Prediction uncertainty assessment of a systems biology model requires a sample of the full probability distribution of its parameters

Simon van Mourik[1], Cajo ter Braak[2], Hans Stigter[3], and Jaap Molenaar[4]

[1,2,3,4]**Biometris, Wageningen University and Research Center, Wageningen, The Netherlands**
[1,4]**Netherlands Consortium for Systems Biology, Amsterdam, The Netherlands**

ABSTRACT

Keywords: Prediction uncertainty, parameter uncertainty, MCMC, Bayesian statistics

INTRODUCTION

By combining experiments and mathematical model analysis, systems biology tries to unravel the key mechanisms behind biological phenomena. This has led to a steadily growing number of experiment-driven modeling techniques (Klipp et al., 2008; Stumpf et al., 2011). Useful models provide insight and falsifiable predictions and the dynamic models used in systems biology are no exception. A complicating factor in multi-parameter models is that often many parameters are largely unknown. Even among similar processes parameter values may vary multiple orders of magnitude, and experimentally they are often hard to infer accurately (Maerkl and Quake, 2007; Buchler and Louis, 2008; Teusink et al., 2000).

A standard procedure for parameter estimation is via a collective fit, i.e., estimating the values of the unknown parameters simultaneously by fitting the model to time series data, resulting in a calibrated model. This approach has to cope with several obstacles, such as measurement errors, biological variation, and limited amounts of data. Another often met problem is that different parameters may have correlated effects on the measured dynamics leading to highly uncertain or even unidentifiable parameter estimates (Zak et al., 2003b,a; Raue et al., 2009). Systems biology models are typically 'sloppy' in that some parameters or combinations of parameters are well defined, whereas many others are not (Brown and Sethna, 2003; Gutenkunst et al., 2007b). It was argued on the basis of this sloppiness that collective fits are more promising for obtaining useful predictions than direct parameter measurements, and that "modelers should focus on predictions rather than on parameters" (Gutenkunst et al., 2007b). One might expect that sloppiness and the associated parameter uncertainty automatically lead to uncertain predictions, but that is not necessarily so. For example, in the literature a model has been reported with 48 parameters calibrated to only 68 data points and with highly uncertain parameter values, but the predictions were quite accurate (Brown et al., 2004).

In this paper we fit an illustrative model and a diverse set of six systems biology models from the BioModels database (Li et al., 2010) to different amounts of (simulated) time series data. For each model, we study the prediction uncertainty of a number of predictions by a full computational uncertainty analysis. We measure the uncertainty in the predicted time course by the dimensionless quantifier $Q_{0.95}$ which has the nice property that $Q_{0.95} < 1$ indicates tight predictions, whereas $Q_{0.95} \geq 1$ implies uncertainty in the dynamics that is likely to obscure biological interpretation. Using this quantifier, we show that some models allow more accurate predictions than others, but, more importantly, that the uncertainty of the predictions may greatly vary within one model. We argue that prediction uncertainty assessment must therefore be performed on a per-prediction basis using a full computational uncertainty analysis (Savage, 2012). We indicate how such an analysis can be performed in a relatively easy way. It requires a sample representing (strictly speaking, approximating) the distribution of the parameters of the model. The importance of such a sample (an ensemble of parameter sets) for uncertainty analysis has been pointed out earlier (Brown and Sethna, 2003; Gutenkunst et al., 2007a), but we go one step further and recommend modellers to make available not only the mathematical model and its typical parameter values and operating conditions, but also a sample representing the distribution of its parameters (Fig. 1). For other reasons, such as verification and reproducibility of the research, data and software to obtain such a sample should also be made available.

Within the Bayesian framework (Jaynes, 2003; MacKay, 2003) this is a sample as generated by an MCMC algorithm (Robert et al., 2011) that collectively fits the parameters of the model to experimental data. We advocate the Bayesian framework as it allows inclusion of prior information (based on prior parameter knowledge, general biological considerations (Grandison and Morris, 2008) and direct parameter measurements) to be combined with the information contained in time series data. If no prior information is available, a noninformative prior can be constructed. For a discussion on prior selection, see Robert et al. (2011).

It is useful at this point to compare the Bayesian approach with the frequentist approach of prediction profile likelihood (PPL) (Kreutz et al., 2013) which is another approach to uncertainty analysis. PPL requires heavy computation for uncertainty analysis of each prediction. In particular, it requires the solution of a large number of nonlinear optimization problems, each under a different nonlinear constraint. By contrast, the Bayesian approach requires (perhaps very) heavy computation to obtain a sample representing the distribution of the parameters. But once this sample has been obtained, the uncertainty analysis is easy, namely by calculating the prediction for each parameter vector in the sample and summarizing the so obtained set of predictions. For this reason, the Bayesian framework has strong appeal when uncertainty analysis is required for many predictions from a single model, as in the current paper. For other aspects in the comparison between the Bayesian and frequentist approaches see Raue et al. (2013).

Parameter uncertainty and prediction uncertainty play an important role also in hydrology (Beck, 1987; Beven and Binley, 1992), ecology (Omlin and Reichert, 1999), and meteorology (Hawkins and Sutton, 2009). What we contribute here is a quantifier for
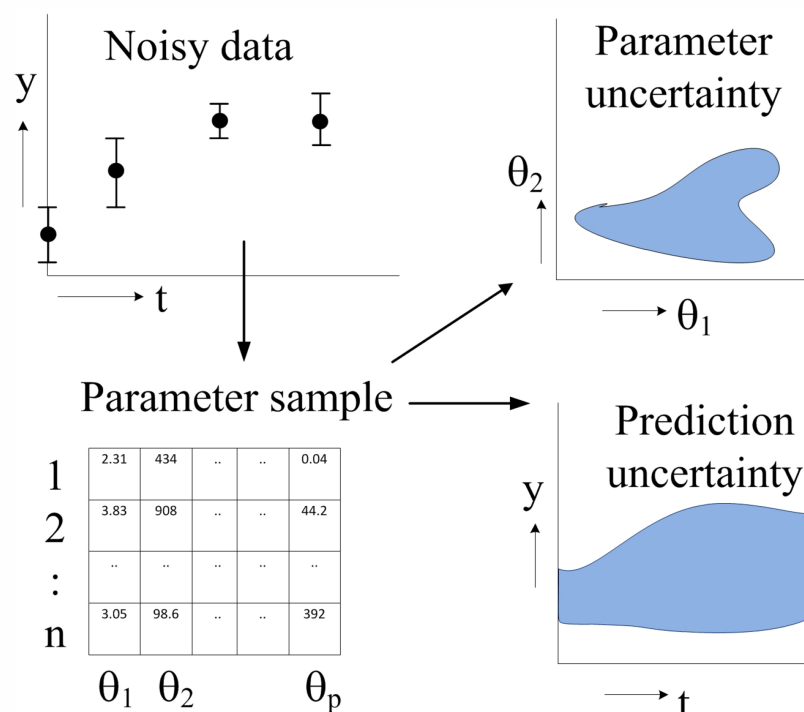
Figure 1. Flowchart of parameter estimation and uncertainty analysis. The model dynamics are fitted to noisy time series data, typically by an MCMC algorithm, resulting in a sample of parameter values representing their posterior distribution. Typical applications of this distribution are the computation of a credible region of parameters (showing parameter uncertainty) and full computational uncertainty analysis of predictions

84   the uncertainty in predicted systems dynamics that expresses the uncertainty in a concise

85   way, and the focus on what is needed to carry out a full computational uncertainty

86   analysis, namely samples representing all sources of uncertainty. In the internet age,

87   such samples can easily be made available for prospective users of the model.

88   MATERIALS AND METHODS

89   In this section we present the methodology to sample the probability distribution of

90   the parameters, and to define the uncertainty of predicted dynamics. To illustrate how

91   this works in practice, we have implemented the used sampling algorithm together with

92   a prediction uncertainty computation algorithm in Matlab® software, together with a

93   sample of the parameters (see the Supplemental Information).

Model class

We consider models that are formulated in terms of differential equations and have the following form:

$$\dot{x}(t) = f(x(t), \theta, u(t))$$
$$y(t) = g(x(t), \theta)$$
$$x(0) = h(\theta). \tag{1}$$

The dynamics of the internal state vector x is a function f of parameter vector $\theta \in \mathbb{R}^p$, with initial condition $x(0)$, and external input vector $u(t)$. The output $y \in \mathbb{R}^m$ is a positive function g of the internal state and θ. To obtain the output time series $y(t)$, we have to numerically integrate $\dot{x}(t) = f(x(t), \theta, u(t))$, which is the time consuming part of the solution.

Estimation of the posterior distribution of the parameters

We adopt the Bayesian framework, in which a prior distribution of the parameters is combined with data to form the posterior distribution $\pi(\theta)$ of the parameters, according to Bayes' theorem

$$\pi(\theta) = p(\theta|y_d) = \frac{p(\theta)p(y_d|\theta)}{p(y_d)}. \tag{2}$$

Here $p(\theta)$ denotes the density of the prior distribution of the parameters based on prior knowledge and $p(y_d|\theta)$ the likelihood of the measured data y_d given θ. The denominator, $p(y_d)$, is the marginal likelihood of the data y_d and acts as a θ-independent normalization constant which is not relevant in MCMC algorithms. We assume that the output $y(t_i), i = 1, .., n$ contains uncorrelated Gaussian noise, with variance σ_i^2 per time point i, so the likelihood of the measured data y_d, given θ, is

$$p(y_d|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(y_i(\theta) - y_{d,i})^2}{2\sigma_i^2}\right), \tag{3}$$

with $y_i(\theta)$ the model output at time t_i given θ. Inserting (3) into (2) and taking the logarithm gives the log-posterior

$$\log(\pi(\theta)) = c - \frac{1}{2}\chi^2(\theta) + \log(p(\theta)), \tag{4}$$

with c a constant independent of θ, and χ^2 a measure of the fitting error:

$$\chi^2(\theta) = \sum_{i=1}^{n} \frac{(y_i(\theta) - y_{d,i})^2}{\sigma_i^2}. \tag{5}$$

The penalized maximum likelihood parameter θ^{PML} maximizes (4). Draws from the posterior $\pi(\theta)$ are obtained by an MCMC algorithm, which, starting from a user-defined initial value of θ, generates a stochastic walk through the parameter space. In iteration k of the walk, a new candidate solution θ is proposed based on the current solution θ_k.

119 We use symmetric proposal distributions centered at θ_k. In this case the proposed θ
120 is accepted or rejected using the Metropolis acceptance probability $\min(1, r)$, where
121 $r = \frac{\pi(\theta)}{\pi(\theta_k)}$. So, the more probable a parameter vector is with respect to the data, the more
122 probable it is to be accepted. In this paper the proposals for the walk are generated
123 by the DE-MCz algorithm, which is an adaptive MCMC algorithm that uses multiple
124 chains in parallel and exploits information from the past to generate proposals (ter Braak
125 and Vrugt, 2008). As consecutive draws are dependent, it may be practical to reduce
126 the dependence by thinning, that is, by storing every K^{th} draw (with $K > 1$). After a
127 number of iterations (the burn-in period) the chain of points $\{\theta_k\}$ will be stationary
128 distributed with a local density that represents the posterior $\pi(\theta)$. The burn-in iterations
129 are discarded.

130 ## A quantifier of uncertainty in predicted systems dynamics
131 Since we want to compare the uncertainty in predictions of time courses over different
132 time intervals and for different models, we need a quantifier that is independent from
133 model specific issues, such as the number, and the dimensions of the variables. Also,
134 biologists are biologists are generally interested in percentage difference rather than
135 absolute difference, so the quantifier should be independent of the typical order of
136 magnitude. To that end we introduce a measure for prediction uncertainty that satisfies
137 these requirements, but we do not claim that this quantifier is unique. However, our
138 present choice has the advantage that it allows a nice interpretation as will be explained
139 below. We first define a measure for the prediction uncertainty of one component of the
140 output and then average the outcomes over the components.
141
142 For a predicted output $y_p(t, \theta)$ on time interval $[0, T]$, the quantifier Q represents the rela-
143 tive deviation of $y_p(t, \theta)$ from the penalized maximum likelihood prediction $y_p(t, \theta^{PML})$,
144 integrated over time:

$$Q_i(\theta) = \frac{1}{T} \int_0^T \left(\log_b \left(\frac{y_{p,i}(t, \theta)}{y_{p,i}(t, \theta^{PML})} \right) \right)^2 dt, \tag{6}$$

145 with $y_{p,i}$ the ith component of y_p, with $i = 1, .., m$, and $Q = \frac{1}{m} \sum_i Q_i$. Q integrates
146 differences over time on a logarithmic scale with base b, and is therefore symmetric
147 with respect to relative over- and underestimations (as opposed to, e.g., a sum of squared
148 errors). The underlying assumption is that y is always positive, which holds true for
149 all models describing concentration dynamics. When only a few time points are of
150 biological interest, the integrand in (6) may be approximated by a summation. The base
151 b characterizes the order of magnitude of the discrepancies that Q is sensitive to. For
152 example, if $\frac{1}{b} < \frac{y_{p,i}}{y_{p,i}^{ML}} < b$ holds for all time points, this will result in $Q < 1$, while a few
153 points outside this range will quickly result in $Q > 1$. In this paper we use $b = 2$, so
154 only differences of a factor two or higher can result in $Q > 1$. The choice of b represents
155 the maximum magnitude of deviations that a prediction is allowed to have, so it should
156 in general be selected based on biological grounds.
157   If $\{\theta\}$ is the collection of sampled θ's reliably representing density $\pi(\theta)$, then the
158 density $\pi(Q(\theta))$ is reliably represented by $Q(\theta)$ with $\theta \in \{\theta\}$ (Robert et al., 2011). We
159 denote the level α prediction uncertainty with Q_α, the α quantile of the distribution

160 of Q. Q_α is therefore the deviation of a prediction relative to the penalized maximum
161 likelihood prediction, at confidence level α (α-level deviation). We use $\alpha = 0.95$
162 throughout.

Algorithm to estimate prediction uncertainty

164 The algorithm for estimating prediction uncertainty naturally falls apart in two sub-
165 algorithms. Part I deals with estimating parameters by exploiting the prior knowledge,
166 the model, and the data. This first part yields the sample of parameter values represent-
167 ing their posterior distribution. Part II is the focus of this paper, and performs the full
168 computational uncertainty analysis by taking as input the sample of parameter values
169 from the first part and by calculating the prediction for each member of this sample. Part
170 I is computationally more intensive than Part II. Because prospective users of the model
171 need to carry out only Part II, it is essential that the sample of parameter values obtained
172 in Part I is stored and made available. In full:

174 Part I: Estimation of the posterior parameter distribution

175 1. To calculate the posterior $\pi(\theta)$ in equation (2), we use equation (3) for the
176   likelihood, together with a log-uniform prior for the parameters, that is, a uniform
177   prior for the logarithm of the parameters.

178 2. A collection of solutions $\{\theta\}$ is generated by MCMC sampling; in this paper we
179   use DE-MCz (ter Braak and Vrugt, 2008). This collection constitutes the sample
180   approximating $\pi(\theta)$.

181 Part II: Computational uncertainty analysis

182 1. $Q(\theta)$ is computed for each $\theta \in \{\theta\}$. The α level prediction uncertainty Q_α is
183   approximated by taking the largest Q after discarding the $100\alpha\%$ largest $Q(\theta)$
184   values.

185 2. For visualization of the uncertainty in the predicted systems dynamics, for each
186      $\theta \in \{\theta\}$ the output $y(t)$ is calculated. For each time point the $\frac{1}{2}(1 - \alpha)100\%$
187   largest and smallest predicted values of $y(t)$ are discarded, whereafter the mini-
188   mum and maximum values are plotted, creating an envelope of predicted dynam-
189   ics.

190   A structural difference between Q and the visualization is that Q integrates the devi-
191 ation over the total function $y(t)$, while the visualization displays prediction uncertainty
192 per time point and per component of $y(t)$. An alternative to definition (6) is to define
193 Q as the deviation of $\overline{\log_b(y_p(t, \theta))}$ from the average log-prediction, thus replacing
194 $\log_b(y_p(\theta^{PML}))$ with $\overline{\log_b(y_{p,i}(\theta))}$.

RESULTS

Illustrative example

197 To demonstrate what kind of effects can be expected when studying prediction uncer-
198 tainty of models with parameter uncertainty, we use a model example that represents a

199 highly simplified case of self-regulation of gene transcription:

$$\dot{y}(t) = \frac{10y(t)^2}{\theta_1 + y(t)^2} - \theta_2 y(t) + y(t)u(t). \tag{7}$$

200 Here, $y(t)$ represents protein concentration and $\dot{y}(t)$ its rate of change. Changes in
201 $y(t)$ stem from self-regulation (the first term in the equation, which is a Hill function
202 (Alon, 2006)), decay (the second term), and an input function (the third term) given by
203 $u(t) = \sin(t)$. The input represents some external signal, e.g., a triggering mechanism
204 (van Mourik et al., 2010). The unknown parameters involved in self-regulation and decay
205 are θ_1 and θ_2, respectively. They are to be estimated from time series data, for which we
206 used simulated data at ten time points (Fig. 2A). We assumed independent Gaussian
207 noise on each measurement, with a mean of 10% of the y value, and a log-uniform prior
208 for the parameter distribution. We estimated θ_1 and θ_2 by generating 1000 draws from
209 their distribution using MCMC sampling. This yielded a sample of 1000 draws. From
210 this sample we derived a 95% credible region of θ_1 and θ_2 (Fig. 2B) and also point-wise
211 95% credible intervals (envelopes) for three sets of predictions (Fig. 2C). The prediction
212 sets differ in starting value $y(0)$ or the input function $u(t)$. The envelopes for prediction
213 (Fig. 2C) are obtained by inserting each of the 1000 draws into the (perturbed) system,
214 calculating the prediction on a fine grid of time points by evaluating the (modified)
215 differential equation (7) and summarizing the 2.5 and 97.5 percentiles of the predictions
216 at each time point. Details are given in the Methods section.

217 The data give much more information about θ_2 than about θ_1 (Fig. 2B, note the
218 different scales on the axes), that is, the dynamics of the time series is sensitive to
219 changes in θ_2, but not sensitive to changes in θ_1; the model thus shows sloppiness. The
220 reason is not hard to understand: since we have used a high initial value, most of the $y(t)$
221 curve has values much bigger than $\theta_1 = 10$, so that the parameter θ_1 has little influence
222 on the dynamics (see (7)). In more complex models, credible regions may have far more
223 complicated shapes than in this illustrative example.

224 The prediction sets Prediction 1 and Prediction 2 (Fig. 2C) are obtained by starting
225 the system for each draw of the sample at a low initial concentration ($y(0) = 1$) and
226 at a very high concentration ($y(0) = 10^4$), respectively. The uncertainty in prediction
227 increases in time for Prediction 1 and is much higher than that for Prediction 2. The
228 predicted dynamics is thus highly sensitive to parameter uncertainty for Prediction 1,
229 but fairly insensitive for Prediction 2. Based on Predictions 1 and 2 one might guess
230 that the uncertainty is related to the magnitude of $y(t)$, but this guess is wrong as the
231 third prediction set shows. In this set, we predict the time course starting at the same low
232 value as in Prediction 1 but now the input amplitude is increased 7-fold ($u(t) = 7\sin(t)$).
233 This drives the concentration dynamics into an oscillating motion which can be predicted
234 rather precisely (Fig. 2C), since now the input $u(t)$ dominates the dynamics.

235 We also estimated prediction uncertainty via a local sensitivity analysis, namely lin-
236 earized covariance analysis (LCA, see the Supplemental Information for details). LCA
237 extrapolates parameter- and prediction uncertainty from a second- and first order Taylor
238 expansion, respectively, and does not take into account the higher order terms that in
239 this case constrain the credible region and the predictions, but note that linearization
240 can also lead to predictions that appear more precise than they really are. Consequently,
241 LCA dramatically overestimates the uncertainties of Predictions 1 and 3 (Fig. 2D).
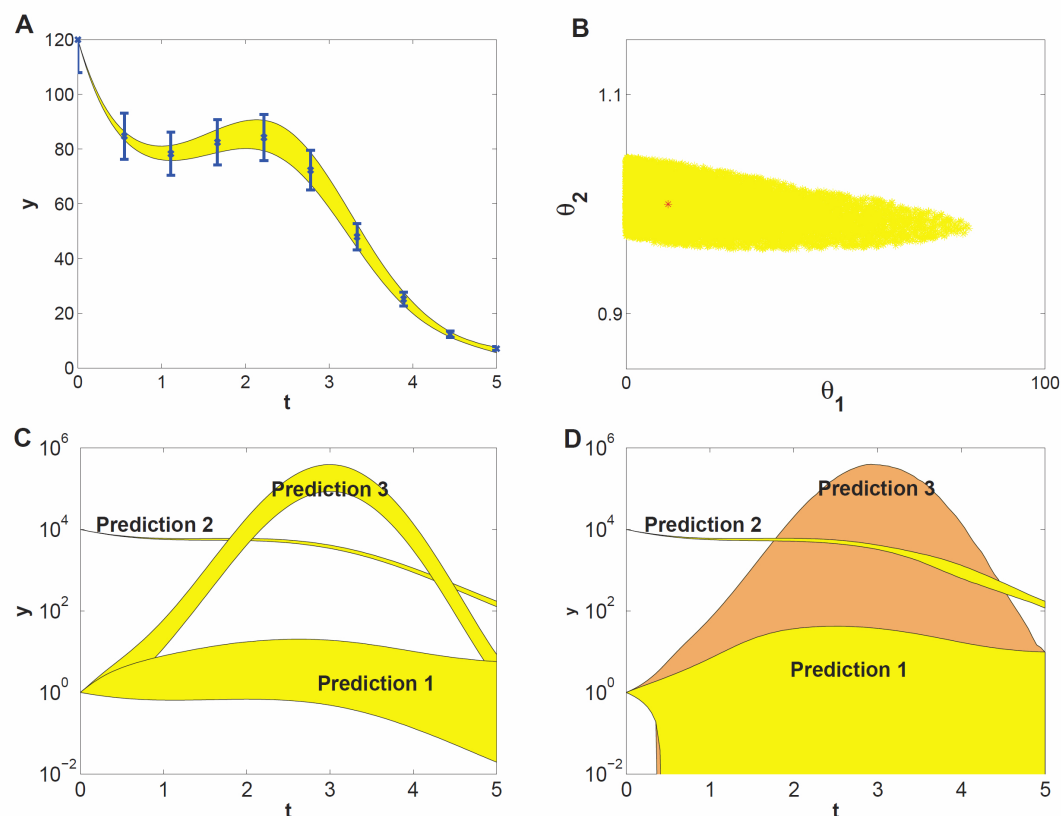
Figure 2. A) The model dynamics fitted to the time series data. The 95 % uncertainty region is displayed in yellow, the error bars denote the standard deviation of the measurement noise. B) Parameter uncertainty represented by the 95 % credible region, containing the true parameter vector (red *). C) Predicted dynamics for different initial concentrations (Predictions 1 and 2), and for a 7 fold amplitude increase of the sinusoidal input (Prediction 3). D) Predicted dynamics via local sensitivities (linearized covariance analysis). The uncertainty of Prediction 3 is plotted in a different shade and includes the uncertainty region of Prediction 1.

This simple example illustrates that prediction uncertainty depends on the type of prediction and is hard to foresee intuitively. We cannot evaluate the prediction uncertainty on the basis of some simple guidelines after only inspecting the credible region of the parameters. Because of the nonlinearities in the system, the credible region is untruthful for evaluating predictions and their uncertainty (Savage, 2012). These results suggest that a full computational uncertainty analysis, as performed here (Fig. 2C), has always to be done in order to estimate the consequences of parameter uncertainty for prediction uncertainty. The key element in this analysis is the sample representing the probability distribution of the parameters.

It is convenient to quantify prediction uncertainty, that is, the variability among predicted time courses. We do so on the basis of the quantity Q defined in equation

254 (7) in the Methods section, which expresses the deviation between two time courses,
255 one being calculated for an arbitrary set of parameter values and the other using the
256 penalized maximum likelihood estimator of the parameters. In our simulation studies,
257 the latter is replaced by the true parameter values. The quantity Q is calculated for each
258 draw of the parameters. The higher the Q value of a prediction, the higher its prediction
259 uncertainty is. Each prediction set yields 1000 values of Q, which are summarized
260 by their 95% percentile, indicated by $Q_{0.95}$. $Q_{0.95}$ values smaller than 1 indicate tight
261 predictions. For example, the $Q_{0.95}$ values for Predictions 1, 2 and 3 in Fig. 2C are 18,
262 0.02 and 1.5, respectively. Even for this simple model the uncertainty among different
263 predictions thus varies by a factor 900 which is an order of 2.9 on logarithmic scale.

Scanning a variety of systems biology models

265 Next, we study prediction uncertainty for six models from the BioModels database
266 (Li et al., 2010). We selected these models to represent some cross section of systems
267 biology models; The systems they represent greatly differ in numbers of variables,
268 parameters, and the types of equations. The parameters in these models represent
269 process rates. The models include circadian rhythm, metabolism, and signalling (Laub
270 and Loomis, 1998; Levchenko et al., 2000; Tyson, 1991; Goldbeter, 1991; Vilar et al.,
271 2002; Leloup and Goldbeter, 1999), see the Supplemental Information for details. For
272 each model, a data set is created in the same manner as described for the simple model
273 above, with parameter values and initial conditions as provided in the database. Each
274 data set was analyzed using an MCMC sampling algorithm yielding a sample of 1000
275 draws representing the distribution of the parameters. This number turned out to be
276 large enough for our prediction purposes as checked in the Supplemental Information.
277 Subsequently, different sets of predictions were made. For each prediction set, the
278 system was perturbed and the time course of the perturbed system was predicted on the
279 basis of each of the 1000 draws of parameter values, whereafter prediction uncertainty
280 was quantified on the basis of $Q_{0.95}$. Perturbations consisted of artificially increasing and
281 decreasing each parameter in turn by a factor of 100 (different factors yielded similar
282 results, see Supplemental Information). For a model with p parameters, we so obtain $2p$
283 different perturbations, prediction sets, and $Q_{0.95}$ values.

284   For the six models the prediction uncertainty is displayed in terms of the $Q_{0.95}$
285 quantifiers (Fig. 3). For the collective fits we used a data set with 10 data points per
286 variable, with an assumed Gaussian noise of $\sigma_i = 0.1 y_{d,i}$ per data point i. (Fig. 3A). The
287 sizes and locations of the $Q_{0.95}$ distribution ranges considerably differ per model. For
288 example, the Tyson model gives mostly precise predictions, and the Laub model gives
289 mostly uncertain predictions. Next, the accuracy of the data was increased artificially by
290 decreasing the noise with a factor 10, and observed the effect on the $Q_{0.95}$ distribution
291 ranges (Fig. 3B). With more accurate data, the $Q_{0.95}$ values are reduced, but not in a
292 very systematic way. For the Vilar model, for example, all predictions have become
293 precise, except for one. More importantly, for all six models the $Q_{0.95}$ values scatter
294 enormously, in ranges that span 2 to 4 orders of magnitude. There is no clear connection
295 between the number of parameters and the maximum $Q_{0.95}$ values or the size of the
296 ranges. By consequence, prediction uncertainty is not a characteristic of a calibrated
297 model, but must be evaluated on a case-by-case basis.
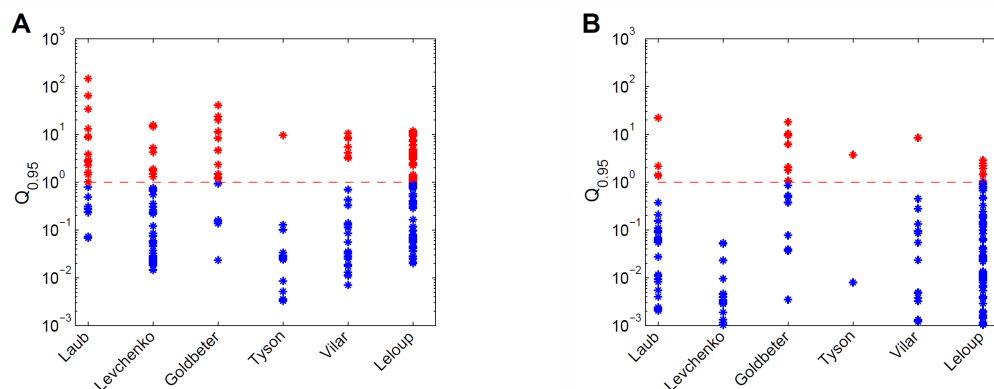
Figure 3. A) Uncertainties of predictions of six different models, expressed in terms of $Q_{0.95}$. A blue * denotes a tight prediction ($Q < 1$), and a red * denotes an uncertain prediction ($Q \geq 1$). B) Prediction uncertainties obtained after reducing parameter uncertainty by a ten-fold decrease in data noise.

DISCUSSION

By introducing the quantifier $Q_{0.95}$ for prediction uncertainty we were able to show that prediction uncertainty is typically not a feature of a calibrated model, but must be calculated on a per-prediction basis using a full computational uncertainty analysis. The adjective *full* stresses that the analysis must use (a representative sample of) the distribution of the parameters and not only the 95% confidence region or credible region of the parameters. The reason is that the parameter values inside the confidence region may be outliers in terms of the prediction, just because of the nonlinearities of the model (Savage, 2012). The key element is thus a faithful representation of the remaining uncertainty in the parameters after fitting the model, in practical terms represented by a sample of parameter values. This sample looks like a dataset with parameters as columns and draws from the distribution as rows (Fig. 1). We used a sample of 1000 nearly independent draws obtained by thinning the MCMC chain. The sample is an approximation of the full posterior distribution of the parameter values. The reliability of this approximation is a point that deserves attention. We checked this by varying the size of the sample and observing whether the results for the quantifier Q showed good convergence. For the models studied in this paper, we found that this convergence was indeed reached if we used a sample size of 1000 points, but for other cases it could happen that this number must be larger.

The value of $Q_{0.95}$ is interpretable. If $Q_{0.95} \geq 1$, the uncertainty is so high that it may obscure biological interpretation of a model prediction, whereas this is unlikely if $Q_{0.95} < 1$. When the uncertainty indeed causes predictions to be biologically ambiguous, $Q_{0.95}$ could be incorporated as a design criterion in experimental design, i.e., additional experiments can be designed or selected to minimize $Q_{0.95}$ (Vanlier et al., 2012; Apgar et al., 2010).

We adopted the Bayesian framework in our analysis. This framework has the advantage that in the collective fits prior information on the parameters can be incorporated, such as from general biological knowledge (Grandison and Morris, 2008) and previous experiments. The prior in a Bayesian analysis is also a means to ensure that a parameter

attains no physically unrealistic values, when fitting the model to data. Current Bayesian MCMC algorithms naturally lead to a sample of the posterior distribution that represents the remaining uncertainty in the parameters. Alternatively, bootstrap sampling (Efron and Tibshirani, 1993) could perhaps be used to create such a sample. MCMC can also be applied to stochastic models. At each iteration the stochastic model has to be run again, so that averaging/summarizing over the chain averages/integrates over the stochastic elements in the model. As this simple approach often leads to very low MCMC acceptance rates, it is advantageous to treat the stochastic elements as additional unknowns in the MCMC with a Metropolis-within-Gibbs approach (ter Braak and Vrugt, 2008). Particle MCMC is a promising method in this area in which proposals for the additional unknowns are made via a particle filter (Andrieu et al., 2010; Vrugt et al., 2013). Intersubject variability is a simple form of a stochastic model in which only the parameters of the model vary among subjects. The populations of models approach (Britton et al., 2013), which is related to the Generalized Likelihood Uncertainty Estimation (GLUE) approach (Beven and Binley, 1992) that is popular in hydrology, is a way to deal with such variability. The formal Bayes approach (and thus MCMC) can also be used. Moreover, it can disentangle inter-subject variability (which cannot be reduced) and uncertainty (which can in principle be reduced by collecting more data) as demonstrated for a pharmacokinetic model in ter Braak and Vrugt (2008). A comparison of formal Bayes and GLUE can be found in Vrugt et al. (2009).

Of course there are computational issues in the Bayesian approach which increase with the number of the parameters and the complexity of the model. Gutenkunst et al. (2007a) speeded up their Monte Carlo analysis (a random walk Metropolis algorithm) by periodically updating the step size based on the current Fisher Information Matrix, but such an updating scheme is generally not recommended as it may change the stationary distribution of the chain (Roberts and Rosenthal, 2007). We used the adaptive MCMC algorithm DE-MCz (ter Braak and Vrugt, 2008), that automatically adapts to the size and shape of the posterior distribution. DE-MCz allows parallel computation when the chains have access to a common store of past-sampled parameter values, and is designed to quickly find and explore very elongated subspaces in the posterior distribution of the parameters, which are due to sloppiness. Such subspaces point to problems with parameter identifiability. An alternative approach to explore these problems is via profile likelihood methods (Raue et al., 2009, 2010).

Systems biology models may be computationally expensive to evaluate, making standard MCMC algorithms impractical as they require many evaluations. For our class of dynamic computational models the likelihood of each sample is based on the time integration of the model, which can be very demanding for large or stiff models. To confine sampling and integration costs, model order reduction techniques may be of help to speed up analysis (see Supplemental Information). However, the reduction errors affect prediction uncertainty differs per model and per prediction.

An alternative approach is to locally approximate the likelihood function using parameter sensitivities. Applications in systems biology of the latter idea include a gene network model in a Drosophila embryo (Ashyraliyev et al., 2008), a microbial growth model (Schittkowski, 2007), and an in-silico gene regulatory network (Zak et al., 2003a). This method is simple and has a very low computational demand since it avoids sampling, but may lead to a highly erroneous estimation of prediction uncertainty

(Gutenkunst et al., 2007a), as we demonstrated (Fig. 2D). A better alternative is to use emulators and use adaptive error modelling within the MCMC algorithm (Cui et al., 2011).

In this paper the analysis is carried out with noiseless data. The Hessian matrix has full rank for each model, so all models are theoretically locally identifiable in the neighbourhood of the penalized maximum likelihood parameter. However, in practice the PML parameter might be hard to find or be non-unique due to noisy data. An alternative is to define Q with respect to the average time course of the log-predictions (that is, averaging across the sample of parameter values), as is mentioned in the end of the Methods section. With noisy data, the uncertainty will increase compared to noiseless data, but the range of uncertainties across different predictions, as expressed by quantifiers such as Q, is likely to be similar to that in Fig. 3.

For biologists it is not always essential to obtain quantitative predictions and a reviewer asked about the uncertainty analysis of qualitative predictions. For this, the first step is to convert any quantitative predicted time course into a qualitative prediction, such as the statement that a certain metabolite or flux is going up, when a gene is over-expressed. Once this step has been decided upon, the prediction for any particular set of parameter values in the posterior sample is then either a simple *yes* (1) or *no* (0) or a value between 0 and 1 that measures the degree to which a single predicted time course corresponds with the statement. The second step is then to determine the mean and variance of the values obtained across the parameter sets in the posterior sample, which measure the extent to which the model predicts what is stated, and the prediction uncertainty, respectively. Note that for crisp (0/1) outcomes the mean is simply the fraction f of the parameter sets that yields a *yes*, and the variance is $f(1-f)$.

CONCLUSIONS

Our survey of a variety of models shows that prediction uncertainty is hard to predict. It turns out that it is practically impossible to establish the predictive power of a model without a full computational uncertainty analysis of each individual prediction. In practice this has strong consequences for the way models are transferred via the literature and databases. We conclude that publication of a model in the open literature or a database should not only involve the listing of the model equations, the parameter values with or without confidence intervals or parameter sensitivities, but should also include a sample of, say, 1000 draws representing the full (posterior) distribution of the parameters. Only then prospective users are able to reliably perform a full computational uncertainty analysis if they intend to use the model for prediction purposes. To this end, a software package is made available online, including an exemplary parameter sample. In addition we encourage to include experimental conditions, model assumptions, and used prior knowledge in order to make the sample reproducible. This will decrease the chances of error propagation, for example due to programming errors.

ACKNOWLEDGEMENTS

The authors would like to thank the reviewers for their comments and W. Kruijer, H. van der Voet, S. Schnabel, and E. Boer for useful discussions.

REFERENCES

Alon, U. (2006). *An Introduction to Systems Biology. Design Principles of Biological Circuits*. Chapman & Hall/CRC.

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342.

Apgar, J. F., Witmer, D. K., White, F. M., and Tidor, B. (2010). Sloppy models, parameter uncertainty, and the role of experimental design. *Molecular BioSystems*, 6(10):1890–1900.

Ashyraliyev, M., Jaeger, J., and Blom, J. G. (2008). Parameter estimation and determinability analysis applied to Drosophila gap gene circuits. *BMC Systems Biology*, 2:83.

Beck, M. B. (1987). Water quality modeling: A review of the analysis of uncertainty. *Water Resources Research*, 23(8):1393–1442.

Beven, K. and Binley, A. (1992). The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes*, 6(3):279–298.

Britton, O. J., Bueno-Orovio, A., Van Ammel, K., Lu, H. R., Towart, R., Gallacher, D. J., and Rodriguez, B. (2013). Experimentally calibrated population of models predicts and explains intersubject variability in cardiac cellular electrophysiology. *Proceedings of the National Academy of Sciences*, 110(23):E2098–E2105.

Brown, K. S., Hill, C. C., Calero, G. A., Myers, C. R., Lee, K. H., Sethna, J. P., and Cerione, R. A. (2004). The statistical mechanics of complex signaling networks: nerve growth factor signaling. *Physical biology*, 1(3):184.

Brown, K. S. and Sethna, J. P. (2003). Statistical mechanical approaches to models with many poorly known parameters. *Physical Review E*, 68(2):021904.

Buchler, N. E. and Louis, M. (2008). Molecular titration and ultrasensitivity in regulatory networks. *Journal of molecular biology*, 384(5):1106–1119.

Cui, T., Fox, C., and O'Sullivan, M. J. (2011). Bayesian calibration of a large-scale geothermal reservoir model by a new adaptive delayed acceptance Metropolis Hastings algorithm. *Water Resources Research*, 47(10).

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*, volume 57. Chapman & Hall/CRC.

Goldbeter, A. (1991). A minimal cascade model for the mitotic oscillator involving cyclin and cdc2 kinase. *Proceedings of the National Academy of Sciences*, 88(20):9107–9111.

Grandison, S. and Morris, R. J. (2008). Biological pathway kinetic rate constants are scale-invariant. *Bioinformatics*, 24(6):741–743.

Gutenkunst, R. N., Casey, F. P., Waterfall, J. J., Myers, C. R., and Sethna, J. P. (2007a). Extracting falsifiable predictions from sloppy models. *Annals of the New York Academy of Sciences*, 1115(1):203–211.

Gutenkunst, R. N., Waterfall, J. J., Casey, F. P., Brown, K. S., Myers, C. R., and Sethna, J. P. (2007b). Universally sloppy parameter sensitivities in systems biology models. *PLoS computational biology*, 3(10):e189.

Hawkins, E. and Sutton, R. (2009). The potential to narrow uncertainty in regional climate predictions. *Bulletin of the American Meteorological Society*, 90(8):1095–

460   1107.

461   Jaynes, E. T. (2003). *Probability theory: the logic of science*. Cambridge university
462   press.

463   Klipp, E., Herwig, R., Kowald, A., Wierling, C., and Lehrach, H. (2008). *Systems*
464   *biology in practice: concepts, implementation and application*. John Wiley & Sons.

465   Kreutz, C., Raue, A., Kaschek, D., and Timmer, J. (2013). Profile likelihood in systems
466   biology. *FEBS Journal*, 280(11):2564–2571.

467   Laub, M. T. and Loomis, W. F. (1998). A molecular network that produces sponta-
468   neous oscillations in excitable cells of Dictyostelium. *Molecular biology of the cell*,
469   9(12):3521–3532.

470   Leloup, J. C. and Goldbeter, A. (1999). Chaos and birhythmicity in a model for circadian
471   oscillations of the PER and TIM proteins in Drosophila. *Journal of Theoretical*
472   *Biology*, 198(3):445–459.

473   Levchenko, A., Bruck, J., and Sternberg, P. W. (2000). Scaffold proteins may biphasically
474   affect the levels of mitogen-activated protein kinase signaling and reduce its threshold
475   properties. *Proceedings of the National Academy of Sciences*, 97(11):5818–5823.

476   Li, C., Donizelli, M., Rodriguez, N., Dharuri, H., Endler, L., Chelliah, V., Li, L., He, E.,
477   Henry, A., Stefan, M. I., et al. (2010). Biomodels database: An enhanced, curated and
478   annotated resource for published quantitative kinetic models. *BMC systems biology*,
479   4(1):92.

480   MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cam-
481   bridge university press.

482   Maerkl, S. J. and Quake, S. R. (2007). A systems approach to measuring the binding
483   energy landscapes of transcription factors. *Science*, 315:233–237.

484   Omlin, M. and Reichert, P. (1999). A comparison of techniques for the estimation of
485   model prediction uncertainty. *Ecological Modelling*, 115(1):45 – 59.

486   Raue, A., Becker, V., Klingmüller, U., and Timmer, J. (2010). Identifiability and
487   observability analysis for experimental design in nonlinear dynamical models. *Chaos:*
488   *An Interdisciplinary Journal of Nonlinear Science*, 20(4):045105.

489   Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U.,
490   and Timmer, J. (2009). Structural and practical identifiability analysis of partially
491   observed dynamical models by exploiting the profile likelihood. *Bioinformatics*,
492   25(15):1923–1929.

493   Raue, A., Kreutz, C., Theis, F. J., and Timmer, J. (2013). Joining forces of Bayesian and
494   frequentist methodology: a study for inference in the presence of non-identifiability.
495   *Philosophical Transactions of the Royal Society A: Mathematical, Physical and*
496   *Engineering Sciences*, 371(1984):20110544.

497   Robert, C. P., Marin, J., and Rousseau, J. (2011). *Bayesian Inference and Computation*,
498   pages 39–65. John Wiley & Sons, Ltd.

499   Roberts, G. O. and Rosenthal, J. S. (2007). Coupling and ergodicity of adaptive Markov
500   chain Monte Carlo algorithms. *Journal of applied probability*, pages 458–475.

501   Savage, S. L. (2012). *The flaw of averages: Why we underestimate risk in the face of*
502   *uncertainty*. Wiley.

503   Schittkowski, K. (2007). Experimental design tools for ordinary and algebraic differen-
504   tial equations. *Industrial & Engineering Chemistry Research*, 46(26):9137–9147.

505   Stumpf, M. P. H., Balding, D. J., and Girolami, M. (2011). *Handbook of Statistical*

Systems Biology. Wiley Online Library.

ter Braak, C. J. F. and Vrugt, J. A. (2008). Differential Evolution Markov Chain with snooker updater and fewer chains. *Statistics and Computing*, 18:435–446.

Teusink, B., Passarge, J., Reijenga, C. A., Esgalhado, E., van der Weijden, C. C., Schepper, M., Walsh, M. C., Bakker, B. M., van Dam, K., Westerhoff, H. V., et al. (2000). Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. *European Journal of Biochemistry*, 267(17):5313–5329.

Tyson, J. J. (1991). Modeling the cell division cycle: cdc2 and cyclin interactions. *Proceedings of the National Academy of Sciences*, 88(16):7328–7332.

van Mourik, S., van Dijk, A. D., de Gee, M., Immink, R. G., Kaufmann, K., Angenent, G. C., van Ham, R. C., and Molenaar, J. (2010). Continuous-time modeling of cell fate determination in Arabidopsis flowers. *BMC systems biology*, 4(1):101.

Vanlier, J., Tiemann, C. A., and Hilbers et al., P. A. J. (2012). A Bayesian approach to targeted experiment design. *Bioinformatics*.

Vilar, J. M., Kueh, H. Y., Barkai, N., and Leibler, S. (2002). Mechanisms of noise-resistance in genetic oscillators. *Proceedings of the National Academy of Sciences*, 99(9):5988–5992.

Vrugt, J. A., ter Braak, C. J., Diks, C. G., and Schoups, G. (2013). Hydrologic data assimilation using particle Markov chain Monte Carlo simulation: Theory, concepts and applications. *Advances in Water Resources*, 51:457–478.

Vrugt, J. A., Ter Braak, C. J., Gupta, H. V., and Robinson, B. A. (2009). Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling? *Stochastic Environmental Research and Risk Assessment*, 23(7):1011–1026.

Zak, D. E., Gonye, G. E., Schwaber, J. S., and Doyle, F. J. (2003a). Importance of input perturbations and stochastic gene expression in the reverse engineering of genetic regulatory networks: insights from an identifiability analysis of an in silico network. *Genome research*, 13(11):2396–2405.

Zak, D. E., Pearson, R. K., Vadigepalli, R., Gonye, G. E., Schwaber, J. S., and Doyle Iii, F. J. (2003b). Continuous-time identification of gene expression models. *Omics A Journal of Integrative Biology*, 7(4):373–386.

Figure 1

Flowchart of parameter estimation and uncertainty analysis

Flowchart of parameter estimation and uncertainty analysis. The model dynamics are fitted to noisy time series data, typically by an MCMC algorithm, resulting in a sample of parameter values representing their posterior distribution. Typical applications of this distribution are the computation of a credible region of parameters (showing parameter uncertainty) and full computational uncertainty analysis of predictions.
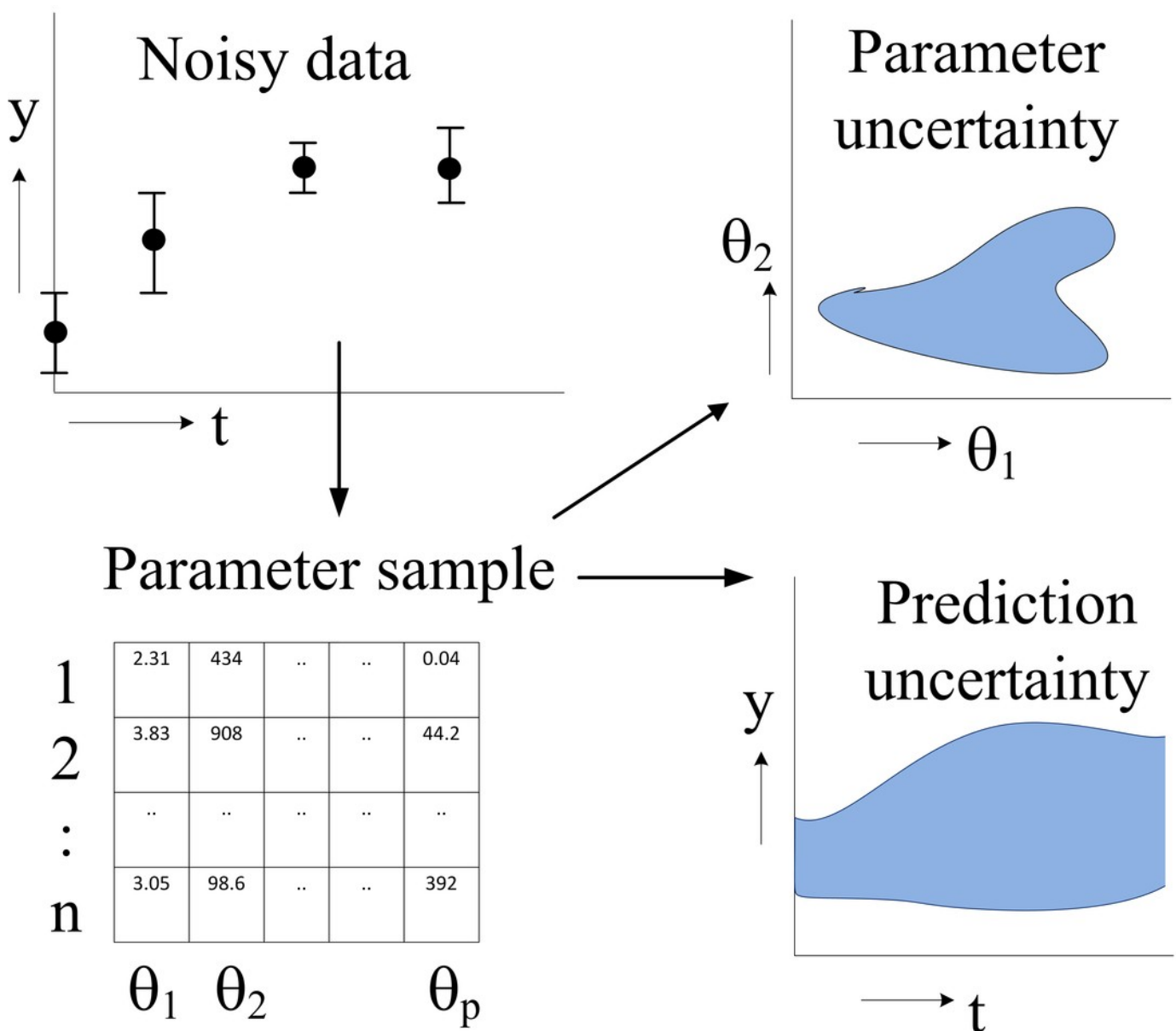
Figure 2

Prediction uncertainty following from parameter uncertainty and computational methods.

A) The model dynamics fitted to the time series data. The 95 % uncertainty region is displayed in yellow, the error bars denote the standard deviation of the measurement noise. B) Parameter uncertainty represented by the 95 % credible region, containing the true parameter vector (red *). C) Predicted dynamics for different initial concentrations (Predictions 1 and 2), and for a 7 fold amplitude increase of the sinusoidal input (Prediction 3). D) Predicted dynamics via local sensitivities (linearized covariance analysis). The uncertainty of Prediction 3 is plotted in a different shade and includes the uncertainty region of Prediction 1.
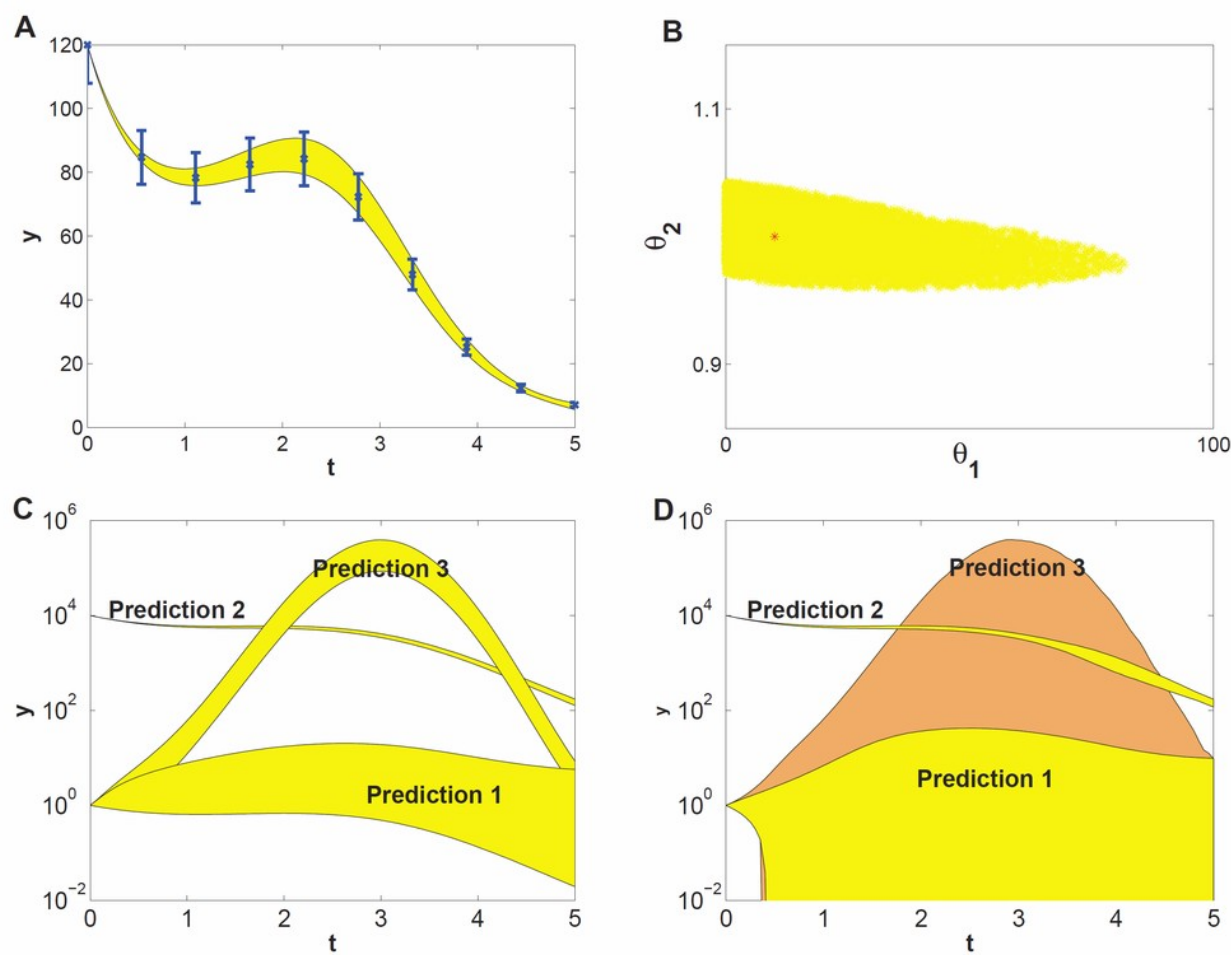
Figure 3

Comparison of prediction uncertainties for six different models

A) Uncertainties of predictions of six different models, expressed in terms of $Q_{0.95}$. A blue * denotes a tight prediction ($Q_{0.95} < 1$), and a red * denotes an uncertain prediction ($Q_{0.95} \geq 1$). B) Prediction uncertainties obtained after reducing parameter uncertainty by a ten-fold decrease in data noise.