}\right]\mathrm{Var}\left[\mathscr{X}^s_{j_2}\right] \tag{17}$$

The standardised difference between the determinant assuming independent inputs and observed determinant is the squared correlation function:

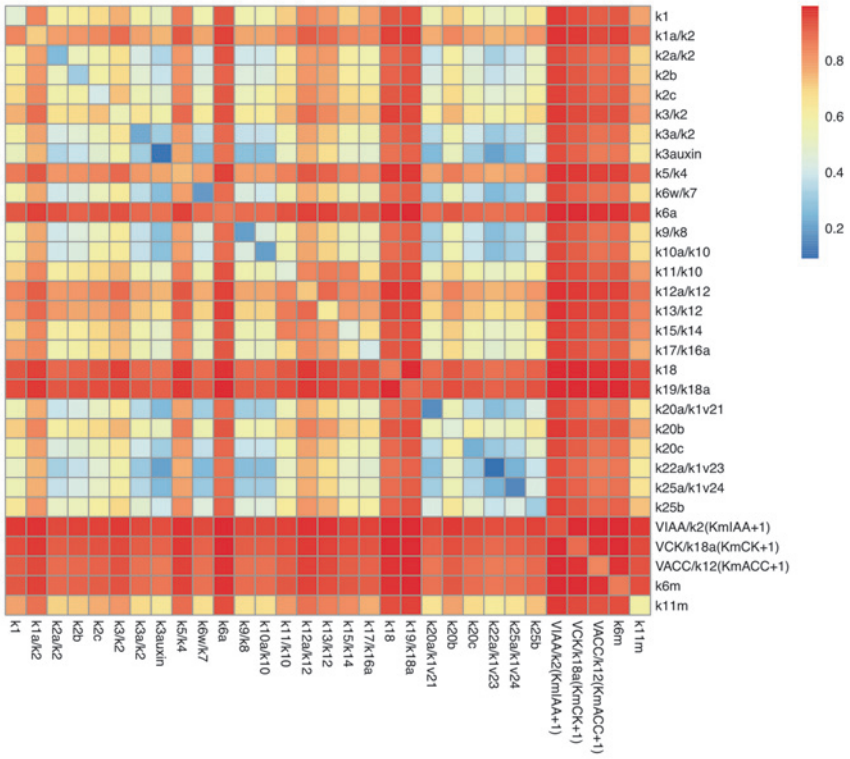

Figure 10: Sample variance resolutions $R_{0C}(\mathscr{X}^s_{j_1,j_2})$ between initial and final non-implausible spaces for each pair of inputs j_1, j_2, represented by colour. Individual input variances, corresponding to the difference between the grey points and green points in Figure 8, are represented along the diagonal. Red indicates high resolution whereas blue represents low resolution. For example, a large amount of variance has been jointly resolved by k_{6a} and k_5/k_4, as indicated by the red square at the intersection of the corresponding row and column.

$$\frac{\mathrm{Var}\left[\mathscr{X}^s_{j_1}\right]\mathrm{Var}\left[\mathscr{X}^s_{j_2}\right] - \det\left(\mathrm{Var}\left[\mathscr{X}^s_{j_1,j_2}\right]\right)}{\mathrm{Var}\left[\mathscr{X}^s_{j_1}\right]\mathrm{Var}\left[\mathscr{X}^s_{j_2}\right]} = \frac{\left(\mathrm{Cov}\left[\mathscr{X}^s_{j_1}, \mathscr{X}^s_{j_2}\right]\right)^2}{\mathrm{Var}\left[\mathscr{X}^s_{j_1}\right]\mathrm{Var}\left[\mathscr{X}^s_{j_2}\right]} \tag{18}$$

This is informative for the dependence, and hence level of constraint, between that pair of inputs in the final non-implausible set, as it alerts us to cases where there has been far more variance resolved jointly than that expected were the inputs just constrained independently. We therefore present these differences between each pair of inputs, represented by colour, in Figure 11. Red represents a larger difference and blue represents a smaller difference. The diagonal elements are necessarily zero. It would appear that there are stronger joint constraints between lots of pairs of inputs, most notably k_{11}/k_{10} and k_{15}/k_{14}, and k_3/k_2 and k_{18}. The first of these pairs, involving the CTR1 protein and ethylene receptor, has the most joint structure of any pair, with a squared sample correlation of 0.46. Since both the CTR1 protein and ethylene receptor take actions in the ethylene signalling module, they relay ethylene signalling. The parameters controlling this relay can be highly interdependent. Therefore, a change in one of these parameters can be compensated by a change in another. Interestingly, Figure 11 would indicate that there is little joint structure between k_{18} and k_{1a}/k_2, with a squared sample correlation of less than 0.05, indicating that the combined variance resolution between k_{1a}/k_2 and k_{18} presented in Figure 10 was not much larger than the independent product of the resolution of each of the individual inputs. Figure 11 can suggest interesting pairs of inputs to analyse in more detail, for example by examining their corresponding pairs plots, as shown in Figure 5.

Figure 12 provides a visualisation of how much each single output informed each input, represented by colour. These were calculated as the standardised difference between the sample variance of the input for all wave 1 runs and those wave 1 runs going through the output error bar. This quantity estimates the sample variance resolution for each input i were we to history match using only output j. Red indicates higher values of this estimated quantity and blue represents lower values. Figure 12 is very informative. We can see that some of the outputs, for example *PLSox_Auxin* and f_c*_Auxin*, are not directly informative about many of the inputs when considered on their own (however this does not preclude the possibility that they could still be

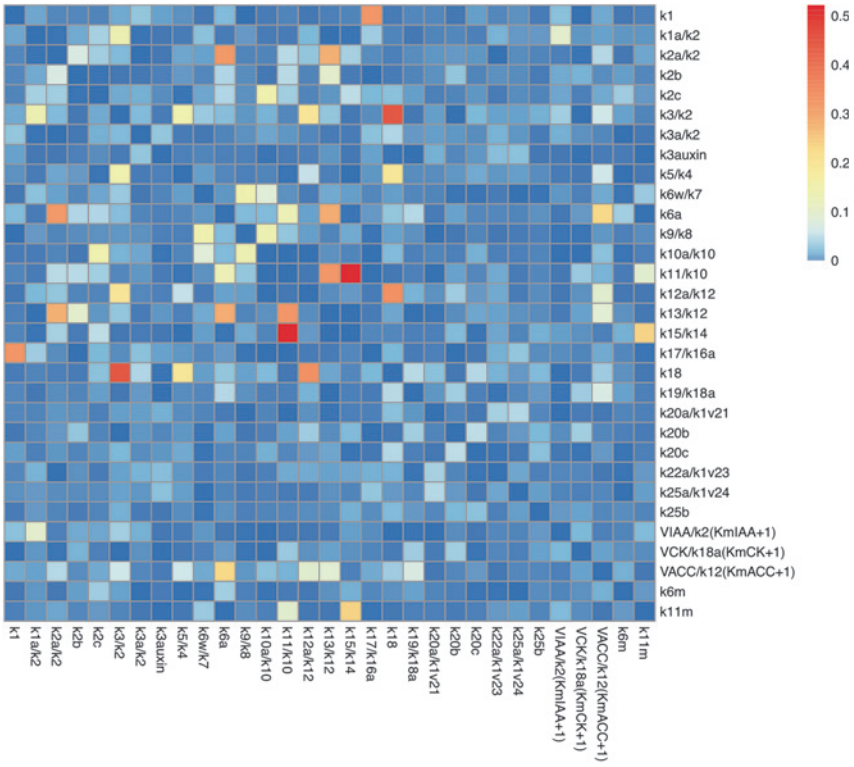

Figure 11: Standardised differences between expected sample variance assuming independence between the inputs and the actual variance in the final non-implausible space for each pair of inputs, represented by colour. The diagonal elements are zero. Red represents a larger difference and blue represents a smaller difference. For example, there is a strong joint constraint between k_{11}/k_{10} and k_{15}/k_{14}, and little joint constraint between k_{18} and k_{1a}/k_2.

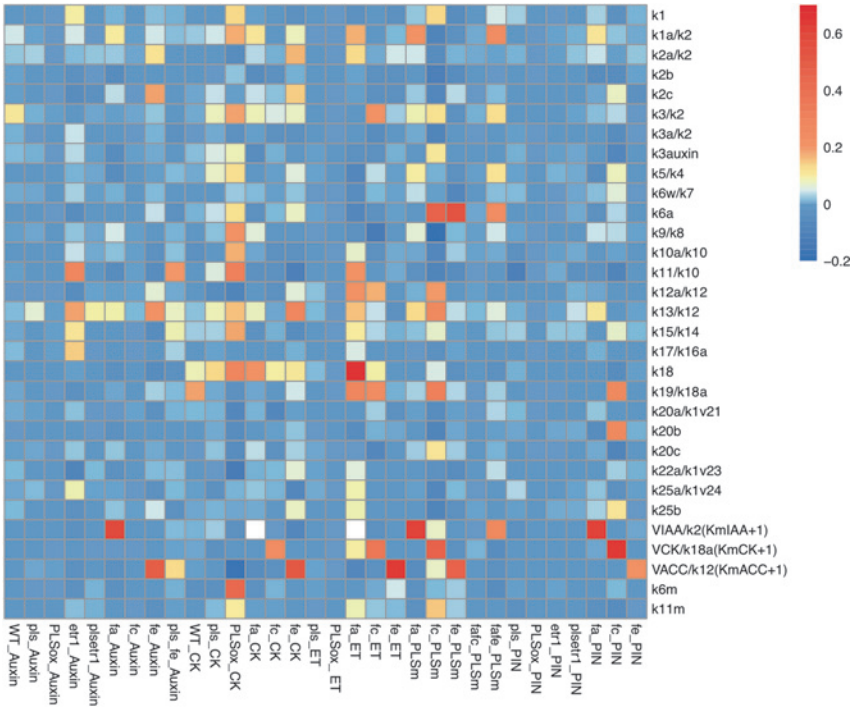

Figure 12: Estimates of how much each output informed each input, represented by colour. These were calculated as the standardised difference between the sample variance of the input for all wave 1 runs and those wave 1 runs going through the output error bar. Red indicates higher values of this estimated quantity and blue represents lower values.

informative about the input parameters when measured in conjunction with other outputs). This is in alignment with Figure 7. Conversely, some outputs are very informative about a few specific inputs, for example f_a_Auxin is particularly informative for k_{1a}/k_2, k_{13}/k_{12} and $V_{IAA}/k_2\,(Km_{IAA}+1)$, with estimated sample variance resolutions of 0.14, 0.13 and 0.52 respectively. It may be unsurprising that matching to the observation of auxin

when feeding auxin is informative for learning about the rate parameters k_{1a}/k_2 or $V_{IAA}/k_2\,(Km_{IAA}+1)$, which represent auxin transport to the cell and the quantity of auxin taken up by the plant respectively. It is more interesting, however, that this experimental observation is also informative for learning about the parameter k_{13}/k_{12}, representing the relationship between biosynthesis and decay of ethylene. Other outputs, for example *etr*1_*Auxin* and *pls_CK*, are slightly informative for a range of the inputs, but not very informative for any of them. This indicates that these outputs are quite informative for learning about the rate parameters and their relationships with each other across the whole network.

Conversely, we can see from Figure 12 that each input is informed about by a different variety of outputs. Some inputs are learnt about by a large number of outputs, for example k_3/k_2 and k_{13}/k_{12} which are the decay rate parameters for the decay of auxin and ethylene respectively from the cell. Interestingly, many of these outputs involved the measurement of cytokinin. Some inputs, for example k_{2b} and k_{3a}/k_2, don't seem to be informed about by many outputs at all. These results are in alignment with Figure 8 which shows the general change in variance for each input. Other inputs are learnt about quite heavily by just a few outputs. For example, k_{6m} which represents the additional *PLS* transcription rate in *PLSox* relative to *WT*, is learnt about heavily after measuring *PLSox_CK*, that is the measurement of cytokinin concentration under the mutant relative to that of WT, with sample variance resolution 0.32. We can see that such an analysis of which output measurements inform us about which input constraints can be insightful. Some of the input-output relationships may be quite intuitive, whilst others inform us about links between the inputs and the outputs of which we were unaware before we started the history matching analysis. Whilst Figure 12 is informative, it is limited to information about the relationship between one input and one output. Information about how single outputs inform us about complex interactions between inputs, or how multiple outputs may be telling us similar information about particular inputs, is not displayed.

4.1 Gaining insight into specific learning objectives from history matching results

Insight into many specific aspects of the model of particular interest can be obtained from the results of a history match. For example, some results in the literature suggest that output f_c_*Auxin*, that is measuring the ratio of Auxin concentration in WT fed cytokinin relative to WT with no feeding, is a down trend, whilst others suggest that it is an up trend (Jones et al. 2010). We therefore separate the final sample of acceptable runs into two groups to analyse whether further measurements of this output would have an effect on our conclusions.

Figure 13 shows boxplots summarising the range of output values for simulator runs $f_i\,(x)$ of all 32 outputs for the final sample of acceptable runs. The light green boxplots are for runs having positive output value for f_c_*Auxin* and dark green boxplots are for runs having negative value for this output. Approximately 80% of the sample runs in the final non-implausible input space have negative values for output f_c_*Auxin* relative to approximately 45% of the initial wave 1 runs. This is a result of matching to other outputs since nearly all initial runs already went through the error bar for f_c_*Auxin*. There are a few outputs, for example f_c_*ET*, which distinguish between runs with positive or negative values of f_c_*Auxin*. However, in general, it would appear that most of the other outputs are relatively independent of f_c_*Auxin*. Therefore, it could be worth taking more careful observations of experiment f_c_*Auxin* in order to learn more about the effect of feeding cytokinin on auxin concentration. Such observations may provide information about the model and physical system which is not being captured by the other experiments.

Figure 14 shows, below the diagonal, for each pair of a subset of inputs for the final simulator acceptable runs, the boundaries of the 0.5 and 0.9 highest density sets as solid and dashed contours respectively. Brown contours indicate runs with positive value for output f_c_*Auxin* and green runs have negative values for this output. We can see that some inputs, for example k_{2b} and k_3/k_2 involving the effects of auxin and cytokinin concentrations on the rate of change of auxin concentration, tend to show a distinction between runs with positive and negative output values for f_c_*Auxin*. This suggests that further measurement of f_c_*Auxin* would be informative for learning about these rate parameters. Above the diagonal are the overall density plots for this subset of outputs for comparison.

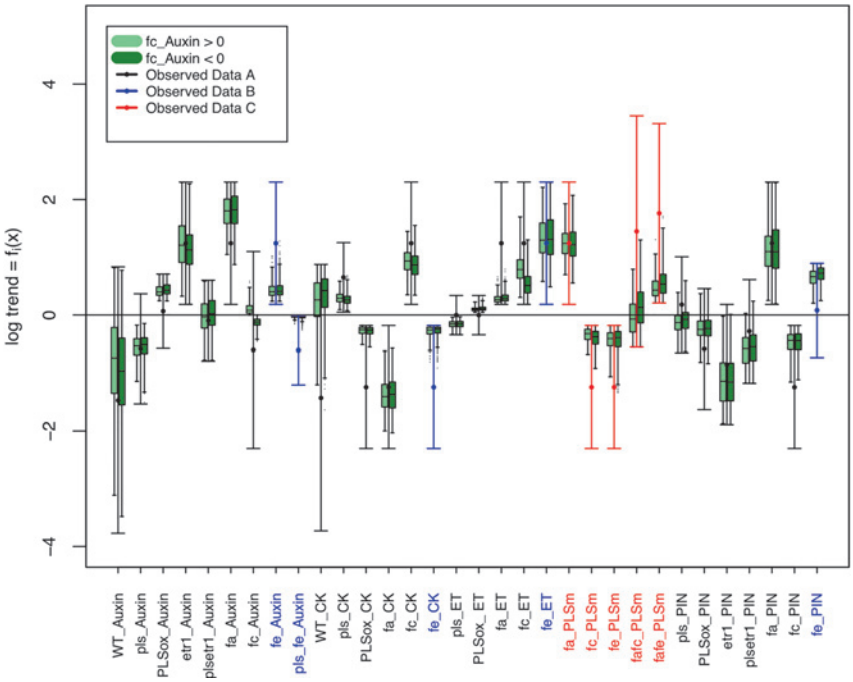

Figure 13: Boxplots summarising the range of output values for simulator runs $f_i(x)$ for all 32 outputs considered that satisfied all of the error bars. The light green boxplots are for runs having positive output value for f_c_*Auxin* and dark green boxplots are for runs having negative value for this output. The targets for the history match, as given by the intervals $z_i \pm 3\sigma_{c_i}$ and the ranges in Table 4, are shown as vertical error bars. Black error bars represent Dataset *A* outputs, blue error bars represent Dataset *B* outputs and red error bars represent Dataset *C* outputs. The horizontal black line at zero represents zero trend.

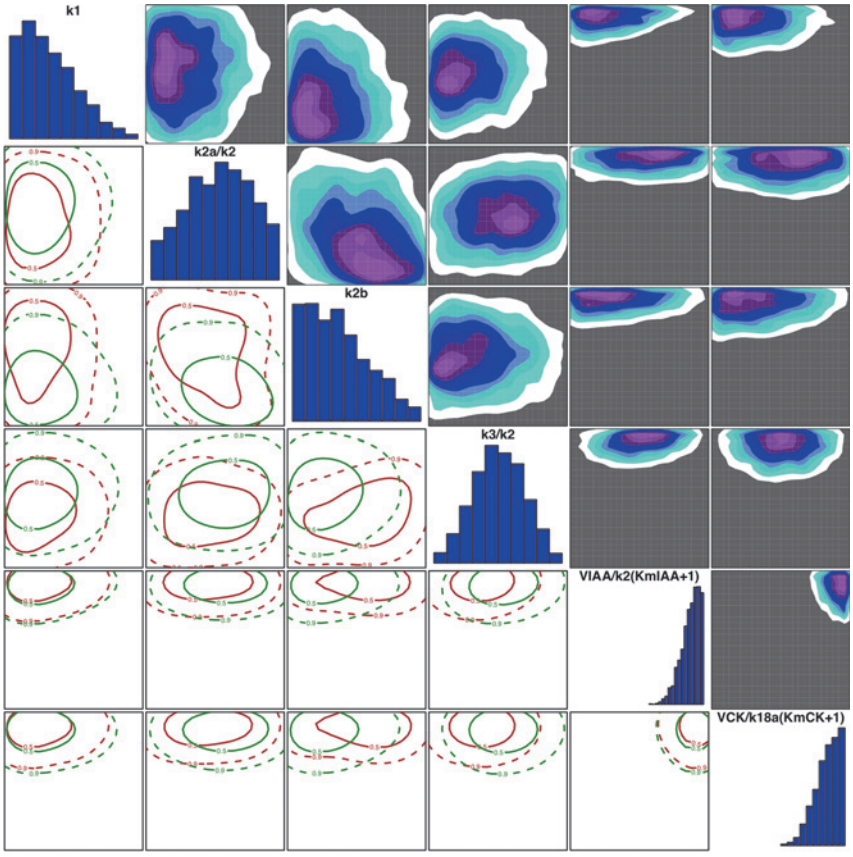

Figure 14: Below diagonal: Contours showing the 0.5 and 0.9 highest density sets for an initial sample of wave 14 runs. Brown contours indicate runs with positive value for output f_c_*Auxin* and green runs have negative values for this output. Above diagonal: 2-dimensional optical depth plots of inputs to runs with acceptable matches to all of the observed data for the same subset of the inputs. The orientation of these plots has been flipped to be consistent with the plots below the diagonal. Along diagonal: 1-dimensional optical depth plots.

Many other interesting features of the model could be analysed in a similar way. In future work we will demonstrate how we can design future experiments using complex models, combined with history matching methodology, in order to choose the set of measurements to perform that will have the best chance of learning about specific scientific criteria of interest.

5 Discussion and conclusions

Understanding how hormones and genes interact to coordinate plant growth is a major challenge in plant developmental biology. Auxin, cytokinin and ethylene are three important hormones that regulate many aspects of plant development. The dynamics of this crosstalk are non-linear and unintuitive (Liu et al. 2014, 2017). Experimental measurements are necessary in order to represent the general dynamics of such a system by formulating kinetic equations. In particular, it is essential to establish how the associated model parameter space can be informed about by experimental observations, since understanding of the rate parameters is essential for a model to be informative for a physical system. In this work, we have shown how comprehensive exploration and understanding of the input rate parameter space can be achieved from sets of experimental observations by applying sequential history matching techniques using Bayesian emulation.

The rate parameter k_{6a} describes how the POLARIS transcriptional rate is regulated by ethylene (Liu et al. 2010). Increasing k_{6a} decreases the strength of this regulation. Figures 5 and 6 suggest that the set of possible values of k_{6a} which satisfy all of the observed data is quite constrained, with large values and the smallest values in the initial range being classed as implausible. Noticeably, this parameter was primarily constrained by the inclusion of dataset C, which measured PLSm.

The parameter ratio k_{6w}/k_7 represents the transcription rate of the POLARIS gene function in WT, and the parameter k_{6m} represents the additional POLARIS transcription when the POLARIS gene is overexpressed. Figure 8 suggests that k_{6w}/k_7 is not highly constrained by the observed measurements, but that k_{6m} is highly constrained after history matching to the observations in Dataset *A*. Figure 12 provides further insight by showing that k_{6m} is particularly constrained by matching to the observation of [*CK*] when POLARIS was overexpressed.

Figure 5 suggests that there is a positive trend between non-implausible values of k_{11}/k_{10}, the ratio for the rate of ethylene receptor conversion from its active to inactive form, to the conversion rate from inactive to active form, and k_{13}/k_{12}, the parameter representing the ratio of ethylene decay rate to biosynthesis rate. This is consistent with current biological understanding that ethylene promotes the conversion from the active form of the ethylene receptor to its inactive form.

The feeding terms $\frac{V_{IAA}[IAA]}{Km_{IAA}+[IAA]}$, $\frac{V_{CK}[cytokinin]}{Km_{CK}+[cytokinin]}$ and $\frac{V_{ACC}[ACC]}{Km_{ACC}+[ACC]}$ are highly constrained by the measurements involving feeding, as can be seen by Figures 8 and 12. In particular, the feeding of ethylene was constrained only after measurements involving the feeding of the ethylene hormone were measured. Although this is unsurprising, strong contradictions to such expected results may be an indication of a problem arising during the history matching procedure, hence these results are indicators that the history match was successful.

In addition, our history matching procedure can also be used to investigate specific aspects of the model. For example, the consequences of two experimentally determined, but opposing, regulatory relationships on constraining the non-implausible parameter space can be determined. Our analysis, summarised in Figures 6 and 12–14, reveals what can be learnt about if further investigation was performed into the trend for f_{c_Auxin}. In particular, we reveal the differences that a confirmed positive or negative trend for this output would have upon constraining the non-implausible parameter space.

In this article, we have developed the study of complex systems biology models using Bayes linear uncertainty analysis, with particular application to an important hormonal crosstalk model of Arabidopsis root development. We have demonstrated the advantages of utilising a formal statistical model to link the biological model to reality. We have also shown that performing a careful history match using implausibility measures, with the assistance of emulators, allows a global exploration of the input parameter space of the

model. In particular, by introducing experimental observations to the history matching procedure sequentially, we can explore constraints imposed by each group of observations, thus aiding the understanding of connections between the inputs and outputs of the model. This in turn allows specific scientific learning objectives to be realised in the context of the model and by the links between the model and the biological system. Being able to understand the contribution of particular experiments for informing us about acceptable input combinations can allow sensible experimental observations to be made relevant to the specific scientific learning objectives of the future.

Plant root developments are regulated by multiple hormones in a coordinated way. Understanding the interdependence of the hormonal regulatory relationships, proteins and gene functions involved in root development is a difficult task. We demonstrate that a combination of experimental observations, a model of hormonal crosstalk in Arabidopsis root development, and a Bayesian history match is able to establish the relationships between physical experiments and parameter space. Thus, following the methodology we have developed in this work, future research should be able to more rationally integrate experimental measurements, model development and determination of non-implausible parameter space, for elucidating the complexity in hormonal signalling systems (Babtie and Stumpf 2017; Liu et al. 2017).

Acknowledgements: JL and KL gratefully acknowledge the Biotechnology & Biological Sciences Research Council (BBSRC) for funding in support of this study. SEJ is in receipt of an Engineering and Physical Sciences Research Council (EPSRC) studentship. IV gratefully acknowledges Medical Research Council (MRC) and EPSRC funding.
Author contribution: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.
Research funding: This research was funded by BBSRC, MRC, and EPSRC.
Conflict of interest statement: The authors declare no conflicts of interest regarding this article.

Appendices

A Extra Parameter λ

In order to perform a full analysis on the Arabidopsis model, we introduced a parameter $\lambda = V_i/V_m$ to represent the ratio of cytosolic volume V_i to the volume of the cell wall V_m. This section outlines why it is necessary to introduce this parameter.

All but one of the outputs of the model represent chemical concentrations within the interior of the cell. Therefore, relative values of chemical volume and chemical concentration of these outputs are the same. On the other hand, however, *PIN*1*pm* should represent the concentration of PIN protein in the exterior of the cell. The volume of the membrane is less than the volume of the interior of the cell, and this needs to be taken into account. We outline an appropriate method to address this issue given the approximate nature of the model, which represents a single cell and membrane. We need conservation of mass to hold for the overall mass of the PIN protein, or equivalently for flux into the membrane to be equal to flux out of the membrane. This can be represented by:

$$\frac{d[PIN1pm]}{dt} = \lambda\left(k_{1,24}[PIN1pi] - \frac{k_{25a}[PIN1pm]}{1+\frac{[Auxin]}{k_{25b}}}\right) \tag{19}$$

where $\lambda = V_i/V_m$ represents the ratio of the volume of the interior of an average cell V_i to the volume of the exterior of an average cell V_m. In general, we can introduce λ as an additional model parameter to the model equation for $[PIN1pm]$ as given by Equation (19), however since we are modelling at equilibrium only, we can

instead just incorporate the effects of this extra model parameter into the rate parameters k_3 and k_{25a}, leaving the equations unaltered. Essentially we investigate $k_3 = \frac{k_3^{'}}{\lambda}$ and $k_{25a} = \frac{k_{25a}^{'}}{\lambda}$, where $k_3^{'}$ and $k_{25a}^{'}$ are the rate parameters assuming equal volume in both cell and membrane.

There are several ways to treat the additional parameter λ. λ could be varied as an extra parameter to the history match over a range of values believed to correspond to cell interior-membrane volume ratios. The effect of varying cell sizes is already a feature of model discrepancy. We therefore believed that it was adequate to fix λ and incorporate the uncertainty of λ as a source of internal model discrepancy (Vernon et al. 2010a, 2010b). Internal model discrepancy refers to aspects of the model discrepancy which can be informed about by running the model. Expert elicitation about the ratio λ led us to fix $\lambda = 6$ and suggested that a reasonable range of possible values for λ was [2, 16]. We made sure that model discrepancy arising from varying the value of λ was captured in each output's model discrepancy assessment.

Nomenclature

ϵ_i	model discrepancy
X	non-implausible set
X_k	non-implausible set after wave k of a history match
$\sigma^2_{w_i}$	constant variance of nugget term
d	dimension of simulator output vector
D_i	set of simulator runs
e_i	measurement error
f	simulator
g_{ij}	known functions of x_{A_i}
$I_i(x)$	implausibility function
k_i	chemical reaction rates
q	dimension of simulator input vector
Q_k	set of outputs emulated at wave k
S_{A_i}	set of indices of the active inputs for output i
$U_i(x)$	standardised prediction errors
$u_i(x_{A_i})$	second-order weakly stationary process
v_i	reactions in the model of Liu et al. (2013)
$w_i(x)$	zero-mean "nugget" term
x	simulator input vector
x^*	best input
x_{A_i}	set of active variables
y_i	physical system value
z_i	observed value

References

Alves, R., Antunes, F., and Salvador, A. (2006). Tools for kinetic modeling of biochemical networks. Nat. Biotechnol., 24: 667–672.

Andrianakis, Y. and Challenor, P. G. (2009). *Parameter estimation and prediction using Gaussian processes*. MUCM Technical report. University of Southampton.

Andrianakis, Y., and Challenor, P.G. (2011). *Parameter estimation for Gaussian process emulators*. Tech. rept. MUCM.

Andrianakis, I., Vernon, I., McCreesh, N., McKinley, T. J., Oakley, J. E., Nsubuga, R. N., Goldstein, M., and White, R. G. (2015). *Bayesian history matching of complex infectious disease models using emulation: a tutorial and a case study on hiv in uganda*.

Andrianakis, I., McCreesh, N., Vernon, I. R., McKinley, T. J., Oakley, J., Nsubuga, R., Goldstein, M., and White, R. G. (2017a). History matching of a high dimensional individual based HIV transmission model. J. Uncertain. Quantification. 5: 694–719.

Andrianakis, I., Vernon, I., McCreesh, N., McKinley, T. J., Oakley, J. E., Nsubuga, R. N., Goldstein, M., and White, R. G. (2017b). History matching of a complex epidemiological model of human immunodeficiency virus transmission by using variance emulation. J. Roy. Stat. Soc. C Appl. Stat., 66: 717–740.

Arendt, P.D., Apley, D.W., and Chen, W. (2012). Quantification of model uncertainty: Calibration, model discrepancy, and identifiability. J. Mech. Des., 134, https://doi.org/10.1115/1.4007390.

Babtie, A. C. and Stumpf, M. P. H. (2017). How to deal with parameters for whole-cell modelling. Interface, 14, https://doi.org/10.1098/rsif.2017.0237.

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

Boogerd, F. C., Bruggeman, F., Hofmeyr, J. H. S., and Westerhoff, H. V. (Eds.) 2007. *Systems biology philosophical foundations*. Elsevier, Amsterdam.

Bower, R.G., Vernon, I., Goldstein, M., Benson, A. J., Lacey, C. G., Baugh, C. M., Cole, S., and Frenk, C. S. 2010 (October). *The parameter space of galaxy formation*. Online link: https://dx.doi.org/10.1111/j.1365-2966.2010.16991.x Also published in the Monthly notices of the Royal Astronomical Society.

Brooks, S., Gelman, A., Jones, G., and Meng, X. (2011). *Handbook of Markov chain monte carlo*. CRC press, Florida.

Brynjarsdottir, J. and O'Hagan, A. (2014). Learning about physical parameters: The importance of model discrepancy. Inverse Probl., 30, https://doi.org/10.1088/0266-5611/30/11/114007.

Castelletti, A., Galelli, S., Ratto, M., Soncini-Sessa, R., and Young, P. C. (2012). A general framework for dynamic emulation modelling in environmental problems. Environ. Model. Software, 34: 5–18.

Castruccio, S., McInerney, D. J., Stein, M. L., Crouch, F. L., Jacob, R. L., and Moyer, E. J. (2014). Statistical emulation of climate model projections based on precomputed GCM runs. J. Clim., 27: 1829–1844.

Conti, S., Gosling, J. P., Oakley, J., and O'Hagan, A. (2009). *Gaussian process emulation of dynamic computer codes*. MUCM.

Craig, P. S., Goldstein, M., Seheult, A. H., and Smith, J. A. (1996). Bayes linear strategies for matching hydrocarbon reservoir history. In: Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (Eds.), *Bayesian statistics*, Vol. 5. Clarendon Press, Oxford, pp. 69–95.

Craig, P. S., Goldstein, M., Seheult, A. H., and Smith, J. A. (1997). Pressure matching for hydrocarbon reservoirs: a case study in the use of Bayes linear strategies for large computer experiments (with discussion). In: Gatsonis, C., Hodges, J. S., Kass, R. E., McCulloch, R. E., Rossi, P., and Singpurwalla, N. D. (Eds.), *Case studies in Bayesian statistics*, Vol. 3. Springer, New York, pp. 36–93.

Cumming, J. and Goldstein, M. (2007). *Multilevel emulation*. MUCM Technical Report 10/07.

Cumming, J. and Goldstein, M. (2010). Bayes linear uncertainty analysis for oil reservoirs based on multiscale computer experiments. In: O'Hagan, A., and Mike, W. (Eds.). *The Oxford handbook of applied Bayesian analysis*. Oxford University Press, pp. 241–270.

Cumming, J. and Goldstein, M. (2009). Small sample Bayesian designs for complex high-dimensional models based on information gained using fast approximations. Technometrics, 51: 377–388.

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. J. Am. Stat. Assoc., 86: 953–963.

de Finetti, B. (1974). *Theory of probability*, Vol. 1. Wiley.

de Finetti, B. (1975). *Theory of probability*, Vol. 2. Wiley.

Farah, M., Birrell, P., Contin, S., and De Angelis, D. (2014). Bayesian emulation and calibration of a dynamic epidemic model for A/H1N1 influenza. J. Am. Stat. Assoc., 109: 1398–1411.

Fisher, R. A. (1937). *The design of experiments*. Oliver and Boyd.

Forrester, A. (2010). Black-box calibration for complex-system simulation. Phil. Trans. Roy. Soc. A, 368: 3567–3579.

Gardner, P., Lord, C., and Barthorpe, R. J. (2018). Bayesian history matching for forward model-driven structural health monitoring. Model Validation Uncertain. Quantification, 3: 175–183.

Goldstein, M. (1999). Bayes linear analysis. Chap. Bayes Linear Analysis. In: Kotz, S., Read, C. B., Balakrishnan, N., and Vidakovic, B. (Eds.), *Encyclopedia of statistical sciences*. Wiley, New York, pp. 29–34.

Goldstein, M. and Rougier, J. C. (2006). Bayes linear calibrated prediction for complex systems. J. Am. Stat. Assoc., 101: 1132–1143.

Goldstein, M. and Rougier, J. C. (2009). Reified Bayesian modelling and inference for physical systems (with Discussion). J. Stat. Plann. Inference, 139: 1221–1239.

Goldstein, M. and Wooff, D. (2007). *Bayes linear statistics*. Wiley, Chichester.

Goldstein, M., Seheult, A., and Vernon, I. (2013). Assessing model adequacy. In: Wainwright, J. and Mulligan, M. (Eds.), *Environmental modelling: Finding simplicity in complexity*. John Wiley and Sons, Chichester.

Gong, Z. and DiazDelaO, F. A. (2017). Sampling schemes for history matching using subset simulations. Proceedings for the 1st international conference on uncertainty quantification in computational sciences and engineering.

Gosling, J. P., Hart, A., Owen, H., Davies, M., Li, J., and MacKay, C. (2013). A Bayes linear approach to weight-of-evidence risk assessment for skin allergy. Bayesian Anal., 8: 169–186.

GPy. (since 2012). *GPy: A Gaussian process framework in Python*. Available at: https://github.com/SheffieldML/GPy.

Hamdi, H., Couckuyt, I., Sousa, M. C., and Dhaene, T. 2017. Gaussian Processes for history-matching: application to an unconventional gas reservoir. Comput. Geosci., 21: 267–287.

Hankin, R. K. S. (2005). Introducing BACCO: an R bundle for Bayesian analysis of computer code output. J. Stat. Software, 14, https://doi.org/10.18637/jss.v014.i16.

Heitmann, K., Higdon, D., White, M., Habib, S., Williams, B. J., Lawrence, E., and Wagner, C. (2010). *The coyote universe II: cosmological models and precision emulation of the nonlinear matter power spectrum*.

Higdon, D., Gattiker, J., Williams, B., and Rightley, M. (2008). Computer model calibration using high-dimensional output. 103: 570–583.

Holden, P. B., Edwards, N. R., Hensman, J., and Wilkinson, R. D. (2018). *Handbook of approximate Bayesian computation*. Chapman and Hall. Chap. ABC for climate: dealing with expensive simulators.

Jamshidi, N. and Palsson, B. O. (2008). Formulating genome-scale kinetic models in the post-genome era. Mol. Syst. Biol., 4, https://doi.org/10.1038/msb.2008.8.

Johnson, J. S., Gosling, J. P., and Kennedy, M. C. (2011). Gaussian process emulation for second-order Monte Carlo simulations. J. Stat. Plann. Inference, 141: 1838–1848.

Jones, B., Gunneras, S. A., Petersson, S. V., Tarkowski, P., Graham, N., May, S., Dolezal, K., Sandberg, G., and Ljung, K. (2010). Cytokinin regulation of auxin synthesis in Arabidopsis involves a homeostatic feedback loop regulated via auxin and cytokinin signal transduction. Plant Cell, 22: 2956–2969.

Journel, A. G. and Huijbregts, C. J. (1978). *Mining geostatistics*. Academic Press, Amsterdam.

Kaye, K., Day, R. D., Hair, E. C., Moore, K. A., Hadley, A. M., Teixeira, P. J., Helmschrott, S., Massing, N. and Ackermann, D. (2009). Parent marital quality and the parent-adolescent relationship: effects on sexual activity among adolescents and youth. Marriage & Family Rev., 45: 270–288.

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models. J. Roy. Stat. Soc., 63: 425–464.

Liu, J., Mehdi, S., Topping, J., Tarkowski, P., and Lindsey, K. (2010). Modelling and experimental analysis of hormonal crosstalk in arabidopsis. Mol. Syst. Biol., 6, https://doi.org/10.1038/msb.2010.26.

Liu, J., Mehdi, S., Topping, J., Friml, J., and Lindsey, K. (2013). Interaction of PLS and PIN and hormonal crosstalk in arabidopsis root development. Front. Plant Sci., 4, https://doi.org/10.3389/fpls.2013.00075.

Liu, J., Rowe, J., and Lindsey, K. (2014). Hormonal crosstalk for root development: a combined experimental and modelling perspective. Front. Plant Sci.: 116, https://doi.org/10.3389/fpls.2014.00116.

Liu, J., Moore, S., Chen, C., and Lindsey, K. (2017). Crosstalk complexities between auxin, cytokinin, and ethylene in arabidopsis root development: From experiments to systems modeling, and back again. Mol. Plant, 10: 1480–1496.

MacDonald, B., Ranjan, P., and Chipman, H. (2015). GPfit: An R package for fitting a Gaussian process model to deterministic simulator outputs. J. Stat. Software, Articles 64: 1–23.

McCreesh, N., Andrianakis, I., Nsubuga, R. N., Strong, M., Vernon, I., McKinley, T. J., Oakley, J. E., Goldstein, M., Hayes, R., and White, R. G. (2017). Universal test, treat, and keep: improving ART retention is key in cost-effective HIV control in Uganda. BMC Infect. Dis., 17: 322.

McKay, M. D., Beckman, R. J., and Conover, W. J. (1979). A Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics, 21: 239–245.

McKinley, T. J., Vernon, I., Andrianakis, I., McCreesh, N., Oakley, J. E., Nsubuga, R., Goldstein, M., and White, R. G. (2018). Approximate Bayesian computation and simulation-based inference for complex stochastic epidemic models. Stat. Sci., 33: 4–18.

Mobius, W. and Laan, L. (2015). Physical and mathematical modelling in experimental papers. Cell, 163: 1577–1583.

Mohamed, L., Calderhead, B., Filippone, M., Christie, M., and Girolami, M. (2012). Population MCMC methods for history matching and uncertainty quantification. Comput. Geosci., 16: 423–436.

Montgomery, D. C. (2009). *Design and analysis of experiments*. Wiley.

Moore, S., Zhang, X., Liu, J., and Lindsey, K. (2015a). Modelling plant hormone gradients. eLS: 1–10, https://doi.org/10.1002/9780470015902.a0023733.

Moore, S., Zhang, X., Liu, J., and Lindsey, K. (2015b). Some fundamental aspects of modeling auxin patterning in the context of auxin-ethylene-cytokinin crosstalk. Plant Signal. Behav., 10: e1056424. PMID: 26237293.

Moore, S., Zhang, X., Mudge, A., Rowe, J. H., Topping, J. F., Liu, J., and Lindsey, K. (2015c). Spatiotemporal modelling of hormonal crosstalk explains the level and patterning of hormones and gene expression in arabidopsis thaliana wild-type and mutant roots. New Phytol., 207: 1110–1122.

Moore, S., Liu, J., Zhang, X., and Lindsey, K. (2017). A recovery principle provides insight into auxin pattern control in the Arabidopsis root. Sci. Rep., 7, https://doi.org/10.1038/srep43004.

Neal, R. M. (1997). *Monte Carlo implementation of Gaussian process models for Bayesian regression and classification*. Tech. rept. University of Toronto.

Nott, D. J., Fan, Y., Marshall, L., and Sisson, S. A. (2014). Approximate Bayesian computation and Bayes' linear analysis: Toward high-dimensional ABC. J. Comput. Graph Stat., 23: 65–86.

O'Hagan, A. (1987). Bayes linear estimators for randomized response models. J. Am. Stat. Assoc., 82: 580–585.

Oliver, D. S. and Chen, Y. (2011). Recent progress on reservoir history matching: a review. Comput. Geosci., 15: 185–221.

Overstall, A. M. and Woods, D. C. (2016). Multivariate emulation of computer simulators: model selection and diagnostics with application to a humanitarian relief model. J. Roy. Stat. Soc. C Appl. Stat., 65, https://doi.org/10.1111/rssc.12141.

Picheny, V., Ginsbourger, D., Richet, Y., and Caplin, G. (2013). Quantile-based optimization of noisy computer experiments with tunable precision. Technometrics, 55, https://doi.org/10.1080/00401706.2012.707580.

Prangle, D., Everitt, R. G., and Kypraios, T. (2018). A rare event approach to high-dimensional approximate Bayesian computation. Stat. Comput., 28: 819–834.

Pukelsheim, F. (1994). The three sigma rule. Am. Statistician, 48: 88–91.
Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press. http://www.gaussianprocess.org/gpml/.
Santner, T. J., Williams, B. J., and Notz, W. I. (2003). *The design and analysis of computer experiments*. Springer, New York.
Smallbone, K., Simeonidis, E., Swainston, N., and Mendes, P. (2010). Towards a genome-scale kinetic model of cellular metabolism. BMC Syst. Biol., 4, https://doi.org/10.1186/1752-0509-4-6.
Smith, A. F. M. and Gelfand, A. E. (1992). Bayesian statistics without tears: A sampling-resampling perspective. Am. Statistician, 46: 84–88.
Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. H. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. J. R. Soc. Interface, 6: 187–202.
Torres, N. V. and Santos, G. (2015). The (Mathematical) modelling process in biosciences. Front. Genet., 6: 354.
Vernon, I., Goldstein, M., and Bower, R. G. (2010a). Galaxy formation: A Bayesian uncertainty analysis. Bayesian Anal., 5: 619–669.
Vernon, I., Seheult, A., and Goldstein, M. (2010b). *Modular dynamic emulation and internal model discrepancy for a rainfall runoff model*. Managing Uncertainty in Complex Models. Durham University, Durham, UK.
Vernon, I., Goldstein, M., and Bower, R. G. (2010c). Rejoinder - galaxy formation: A Bayesian uncertainty analysis. Bayesian Anal., 5: 697–708.
Vernon, I., Goldstein, M., Rowe, J., Topping, J., Liu, J., and Lindsey, K. (2018). Bayesian uncertainty analysis for complex systems biology models: emulation, global parameter searches and evaluation of gene functions. BMC Syst. Biol., 12, https://doi.org/10.1186/s12918-017-0484-3.
Wang, X., Nott, D. J., Drovandi, C. C., Mengersen, K., and Evans, M. (2018). Using history matching for prior choice. Technometrics, 60: 445–460.
Whittle, P. (1958). On the smoothing of probability density functions. J. Roy. Stat. Soc. B Methodol., 20: 334–343.
Wilkinson, R. D. (2013). Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. Stat. Appl. Genet. Mol. Biol., 12: 129–141.
Wilkinson, R. D. (2014). Accelerating ABC methods using Gaussian processes. In: *JMLR: Workshop and conference proceedings*, Vol. 33, pp. 1015–1023.
Williams, C. K. I. and Rasmussen, C. E. (1996). Gaussian processes for regression. Adv. Neural Inf. Process. Syst., 8: 514–520.
Williamson, D. and Vernon, I. (2013). *Efficient uniform designs for multi-wave computer experiments*.
Williamson, D., Goldstein, M., Allison, L., Blaker, A., Challenor, P., Jackson, L., and Yamazaki, K. (2013). History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble. Clim. Dynam., 41: 1703–1729.
Zhang, X., Hou, J., Wang, D., Mu, T., Wu, J., and Lu, X. (2012). An automatic history matching method of reservoir numerical simulation based on improved genetic algorithm. *Proceedings to the 29th international workshop on information and electronic engineering*, Vol. 29, pp. 3924–3928.