Selecting fitted models under epistemic uncertainty using a stochastic process on quantile functions

PUBLISHED IN NATURE COMMUNICATIONS ON 23 OCT. 2025

Alexandre René[*,1, 2, 3] and **André Longtin**[1, 4, 5]

[1]Department of Physics, University of Ottawa, Canada
[2]Informatik, Fakultät 1, RWTH Aachen, Germany
[3]Physik, Fakultät 1, RWTH Aachen, Germany
[4]Department of Cellular and Molecular Medicine, University of Ottawa, Canada
[5]Center for Neural Dynamics, University of Ottawa, Canada

Fitting models to data is an important part of the practice of science. Advances in machine learning have made it possible to fit more — and more complex — models, but have also exacerbated a problem: when multiple models fit the data equally well, which one(s) should we pick? The answer depends entirely on the modelling goal. In the scientific context, the essential goal is *replicability*: if a model works well to describe one experiment, it should continue to do so when that experiment is replicated tomorrow, or in another laboratory. The selection criterion must therefore be robust to the variations inherent to the replication process. In this work we develop a nonparametric method for estimating uncertainty on a model's empirical risk when replications are non-stationary, thus ensuring that a model is only rejected when another is *reproducibly* better. We illustrate the method with two examples: one a more classical setting, where the models are structurally distinct, and a machine learning-inspired setting, where they differ only in the value of their parameters. We show how, in this context of replicability or "epistemic uncertainty", it compares favourably to existing model selection criteria, and has more satisfactory behaviour with large experimental datasets.

1 Introduction

Much of our understanding of the natural world is built upon mathematical models. But how we build those models evolves as new techniques and techonologies are developed. With the arrival of machine learning methods, it has become even more feasible to solve inverse problems and learn complex descriptions by applying data-driven methods to scientifically-motivated models.[10,54,27] However, even moderately complex models are generally *non-identifiable*: many different parameter sets may yield very similar outputs.[7,65,48] With imperfect real-world data, it is often unclear which — if any — of these models constitutes a trustable solution to the inverse problem.

This presents us with a model selection problem which is somewhat at odds with the usual statistical theory. First, we require a criterion which can distinguish between structurally identical models differing only in their *specific parameters*. Moreover, even with the most tightly controlled experiments, we expect all candidate models to be *misspecified* due to experimental variations. Those variations also *change* when the experiment is repeated in the same lab or another lab; in other words, the replication process is *non-stationary*, even when the data-generation process is stationary within individual trials. While there has been some work to develop model selection criteria which are robust to misspecification,[22,39,29] and some consideration of locally stationary processes,[68] the question of non-stationary replications has received little attention. Criteria which assume the data generating process to be stationary therefore underestimate *epistemic uncertainty*, which has important repercussions when they are used for scientific induction, where the goal is to learn a model which *generalises*. In recent years, a few authors have advanced similar arguments in the context of machine learning models.[13,74]

To support data-driven scientific methods, there is therefore a need for a criterion which more accurately estimates this uncertainty due to differences between our model and the true data generating process. Epistemic uncertainty may be due to learning the model from limited samples, the model being misspecified or non-identifiable, or the data generating process being non-stationary. We distinguish this from *aleatoric* uncertainty (also called sampling uncertainty),[33,30] which is due to the intrinsic stochasticity of the process.

Some of the pioneering work on this front came from the geosciences, which methods like GLUE[6,62,8] and Bayesian calibration.[32] More recently, we have also seen strategies to account for epistemic uncertainty in machine learning,[15,30] astrophysics[37] and condensed matter physics.[31] All of these employ some form of ensemble: epistemic uncertainty is represented as a concrete distribution over models, and predictions are averaged over this distribution.

Unfortunately, ensemble models are difficult to interpret since they are generally invalid in a Bayesian sense.[18] In our view, if the goal is to find interpretable models, then an approach like that advocated by Tarantola[65] seems more appropriate: instead of assigning probabilities to an ensemble of

*Corresponding author: a.rene@physik.rwth-aachen.de

plausible models, simply treat that ensemble as the final inference result. With a finite (and hopefully small) number of plausible models, each can be interpreted individually.

The high-level methodology Tarantola proposes comes down to the following: 1. Construct a (potentionally large) set of *candidate models*. 2. Apply a *rejection criterion* to each candidate model in the set. 3. Retain from the set of candidate models those that satisfy the criterion.

Step 1 can be accomplished in different ways; for example, Prinz *et al.*[48] performed a grid search using a structurally fixed model of the lobster's pyloric rhythm, and found thousands of distinct parameter combinations which reproduce chosen features of the recordings. More recently, machine learning methods have also been used to learn rich mechanistic models of molecular forces,[10] cellular alignment[27] and neural circuits.[24,54] Since these are intrinsically nonlinear models, their objective landscape contains a multitude of local minima, which translates to a multitude of candidate models. For a general formulation of the learning problem which fits our framing especially well, see Vapnik's *Function estimation model*.[69]

The focus of our work is to present a practical criterion for step 2: **we therefore assume that we have already obtained a set of candidate models**. We make only three hard requirements. First, a candidate model $\mathcal{M}$ must be probabilistic, taking the form

$$p(y_i \mid x_i; \mathcal{M}) \,, \quad (x_i, y_i) \in \mathcal{D}_{\text{test}} \,, \tag{1}$$

where $\mathcal{D}_{\text{test}}$ is the observed dataset and $(x_i \in \mathbb{R}^n, y_i \in \mathbb{R}^m)$ is the i-th input/output pair (or independent/dependent variables) that was observed. Often this takes the form of a mechanistic model with some additional observation noise. Second, it must be possible to generate arbitrarily many synthetic sample pairs (x, y) following each candidate model's distribution. And third, any candidate model $\mathcal{M}$ must assign a non-vanishing probability to each of the observations:

$$p(y_i \mid x_i; \mathcal{M}) > 0 \,, \quad \forall (x_i, y_i) \in \mathcal{D}_{\text{test}} \,. \tag{2}$$

For example, a model whose predictions y are restricted to an interval $[a, b]$ is only allowed if all observations are within that interval. We also require the definition of a loss function Q to quantify the accuracy of model predictions y_i. Taking the expectation of Q over data samples yields the *risk* R, a standard measure of performance used to fit machine learning models.

Key to our approach is a novel method for assigning to each model a *risk*- or *R-distribution* representing epistemic uncertainties due either to finite samples, model misspecification or non-stationarity. Only when two models have sufficiently non-overlapping R-distributions do we reject the one with higher R. This approach allows us to compare any number of candidate models, by reducing the problem to a sequence of pairwise comparisons. We make no assumption on the structure of those models: they may be given by two completely different sets of equations, or they may have the same equations but differ only in their parameters, as long as they can be cast in the form of equation (1).

In many cases, models will contain a "physical" component — the process we want to describe — and an "observation" component — the unavoidable experimental noise. Distinguishing these components is often useful, but it makes no difference from the point of view of our method: only the combined "physical + observation" model matters. In their simplest forms, the physical component may be deterministic and the observation component may be additive noise, so that the model is written as a deterministic function f plus a random variable ξ affecting the observation:

$$y_i = f(x_i; \mathcal{M}) + \xi_i \,.$$

In such a case, the probability in equation (1) reduces to $p(\xi_i = y_i - f(x_i; \mathcal{M}))$; if ξ_i is Gaussian with variance σ^2, this further reduces to $p(y_i \mid x_i; \mathcal{M}) \propto \exp\!\big(-(y_i - f(x_i; \mathcal{M}))^2 / 2\sigma^2\big)$. Of course, in many cases the model itself is stochastic, or the noise is neither additive nor Gaussian. Moreover, even when such assumptions seem justified, we should be able to test them against alternatives.

We will illustrate our proposed criterion using two examples from biology and physics. The first compares candidate models of neural circuits (from the aforementioned work of Prinz *et al.*[48]) with identical structure but different parameters; the second compares two structually different models of black body radiation, inspired by the well-known historical episode of the "ultraviolet catastrophe".

In the next few sections, we present the main steps of our analysis using the neural circuit model for illustration. We start by arguing that modelling discrepancies are indicative of epistemic uncertainty. We then use this idea to assign uncertainty to each model's risk, and thereby define the *EMD rejection rule*. Here EMD stands for *empirical modelling discrepancy*, which describes the manner in which we estimate epistemic uncertainty; this mainly involves three steps, schematically illustrated in Fig. 1. First, we represent the prediction accuracy of each model with a quantile function q^* of the loss. Second, by measuring the self-consistency of q^* with the model's own predictions $(\tilde{q})$, we obtain a measure (δ^{EMD}) of epistemic uncertainty, from which we construct a stochastic process $\mathfrak{Q}$ over quantile functions. Since each realisation of $\mathfrak{Q}$ can be integrated to yield the risk, $\mathfrak{Q}$ thus induces the R-distributions we seek. This requires however the introduction of a new type of stochastic process, which we call *hierarchical beta process*, in order to ensure that realizations are valid quantile functions. Those R-distributions are used to compute the tail probabilities, denoted B^{EMD}, in terms of which the rejection rule is defined. The third step is a calibration procedure, where we validate the estimated probabilities B^{EMD} on a set of simulated experiments. To ensure the soundness of our results, we used over 24,000 simulated experiments, across 16 forms of experimental variation.

The following sections then proceed with a systemic study of the method, using the simpler model of black body radiation. We illustrate how modelling errors and observation noise interact to affect the shape of R distributions and ultimately the EMD rejection criterion. Finally, we compare our method with a number of other popular criteria — AIC, BIC, MDL, elpd and Bayes factors. We highlight the major modelling assumptions each method makes, and therefore in which cir-

cumstances it is likely most appropriate. Special attention is paid to the behaviour of the model selection statistics as the sample size grows, which is especially relevant to the scientific context.

2 Results

2.1 The risk as a selection criterion for already fitted models

As set out in the Introduction, we start from the assumption that models of the natural phenomenon of interest have already been fitted to data; this anterior step is indicated by the grey faded rectangle in Fig. 1. The manner in which this is done is not important; model parameters for example could result from maximizing the likelihood, running a genetic algorithm or performing Bayesian inference.

Our starting points are the true data-generating process $\mathcal{M}_{\text{true}}$ and the pointwise loss Q. The loss depends on the model $\mathcal{M}_A$, evaluates on individual data samples (x_i, y_i) and returns a real number:

$$Q(x_i, y_i; \mathcal{M}_A) \in \mathbb{R}\,. \tag{3}$$

Often, but not always, the negative log likelihood is chosen as the loss. What we seek to estimate is the risk, i.e. the expectation of Q:

$$R_A = \mathbb{E}_{(x_i,y_i)\sim\mathcal{M}_{\text{true}}}[Q(x_i, y_i; \mathcal{M}_A)]; \tag{4}$$

we use the notation $(x_i, y_i) \sim \mathcal{M}_{\text{true}}$ to indicate that the samples (x_i, y_i) are drawn from $\mathcal{M}_{\text{true}}$. For simplicity, our notation assumes that model parameters are point estimates. For Bayesian models, one can adapt the definition of the loss Q to include an expectation over the posterior; equation (4) would then become the Bayes risk. Our methodology is agnostic to the precise definition of Q, as long as it is evaluated pointwise; it can differ from the objective used to fit the models.

The risk describes the expected performance of a model on replicates of the dataset, when each replicate is drawn from the same data-generating process $\mathcal{M}_{\text{true}}$. It is therefore a natural basis for comparing models based on their ability to generalise. It is also the gold standard objective for a machine learning algorithm,[69,70] for the same reason, and has been used in classical statistics to analyse model selection under misspecification.[14] In practice we usually estimate R_A from finite samples with the *empirical risk*:

$$\hat{R}_A = \frac{1}{|\mathcal{D}|}\sum_{(x_i,y_i)\in\mathcal{D}} Q(x_i, y_i; \mathcal{M}_A)\,. \tag{5}$$

For the types of problems we are interested in, it is safe to assume that $\hat{R}_A$ converges to R_A in probability.

A common consideration with model selection criteria is whether they are *consistent*, i.e. whether they eventually select the true model when the number of samples grows. When models are misspecified, none of the candidates are actually the true model, so the definition of consistency is often generalised to mean convergence to the "pseudo-true" model: the one with the smallest Kullback-Leibler divergence D_{KL} from the true model.[11,26] If consistency in this sense is desired, then using the log likelihood for Q is recommended, since for a fixed true model, minimizing the D_{KL} is equivalent to minimizing the log likelihood. The log likelihood is also known to be *proper* (see e.g. references 21, 46 or 19 (Chap. 7.1)); a selection rule which minimizes the log likelihood is thus also consistent in the original sense of asymptotically selecting the true model when it is among the candidates.

The statistical learning literature instead defines a "learning machine" as *consistent* if it asymptotically minimizes the true risk R when trained with the empirical risk $\hat{R}_A$.[69,70] As long as we have a finite number of candidate models, our procedure will satisfy this notion of consistency.

Convergence to R_A also means that the result of a risk-based criterion is insensitive to the size of the dataset, provided we have enough data to estimate the risk of each model accurately. This is especially useful in the scientific context, where we want to select models which can be used on datasets of any size.

In addition to defining a consistent, size-insensitive criterion, risk also has the important practical advantage that it remains informative when the models are structurally identical. We contrast this with the marginal likelihood (also known as the model evidence), which for some dataset $\mathcal{D}$, model $\mathcal{M}_A$ parametrised by θ, and prior π_A, would be written

$$\mathcal{E}_A = \int p(\,\mathcal{D}\,|\,\mathcal{M}_A(\theta)\,)\,\pi_A(\theta)d\theta\,. \tag{6}$$

Since it is integrated over the entire parameter space, $\mathcal{E}_A$ characterizes the family of models $\mathcal{M}_A(\cdot)$, but says nothing of a specific parametrisation $\mathcal{M}_A(\theta)$. We include a comparison of our proposed yperref[thm:emdrejectionrule].

A criterion based on risk however will not account for a model which overfits the data, which is why it is **important to use separate datasets for fitting and comparing models**. This in any case better reflects the usual scientific paradigm of recording data, forming a hypothesis, and then recording new data (either in the same laboratory or a different one) to test that hypothesis. Splitting data into training and test sets is also the standard practice in machine learning to avoid overfitting and ensure better generalisation. In the following therefore we will always denote the observation data as $\mathcal{D}_{\text{test}}$ to stress that they should be independent from the data to which the models were fitted.

In short, the EMD rejection rule we develop in the following sections ranks models based on their empirical risk equation (5), augmented with an uncertainty represented as a risk-distribution. Comparisons based on the risk are consistent and insensitive to the number of test samples, but require a separate test dataset to avoid overfitting.

2.2 Example application: Selecting among disparate parameter sets of a biophysical model

To explain the construction of the EMD rejection rule, we use the dynamical model for a neuron membrane potential described by Prinz *et al.*[48]~.[48] This choice was motivated by the fact that fitting a neuron model is a highly ill-posed problem, and therefore corresponds to the situation we set out in the

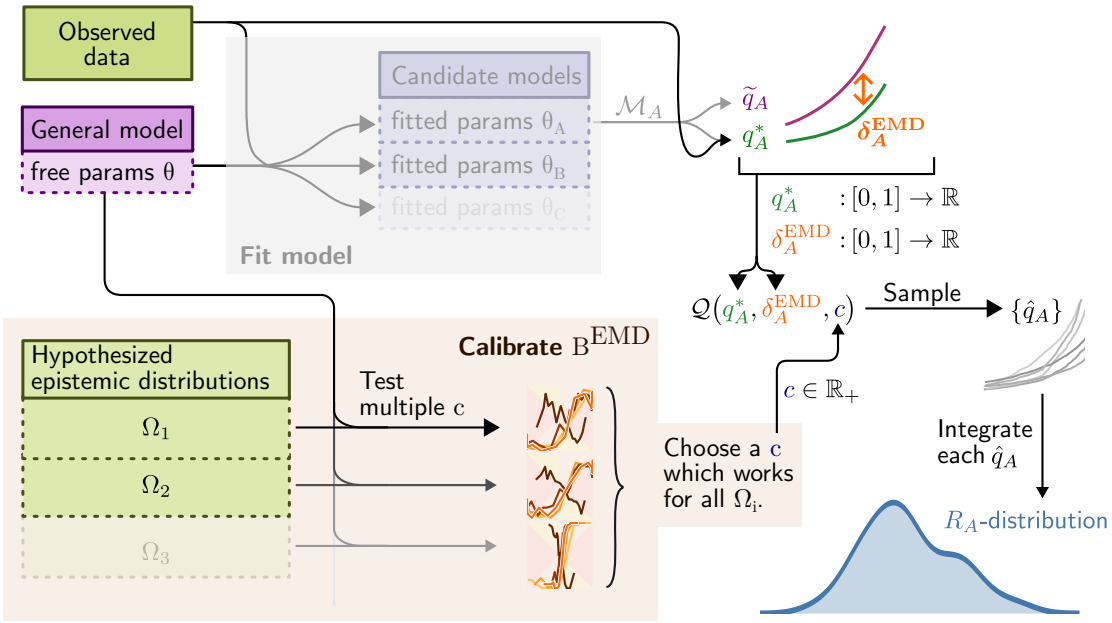


Figure 1: **Overview of the approach for computing R-distributions.** We assume the model has already been fitted to obtain a set of *candidate parameter sets* θ_A, θ_B…(not shown: the models may also be structurally different). Each candidate parameter set Θ_A defines a *candidate model* $\mathcal{M}_A$, for which we describe the statistics of the loss with two different quantile functions: the purely synthetic $\tilde{q}_A$ (equation (22)) which depends only on the model, and the mixed q_A^* (equation (19)) which depends on both the model and the data. A small discrepancy δ_A^{EMD} (equation (23)) between those two curves indicates that model predictions concord with the observed data. Both q_A^* and δ_A^{EMD} are then used to parametrise a stochastic process $\mathfrak{Q}$ which generates random quantile functions. This induces a distribution for the risk R_A of model $\mathcal{M}_A$, which we ascribe to *epistemic uncertainty*. The stochastic process $\mathfrak{Q}$ also depends on a global scaling parameter c. This is independent of the specific model, and is obtained by calibrating the procedure with simulated experiments Ω_i that reflect variations in laboratory conditions. The computation steps on the right (white background) have been packaged as available software.[52]

Introduction: disparate sets of parameters for a structurally fixed model which nevertheless produce similar outputs. We will focus on the particular LP neuron type; Prinz *et al.*[48] find five distinct parameter sets which reproduce its experimental characteristics. Throughout this work, we reserve `LP 1` as the model that generates through simulation the *true data* $\mathcal{D}_{\mathrm{test}}$, and use `LP 2` to `LP 5` to define the *candidate models* $\mathcal{M}_A$ to $\mathcal{M}_D$ which we compare against $\mathcal{D}_{\mathrm{test}}$. We intentionally exclude `LP 1` from the candidate parameters to emulate the typical situation where none of the candidate models fit the data perfectly. Visual inspection of model outputs suggests that two candidates (models $\mathcal{M}_A$ and $\mathcal{M}_B$) are more similar to the true model output (Fig. 2). We will show that our method not only concords with those observations, but makes the similarity between models quantitative.

It is important to note that we treat the models $\mathcal{M}_A$ to $\mathcal{M}_D$ as simply given. We make no assumption as to how they were obtained, or whether they correspond to the maximum of a likelihood. We chose this neural example, where parameters are obtained with a multi-stage, semi-automated grid search, partly to illustrate that our method is agnostic to the fitting procedure.

The possibility of comparing models visually was another factor in choosing this example. Since numbers can be misleading, this allows us to confirm that the method works as expected. Especially for our target application where all candidate models share the same equation structure, a visual validation is key to establishing the soundness of the B^{EMD}, since none of the established model selection like the Bayes factor, AIC or WAIC[71,19] are applicable. Indeed, these other methods only compare alternative equation structures, not alternative parameter sets. We include a more in-depth comparison to these other methods at the end of the Results.

The datasets in this example take the form of one-dimensional time series, with the time $t \in \mathcal{T}$ as the independent variable and the membrane potential $V^{\mathrm{LP}} \in \mathcal{V}$ as the dependent variable. We denote the space of all possible time-potential tuples $\mathcal{T}\times\mathcal{V}$. Model specifics are given in the Methods; from the point of view of model selection, what matters is that we have ways of generating series of these time-potential tuples: either using the true data-generating process ($\mathcal{M}_{\mathrm{true}}$) or one of the candidate models ($\mathcal{M}_A$ to $\mathcal{M}_D$).

We will assume that the dataset used to evaluate models is composed of L samples, with each sample a $(t, \tilde{V}^{\mathrm{LP}})$ tuple:

$$\mathcal{D}_{\mathrm{test}} = \left\{ \left(t_k, \tilde{V}^{\mathrm{LP}}(t_k; \mathcal{M}_{\mathrm{true}}, I_{\mathrm{ext}})\right) \in \mathcal{T}\times\mathcal{V} \right\}_{k=1}^{L} . \tag{7}$$

The original model by Prinz *et al.*[48] produced deterministic traces V^{LP}. Experimental measurements however are variable, and our approach depends on that variability. For this work we therefore augment the model with two sources of stochasticity. First, the system as a whole receives a coloured

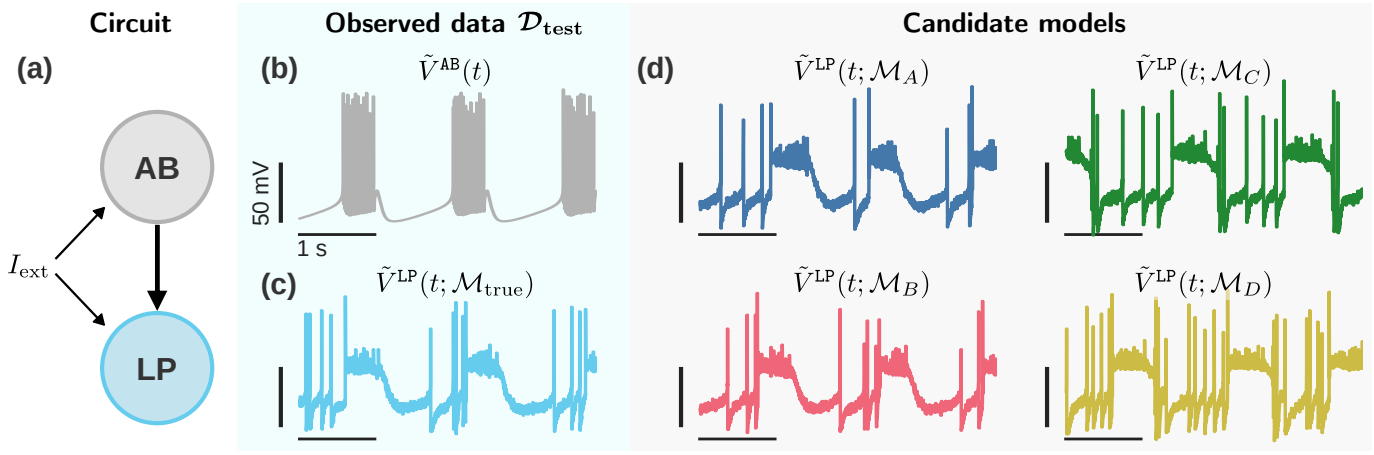


Figure 2: **Comparison of LP neuron responses** The goal is to estimate a good set of candidate models for neuron LP. **a)** We consider a simple circuit with a known pacemaker neuron (AB) and a post-synaptic neuron of unknown response (LP). **b)** Output of the AB neuron. This serves as input to the LP neuron. In practice, these data would more likely be provided by an experimental recording of neuron AB, but here we assume it is a simulatable known model for convenience. **c)** Response of neuron model `LP 1` to the input in (b). Together, (b) and (c) serve as our observed data $\mathcal{D}_{\text{test}}$. **d)** Response of neuron models `LP 2` to `LP 5` to the input in (b). These are our four candidate models of neuron LP.

noise input I_{ext}, representing an external current received from other neurons. (External currents may also be produced by the experimenter, to help model the underlying dynamics.) Second, we think of the system as the combination of a biophysical model — described by equations (50) to (52) — and an observation model which adds Gaussian noise. The observation model represents components which don't affect the biophysics — like noise in the recording equipment — and can be modelled to a first approximation as:

$$\tilde{V}^{\text{LP}}(t_k; \mathcal{M}_{\text{true}}) = \underbrace{V^{\text{LP}}\big(t_k, I_{\text{ext}}(t_k; \tau, \sigma_i); \mathcal{M}_{\text{true}}\big)}_{\text{biophysical model}} + \underbrace{\xi(t_k; \sigma_o)}_{\text{observation model}} . \tag{8}$$

The parameters of the external input I_{ext} are τ and σ_i, which respectively determine its autocorrelation time and strength (i.e. its amplitude). The observation model has only one parameter, σ_o, which determines the strength of the noise. The symbol $\mathcal{M}_{\text{true}}$ is shorthand for everything defining the data-generating process, which includes the parameters τ, σ_i and σ_o. The indices i and o are used to distinguish *input* and *output* model parameters.

Since the candidate models in this example assume additive observational Gaussian noise, a natural choice for the loss function is the negative log likelihood:

$$\begin{aligned} Q(t_k, \tilde{V}^{\text{LP}}; \mathcal{M}_a) &= -\log p(\tilde{V}^{\text{LP}} \mid t_k, \mathcal{M}_a) \\ &= \tfrac{1}{2}\log(2\pi\sigma_o) - \frac{(\xi_k)^2}{2\sigma} , \\ &= \tfrac{1}{2}\log(2\pi\sigma_o) - \frac{\big(\tilde{V}^{\text{LP}} - V^{\text{LP}}(t_k; \mathcal{M}_a)\big)^2}{2\sigma_o} . \end{aligned} \tag{9}$$

However one should keep in mind that a) it is not necessary to define the loss in terms of the log likelihood, and b) the loss used to *compare* models need not be the same used to *fit* the models. For example, if the log likelihood of the assumed observation model is non-convex or non-differentiable, one might use a simpler objective for optimization. Alternatively, one might fit using the log likelihood, but compare models based on global metrics like the interspike interval. Notably, Prinz *et al.*[48] do not directly fit to potential traces $\tilde{V}$, but rather use a set of data-specific heuristics and global statistics to select candidate models. Gonçalves *et al.*[24] similarly find that working with specially crafted summary statistics — rather than the likelihoods — is more practical for inferring this type of model. In this work we nevertheless stick with the generic form of equation (9), which makes no assumptions on the type of data used to fit the models and is thus easier to generalise to different scenarios.

The definition of the risk then follows immediately:

$$R_a \coloneqq \mathbb{E}_{(t,\tilde{V}^{\text{LP}})\sim\mathcal{M}_{\text{true}}}\big[Q(t, \tilde{V}^{\text{LP}}; \mathcal{M}_a)\big] . \tag{10}$$

Here a stands for one of the model labels A, B, C or D. The notation $(t, \tilde{V}^{\text{LP}}) \sim \mathcal{M}_{\text{true}}$ denotes that the distribution from which the samples $(t, \tilde{V}^{\text{LP}})$ are drawn is $\mathcal{M}_{\text{true}}$. We can think of $\mathcal{M}_{\text{true}}$ as producing an infinite sequence of data points by integrating the model dynamics and appyling equation (8) multiple times, or by going to the laboratory and performing the experiment multiple times.

As alluded to above, the candidate model traces in Fig. 2**d** suggest two groups of models: the $\mathcal{M}_A$ and $\mathcal{M}_B$ models seem to better reproduce the data than $\mathcal{M}_C$ or $\mathcal{M}_D$. Within each group however, it is hard to say whether one model is better than the other; in terms of the risks, this means that we expect $R_A, R_B < R_C, R_D$. This also means that we should be wary of a selection criterion which unequivocally ranks $\mathcal{M}_A$ better than $\mathcal{M}_B$, or $\mathcal{M}_C$ better than $\mathcal{M}_D$.

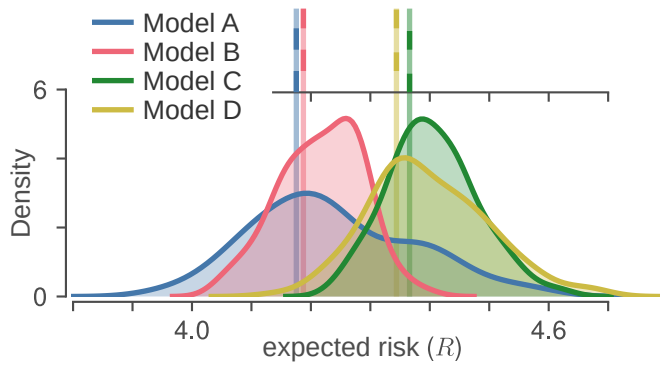


Figure 3: **Empirical risk vs. R-distributions for the four candidate LP models. (top) The empirical risk** (5) for each of the four candidate LP models. **(bottom) Our proposed B^{EMD} criterion** replaces the risk by an *R-distribution*, where the spread of each distribution is due to the *replication uncertainty* for that particular model. R-distributions are distributions of the R functional in equation (20); we estimate them by sampling the quantile function (i.e. inverse cumulative density function) q according to a stochastic process $\mathfrak{Q}$ on quantile functions. We used an EMD sensitivity factor of $c = 2^{-2}$ (see later section on calibration) for $\mathfrak{Q}$ and drew samples $\hat{q} \sim \mathfrak{Q}$ until the relative standard error on the risk was below 3 %. A kernel density estimate (KDE) is used to display those samples as distributions. The R-distribution for the true model is much narrower (approximately Dirac) and far to the left, outside the plotting bounds.

In other words, we expect that uncertainties on R_A and R_B should be at least commensurate with the difference $|R_A - R_B|$, and similarly for the uncertainties on R_C and R_D.

2.3 A conceptual model for replicate variations

Evaluating the risk (10) for each candidate model on a dataset $\mathcal{D}_{\text{test}}$ yields four scalars $\hat{R}_A$ to $\hat{R}_D$; since a lower risk should indicate a better model, a simple naive model selection rule would be

$$\begin{cases} \text{reject } \mathcal{M}_b & \text{if } \hat{R}_a < \hat{R}_b\,, \\ \text{reject } \mathcal{M}_a & \text{if } \hat{R}_a > \hat{R}_b\,, \\ \text{no rejection} & \text{if } \hat{R}_a = \hat{R}_b\,, \end{cases} \quad \text{for } a, b \in \{A, B, C, D\}\,. \tag{11}$$

As with many selection criteria (e.g. AIC,[2,1] BIC,[57] DIC[60,61]), this rule is effectively binary: the third option, to reject neither model, has probability zero. It therefore *always* selects either $\mathcal{M}_a$ or $\mathcal{M}_b$, even when the evidence favouring one of the two is extremely weak. Another way to see this is illustrated at the top of Fig. 3: the lines representing the four values R_A through R_D have no error bars, so even minute differences suffice to rank the models.

The problem is that from a scientific standpoint, if the evidence is too weak, this ranking is likely irrelevant — or worse, misinformative. Hence our desire to assign uncertainty to each estimate of the risk. Ideally we would like to be able to compute tail probabilities like $P(R_A < R_B)$, which would quantify the strength of the evidence in favour of either $\mathcal{M}_A$ or $\mathcal{M}_B$. Selecting a minimum evidence threshold ϵ would then allow us to convert equation (11) into a true ternary decision rule with non-zero probability of keeping both models:

$$\begin{cases} \text{reject } \mathcal{M}_b & \text{if } P(R_a < R_b) > \epsilon\,, \\ \text{reject } \mathcal{M}_a & \text{if } P(R_a < R_b) < (1-\epsilon)\,, \\ \text{no rejection} & \text{if } (1-\epsilon) \leq P(R_a < R_b) \leq \epsilon\,. \end{cases} \tag{12}$$

As we explain in the Introduction, our goal is to select a model which can describe not just the observed data $\mathcal{D}_{\text{test}}$, but also new data generated in a replication experiment. In order for our criterion to be *robust*, the tail probabilities $P(R_a < R_b)$ should account for uncertainty in the replication process.

Note that for a given candidate model $\mathcal{M}_a$ and fixed $\mathcal{M}_{\text{true}}$, we can always estimate the risk with high accuracy if we have enough samples, irrespective of the amount of noise intrinsic to $\mathcal{M}_a$ or of misspecification between $\mathcal{M}_a$ and $\mathcal{M}_{\text{true}}$. Crucially however, we *also do not assume the replication process to be stationary*, so a replicate dataset $\mathcal{D}'_{\text{test}}$ may be drawn from a slightly different data-generating process $\mathcal{M}'_{\text{true}}$. This can occur even when $\mathcal{M}_{\text{true}}$ itself is stationary, reflecting the fact that it is often easier to control variability within a single experiment than across multiple ones. Ontologically therefore, the uncertainty on $\hat{R}$ is a form of *epistemic uncertainty* arising from the variability across (experimental) replications.

To make this idea concrete, consider that the input I_{ext} and noise ξ might not be stationary over the course of an experiment with multiple trials. We can represent this by making their parameters random variables, for example

$$\Omega \coloneqq \begin{cases} \log \sigma_o \sim \mathcal{N}(0.0\,\text{mV}, (0.5\,\text{mV})^2) \\ \log \sigma_i \sim \mathcal{N}(-15.0\,\text{mV}, (0.5\,\text{mV})^2) \\ \log_{10} \tau \sim \text{Unif}([0.1\,\text{ms}, 0.2\,\text{ms}]) \end{cases}, \tag{13}$$

and drawing new values of $(\sigma_o, \sigma_i, \tau)$ for each trial (i.e. each replicate). Since it is a distribution over data-generating processes, we call Ω an *epistemic distribution*. For illustration purposes, here we have parametrised Ω in terms of two parameters of the biophysical model and one parameter of the observation model, thus capturing epistemic uncertainty within a single experiment. In general the parametrisation of Ω is a modelling choice, and may represent other forms of non-stationarity — for example due to variations between experimental setups in different laboratories.

Conceptually, we could estimate the tail probabilities $P(R_a < R_b)$ by sampling J different data-generating processes $\mathcal{M}^j_{\text{true}}$ from Ω, for each then drawing a dataset $\mathcal{D}^j_{\text{test}} \sim \mathcal{M}^j_{\text{true}}$, and finally computing the empirical risks $\hat{R}^j_a$ and $\hat{R}^j_b$ on $\mathcal{D}^j_{\text{test}}$. The fraction of datasets for which $\hat{R}^j_a < \hat{R}^j_b$ would then estimate the tail probability:

$$P(R_a < R_b) \approx \frac{1}{J} \left| \left\{ j \;\middle|\; \hat{R}^j_a < \hat{R}^j_b \right\}_{j=1}^{J} \right|. \tag{14}$$

The issue of course is that we cannot know Ω. First because we only observe data from a single $\mathcal{M}_{\text{true}}$, but also because there may be different contexts to which we want to generalise: one researcher may be interested in modelling an

individual LP neuron, while another might seek a more general model which can describe all neurons of this type. These two situations would require different epistemic distributions, with the latter one being in some sense broader.

However *we need not commit to a single epistemic distribution*: if we have two distributions, and we want to ensure that conclusions hold under both, we can instead use the condition

$$\min_{\Omega\in\{\Omega_1,\Omega_2\}} P_\Omega(R_a < R_b) > \epsilon \tag{15}$$

to attempt to reject $\mathcal{M}_b$. In general, considering more epistemic distributions will make the model selection more robust, at the cost of discriminatory power. *Epistemic distributions are not therefore prior distributions*, since a Bayesian calculation is always tied to a specific choice of prior. (The opposite however holds: a prior can be viewed as a particular choice of epistemic distribution.)

We view epistemic distributions mostly as conceptual tools. For calculations, we will instead propose in the next sections a different type of distribution ($\mathfrak{Q}$; technically a stochastic process), which is not on the data-generating process, but on the distribution of pointwise losses. Being lower-dimensional and more stereotyped than Ω, we will be able to construct $\mathfrak{Q}$ entirely non-parametrically, up to a scaling constant c (c.f. Fig. 1 and the section listing desiderata for $\mathfrak{Q}$). A later section will then show, through numerical calibration and validation experiments, that $\mathfrak{Q}$ also has nice universal properties, so that the only thing that really matters is the overall scale of the epistemic distributions. The constant c is matched to this scale by numerically simulating epistemic distributions as part of the calibration experiments.

2.4 Model discrepancy as a baseline for non-stationary replications

To keep the notation in the following sections more general, we use the generic x and y as independent and dependent variables. To recover expressions for our neuron example, substitute $x \to t$, $y \to \tilde{V}^{\mathrm{LP}}$, $\mathcal{X} \to \mathcal{T}$ and $\mathcal{Y} \to \mathcal{V}$. Where possible we also use A and B as generic placeholders for a model label.

Our goal is to define a selection criterion which is robust against variations between experimental replications, but which can be computed *using knowledge only of the candidate models and the observed empirical data.* To do this, we make the following assumption:

EMD assumption (version 1). *Candidate models represent that part of the experiment which we understand and control across replications.*

More precisely, in the next section we define the *empirical model discrepancy* function $\delta_A^{\mathrm{EMD}} : (0,1) \to \mathbb{R}_{\geq 0}$ such that if model $\mathcal{M}_A$ exactly reproduces the observations, then δ_A^{EMD} is identically zero. This function therefore measures the discrepancy between model predictions and actual observations. Since we expect misspecified models to have positive discrepancy ($\int_0^1 \delta_A^{\mathrm{EMD}}(\Phi)d\Phi > 0$), in the following we treat the discrepancy δ_A^{EMD} as a measure of *misspecification.*

Under our EMD assumption, misspecification in a model corresponds to experimental conditions we don't fully control, and which could therefore vary across replications of the experiment. Concretely this means that we can reformulate the EMD assumption as

EMD assumption (version 2). *The variability of R_A across replications is predicted by the model discrepancy δ_A^{EMD}.*

We also assume that the data-generating process $\mathcal{M}_{\mathrm{true}}$ within one replication is *strictly stationary* with finite correlations, i.e. that all observations are identically distributed but may be correlated. For simplicity in fact we treat the samples as i.i.d., since if necessary this could be done by thinning. See references 22 or 38 for discussions on constructing estimators from correlated time series.

Below we further assume a particular *linear* relationship between δ_A^{EMD} and the replication uncertainty: the function δ_A^{EMD} is scaled by the aforementioned sensitivity factor c to determine the variance of the stochastic process $\mathfrak{Q}_A$ (which we recall induces the R_A-distribution for the risk of model $\mathcal{M}_A$). The parameter $c \in \mathbb{R}_+$ therefore represents a conversion from *model discrepancy* to *epistemic uncertainty*. A practitioner can use this parameter to adjust the sensitivity of the criterion to misspecification: a larger value of c will emulate more important experimental variations. Since the tail probabilities (equation (14)) we want to compute will depend on c, we will write them

$$B_{AB;c}^{\mathrm{EMD}} \coloneqq P(R_A < R_B \mid c)\,. \tag{16}$$

The EMD rejection rule. *For a chosen rejection threshold $\epsilon \in (0.5, 1]$, reject model $\mathcal{M}_A$ if there exists a model $\mathcal{M}_B$ such that $B_{AB;c}^{\mathrm{EMD}} < \epsilon$ and $\hat{R}_A > \hat{R}_B$.*

The second condition ($\hat{R}_A > \hat{R}_B$) ensures the rejection rule remains consistent, even if the R-distributions become skewed. (See comment below equation (5).)

As an illustration, Table 1 gives the value of B^{EMD} for each candidate model pair in our example from Figs. 2 and 3. As expected, models that were visually assessed to be similar also have B^{EMD} values close to $\frac{1}{2}$. In practice one would not necessarily need to compute the entire table, since the B^{EMD} satisfy *dice transitivity*.[12,4] In particular this implies that for any threshold $\epsilon > \varphi^{-2}$ (where φ is the golden ratio), we have

$$\left.\begin{aligned} B_{AB;c}^{\mathrm{EMD}} &> \sqrt{\epsilon} \\ B_{BC;c}^{\mathrm{EMD}} &> \sqrt{\epsilon} \end{aligned}\right\} \Rightarrow B_{AC;c}^{\mathrm{EMD}} > \epsilon\,. \tag{17}$$

Since any reasonable choice of threshold will have $\epsilon > 0.5 > \varphi^{-2}$, we can say that whenever the comparisons $B_{AB;c}^{\mathrm{EMD}}$ and $B_{BC;c}^{\mathrm{EMD}}$ are both greater than $\sqrt{\epsilon}$, we can treat them as transitive. We give a more general form of this result in the Supplementary Methods.

2.5 δ^{EMD}: Expressing misspecification as a discrepancy between CDFs

We can treat the loss function for a given model $\mathcal{M}_A$ as a random variable $Q(x, y; \mathcal{M}_A)$ where the (x, y) are sampled

Table 1: B^{EMD} **probabilities for candidate LP models.** Values are the probabilities given by equation (16) with the candidate labels a and b corresponding to rows and columns respectively. Candidate models being compared are those of Fig. 2. Probabilities are computed for the R-distributions shown in Fig. 3, which used an EMD constant of $c = 2^{-2}$. Since $P(R_a < R_b) = 1 - P(R_b < R_a)$, the sum of symmetric entries equals 1.

	A	B	C	D
A	0.500	0.483	0.846	0.821
B	0.517	0.500	0.972	0.940
C	0.154	0.028	0.500	0.463
D	0.179	0.060	0.537	0.500

from $\mathcal{M}_{\text{true}}$. A key realisation for our approach is that *the CDF (cumulative distribution function) of the loss suffices to compute the risk.* Indeed, we have for the CDF of the loss

$$\begin{aligned}\Phi_A^*(q) &\coloneqq p\big(Q(x,y;\mathcal{M}_A) \le q \mid x,y \sim \mathcal{M}_{\text{true}}\big) \\ &= \int_{\mathcal{X}\times\mathcal{Y}} H\big(q - Q(x,y;\mathcal{M}_A)\big)\, p\big(x,y \mid \mathcal{M}_{\text{true}}\big)\, dxdy \\ &\approx \frac{1}{L} \sum_{x_i,y_i \in \mathcal{D}_{\text{test}}} H\big(q - Q(x_i,y_i;\mathcal{M}_A)\big)\,, \end{aligned} \tag{18}$$

where H is the Heaviside function with $H(q) = 1$ if $q \ge 0$ and 0 otherwise.

Since $\mathcal{M}_{\text{true}}$ is unknown, a crucial feature of equation (18) is that $\Phi_A^*(q)$ can be estimated without needing to evaluate $p\big(x,y \mid \mathcal{M}_{\text{true}}\big)$; indeed, all that is required is to count the number of observed data points in $\mathcal{D}_{\text{test}}$ which according to $\mathcal{M}_A$ have a loss less than q. Moreover, since Q returns a scalar, the data points (x_i, y_i) can have any number of dimensions: the number of data points required to get a good estimate of $\Phi_A^*(q)$ does not depend on the dimensionality of the data, for the same reason that estimating marginals of a distribution requires much fewer samples than estimating the full distribution. Finally, because the loss is evaluated using $\mathcal{M}_A$, but the expectation is taken with respect to a distribution determined by $\mathcal{M}_{\text{true}}$, we call Φ_A^* the *mixed CDF*.

We can invert Φ_A^* to obtain the *mixed PPF* (*percent point function*, also known as *quantile function*, or *percentile function*):

$$q_A^*(\Phi) \coloneqq {\Phi_A^*}^{-1}(\Phi)\,, \tag{19, \textit{mixed PPF}}$$

which is also a 1-d function, irrespective of the dimensionality of $\mathcal{X}$ or $\mathcal{Y}$. We can then rewrite the risk as a one dimensional integral in q_A^*:

$$\begin{aligned} R_A = R[q_A^*] &= \int_0^1 q_A^*(\Phi)\, d\Phi \\ &\approx \frac{1}{L} \sum_{x_i,y_i \in \mathcal{D}_{\text{test}}} Q(x_i,y_i;\mathcal{M}_A)\,. \end{aligned} \tag{20}$$

To obtain equation (20), we simply used Fubini's theorem to reorder the integral of equation (18) and marginalized over all slices of a given loss $q_A^*(\Phi)$. The integral form (first line of equation (20)) is equivalent to averaging an infinite number of samples, and therefore to the *(true) risk*, whereas the average over observed samples (second line of equation (20)) is exactly the *empirical risk* defined in equation (5).

In practice, to evaluate equation (20), we use the observed samples to compute the sequence of per-sample losses $\{Q(x_i,y_i;\mathcal{M}_A)\}_{x_i,y_i\in\mathcal{D}_{\text{test}}}$. This provides us with a sequence of losses, which we use as ordinate values. We then sort this sequence so that we have $\{Q_i\}_{i=1}^L$ with $Q_i \le Q_i + 1$, and assign to each the abscissa $\Phi_i = i/L + 1$, such that losses are motonically increasing and uniformly distributed on the [0, 1] interval. This yields the *empirical PPF of the loss* — the "empirical" qualifier referring to this construction via samples, as opposed to an analytic calculation. Interpolating the points then yields a continuous function which can be used in further calculations. All examples in this paper linearly interpolate the PPF from $2^{10} = 1024$ points.

In Fig. 4 we show four examples of empirical PPFs, along with their associated empirical CDFs. We see that the statistics of the additive observational noise affects the shape of the PPF: for noise with exponential tails, as we get from Gaussian or Poisson distributions, we have strong concentration around the minimum value of the loss followed by a sharp increase at $\Phi = 1$. For heavier-tailed distributions like Cauchy, loss values are less concentrated and the PPF assigns non-negligible probability mass to a wider range of values. The dimensionality of the data also matters. High-dimensional Gaussians are known to place most of their probability mass in a thin shell centered on the mode, and we see this in the fourth column of Fig. 4: the sharp increase at $\Phi = 0$ indicates that very low probability is assigned to the minimum loss.

Since by construction, the abscissae Φ of an empirical PPF are spaced at intervals of $1/L$, the Riemann sum for the integral in equation (20) reduces to the sample average. More importantly, we can interpret the risk as a functional in $q_A^*(\Phi)$, which will allow us below to define a generic stochastic process that accounts for epistemic uncertainty.

Up to this point with equation (20) we have simply rewritten the usual definition of the risk. Recall now that in the previous section, we proposed to equate replication uncertainty with misspecification; specifically we are interested in how differences between the candidate model $\mathcal{M}_A$ and the true data-generating process $\mathcal{M}_{\text{true}}$ affect the loss PPF, since this determines the risk. Therefore we also compute the PPF of $Q(x,y;\mathcal{M}_A)$ under its own model (recall from equation (1) that $\mathcal{M}_A$ must be a probabilistic model):

$$\begin{aligned}\widetilde{\Phi}_A(q) &\coloneqq p\big(Q(x,y;\mathcal{M}_A) \le q \mid x,y \sim \mathcal{M}_A\big) \\ &= \int_{\mathcal{X}\times\mathcal{Y}} dxdy\; H\big(q - Q(x,y;\mathcal{M}_A)\big)\, p\big(x,y \mid \mathcal{M}_A\big) \\ &\approx \frac{1}{L_{\text{synth},A}} \sum_{x_i,y_i \in \mathcal{D}_{\text{synth},A}} H\big(q - Q(x_i,y_i;\mathcal{M}_A)\big)\,, \end{aligned} \tag{21}$$

from which we obtain the PPF:

$$\tilde{q}_A(\Phi) \coloneqq \widetilde{\Phi}_A^{-1}(\Phi)\,. \tag{22, \textit{synth PPF}}$$

The only difference between $\tilde{q}_A$ and q_A^* is the use of $p(x,y \mid \mathcal{M}_A)$ instead of $p(x,y \mid \mathcal{M}_{\text{true}})$ in the integral. In practice

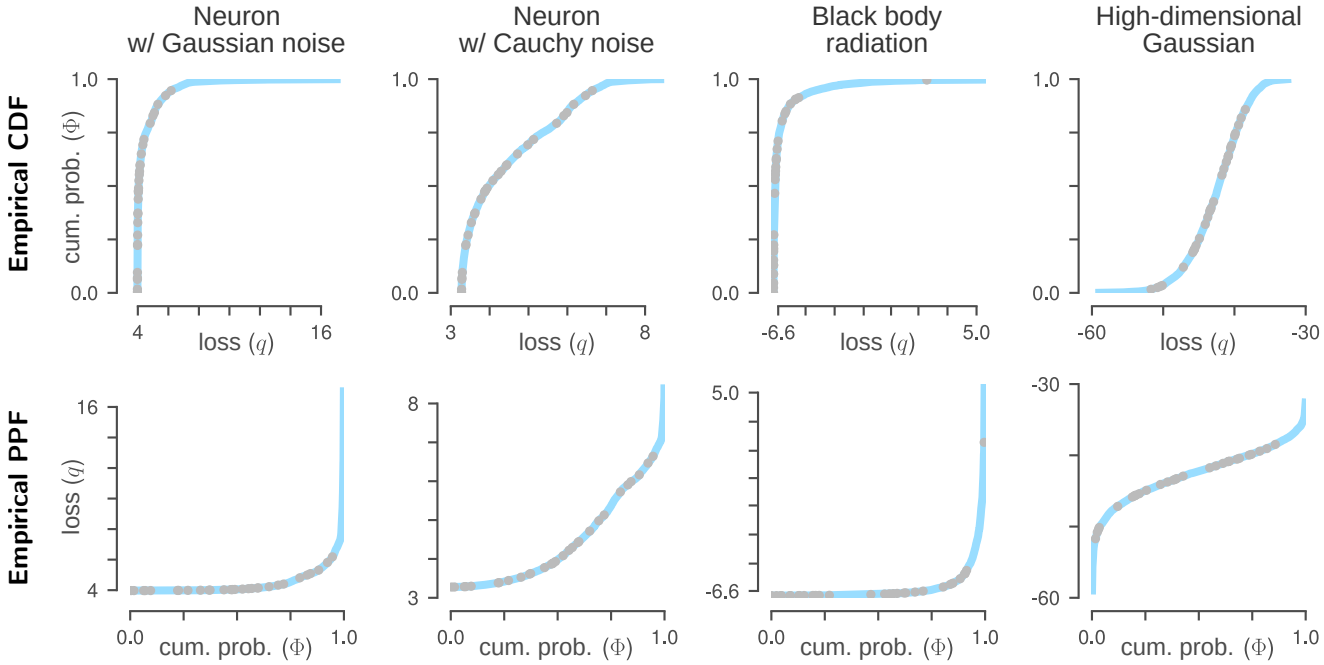


Figure 4: **Loss PPF for different models:** each column corresponds to a different model. The PPF (bottom row) is the inverse of the CDF (top row). For calculations we interpolate $2^{10} = 1024$ points (cyan line) to obtain a smooth function; for illustration purposes here only 30 points are shown. The data for the first two columns were generated with the neuron model described at the top of our Results, where the additive noise follows either a Gaussian or Cauchy distribution. The black body radiation data for the third column were generated from a Poisson distribution using equation (38) with $s = 2^{14}$ and λ in the range 6 µm to 20 µm. Here the true noise is binomial, but the loss assumes a Gaussian. The fourth column shows an example where the data are high-dimensional; the same 30 dimensional, unit variance, isotropic Gaussian is used for both generating the data and evaluating the loss. In all panels the loss function used is the log likelihood under the model.

this integral would also be evaluated by sampling, using $\mathcal{M}_A$ to generate a dataset $\mathcal{D}_{\text{synth},A}$ with $L_{\text{synth},A}$ samples. Because in this case the candidate model is used for both generating samples and defining the loss, we call $\tilde{q}_A$ ($\widetilde{\Phi}_A$) the *synthetic PPF* (*CDF*).

The idea is that the closer $\mathcal{M}_A$ is to $\mathcal{M}_{\text{true}}$, the closer also the synthetic PPF should be to the mixed PPF — indeed, equality of the PPFs ($\tilde{q}_a = q^*_A$) is a necessary condition for equality of the models ($\mathcal{M}_A = \mathcal{M}_{\text{true}}$). Therefore we can quantify the uncertainty due to misspecification, at least insofar as it affects the risk, as the absolute difference between $\tilde{q}_A$ and q^*_A:

$$\begin{aligned} \delta^{\text{EMD}}_A : [0,1] &\to \mathbb{R} \\ \Phi &\mapsto \left| \tilde{q}_A(\Phi) - q^*_A(\Phi) \right|. \end{aligned} \tag{23}$$

We refer to δ^{EMD}_A as the *empirical model discrepancy* (EMD) function because it measures the discrepancy between two empirical PPFs.

It is worth noting that even a highly stochastic data-generating process $\mathcal{M}_{\text{true}}$, with a lot of aleatoric uncertainty, still has a well defined PPF — which would be matched by an equally stochastic candidate model $\mathcal{M}_A$. Therefore the discrepancies measured by δ^{EMD}_A are a representation of the *epistemic* uncertainty. These discrepancies can arise either from a mismatch between $\mathcal{M}_A$ and $\mathcal{M}_{\text{true}}$, or simply having too few samples to estimate the mixed PPF q^*_A exactly; either mechanism contributes to the uncertainty on the expected risk of replicate datasets. We illustrate some differences between aleatoric and epistemic uncertainty in Supplementary Fig. 3, and include further comments on how different types of uncertainties relate to risk in the Supplementary Discussion.

2.6 Ω: A stochastic process on quantile functions

When replication is non-stationary, each replicate dataset will produce a different loss PPF. A distribution over replications (like the previously defined epistemic distribution Ω) therefore corresponds to a distribution over PPFs. This leads us to the final formulation of our Emdassumptionversion2 2.4, which we now operationalize by assuming a linear relation between empirical model discrepancy and epistemic uncertainty:

EMD principle. *For a given model $\mathcal{M}_A$, differences between its PPFs on two replicate datasets should be proportional to δ^{EMD}_A.*

Formalising this idea is made considerably simpler by the fact that PPFs are always scalar functions, irrespective of the model or dataset's dimensionality. We do this as follows, going through the steps schematically represented by downward facing arrows on the right of Fig. 1:

> For a given model $\mathcal{M}_A$, we treat the PPFs of different replicates as realisations of a stochastic process $\mathfrak{Q}_A$ on the interval $[0,1]$; a realisation of $\mathfrak{Q}_A$ (i.e. a PPF) is a function $\hat{q}_A(\Phi)$ with $\Phi \in [0,1]$. The stochastic process $\mathfrak{Q}_A$ is defined to satisfy the desiderata listed below, which ensure that it is centered on the observed PPF q_A^* and that its variance at each Φ is governed by δ_A^{EMD}.

Obtaining the R-distributions shown in Fig. 3 is then relatively straightforward: we simply sample an ensemble of $\hat{q}_A$ from $\mathfrak{Q}_A$ and integrate each to obtain an ensemble of risks.

Working with PPFs has the important advantage that we can express our fundamental assumption as a simple linear proportionality relation between empirical model discrepancy and epistemic uncertainty. While this is certainly an approximation, it simplifies the interpretability of the resulting R-distribution. As we show in the Supplementary Discussion, the resulting criterion is also more robust than other alternatives. The simplicity of the assumption however belies the complexity of defining a stochastic process $\mathfrak{Q}$ directly on PPFs.

2.6.1 *Desiderata for* $\mathfrak{Q}$

In order to interpret them as such, realisations of $\hat{q}_A \sim \mathfrak{Q}_A$ must be valid PPFs. This places quite strong constraints on those realisations, for example:

- All realisations $\hat{q}_A(\Phi) \sim \mathfrak{Q}_A$ must be *monotone*.
- All realisations $\hat{q}_A(\Phi) \sim \mathfrak{Q}_A$ must be *integrable*.

Monotonicity follows immediately from definitions: a CDF is always monotone because it is the integral of a positive function (equation (18)), and therefore its inverse must also be monotone.

Integrability simply means that the integral in equation (20) exists and is finite. Concretely this is enforced by ensuring that the process $\mathfrak{Q}_A$ is *self-consistent*,[20] a property which we explain in the Methods.

Interpreting the realisations $\hat{q}$ as PPFs also imposes a third constraint, more subtle but equally important:

- The process $\mathfrak{Q}_A$ must be *non-accumulating*.

A process which is accumulating would start at one end of the domain, say $\Phi = 0$, and sequentially accumulate increments until it reaches the other end. Brownian motion over the interval $[0, T]$ is an example of such a process. In contrast, consider the process of constructing a PPF for the data in Fig. 2: initially we have few data points and the PPF of their loss is very coarse. As the number of points increases, the PPF gets refined, but since loss values occur in no particular order, this happens *simultaneously across the entire interval*.

The accumulation of increments strongly influences the statistics of a process; most notably, the variance is usually larger further along the domain. This would not make sense for a PPF: if $\mathbb{V}[\mathfrak{Q}_A(\Phi)]$ is smaller than $\mathbb{V}[\mathfrak{Q}_A(\Phi')]$, that should be a consequence of $\delta^{\mathrm{EMD}}(\Phi)$ being smaller than $\delta^{\mathrm{EMD}}(\Phi')$ — not of Φ occurring "before" Φ'.

This idea that a realisation $\hat{q}_A(\Phi)$ is generated simultaneously across the interval led us to define $\mathfrak{Q}_A$ as a sequence of *refinements*: starting from an initial increment $\Delta\hat{q}_1(0) = \hat{q}_A(1) - \hat{q}_A(0)$ for the entire Φ interval $[0,1]$, we partition $[0,1]$ into n subintervals, and sample a set of n subincrements in such a way that they sum to $\Delta\hat{q}_1(0)$. This type of distribution, where n random variables are drawn under the constraint of a fixed sum, is called a *compositional* distribution.[42] Note that the constraint reduces the number of dimensions by one, so a pair of increments would be drawn from a 1-d compositional distribution. A typical 1-d example is the beta distribution for $x_1 \in [0,1]$, with $x_2 = (1 - x_1)$ and $\alpha, \beta > 0$:

$$\text{if } \; x_1 \sim \mathrm{Beta}(\alpha, \beta)\,, \quad \text{then } \; p(x_1) \propto x_1^{\alpha-1}(1-x_1)^{\beta-1}\,. \tag{24}$$

Interestingly, the most natural statistics for compositional distributions are not the mean and variance, but analogue notions of *centre* and *metric variance*;[42,47] for the beta distribution defined above, these are

$$\mathbb{E}_a[(x_1, x_2)] = \frac{1}{e^{\psi(\alpha)} + e^{\psi(\beta)}}\left(e^{\psi(\alpha)}, e^{\psi(\beta)}\right) \tag{25a}$$

$$\mathrm{Mvar}[(x_1, x_2)] = \frac{1}{2}\left(\psi_1(\alpha) + \psi_1(\beta)\right), \tag{25b}$$

where ψ and ψ_1 are the digamma and trigamma functions respectively, and $\mathbb{E}_a$ denotes expectation with respect to the Aitchison measure.[41,42] In essence, equations (25a) and (25b) are obtained by mapping x_1 and x_2 to the unbounded domain $\mathbb{R}$ via a logistic transformation, then evaluating moments of the unbounded variables. Of particular relevance is that — in contrast to the variance — the metric variance Mvar of a compositional distribution is therefore unbounded, which simplifies the selection of α and β (see Choosing beta distribution parameters in the Methods).

Of course, we not only want the $\hat{q}_A$ to be valid PPFs, but also descriptive of the model and data. We express this as two additional constraints, which together define a notion of closeness to q_A^*:

- At each intermediate point $\Phi \in (0,1)$, the *centre* of the process is given by $q_A^*(\Phi)$:

$$\mathbb{E}_a\left[\hat{q}(\Phi)\right] = q_A^*(\Phi)\,. \tag{26}$$

- At each intermediate point $\Phi \in (0,1)$, the *metric variance* is proportional to the square of $\delta_A^{\mathrm{EMD}}(\Phi)$:

$$\mathrm{Mvar}\left[\hat{q}(\Phi)\right] = c\,\delta_A^{\mathrm{EMD}}(\Phi)^2\,. \tag{27}$$

Note that equation (27) formalises the Emdprinciple 2.6 stated at the top of this section, with the sensitivity factor $c > 0$ determining the conversion from *empirical model discrepancy* (which we interpret as misspecification) to *epistemic uncertainty*. Note also that equations (26) and (27) are akin to a variational approximation of the stochastic process around q_A^*. Correspondingly, we should expect these equations (and therefore $\mathfrak{Q}$) to work best when c is not too large. We describe how one might determine an appropriate value for c in the later section on calibration.

In addition to the above desiderata, for reasons of convenience, we also ask that

- The end points are sampled from Gaussian distributions:

$$\begin{aligned}\hat{q}(0) &\sim \mathcal{N}(q_A^*(0), c\,\delta_A^{\mathrm{EMD}}(0)^2)\,,\\ \hat{q}(1) &\sim \mathcal{N}(q_A^*(1), c\,\delta_A^{\mathrm{EMD}}(1)^2)\,.\end{aligned} \tag{28}$$

Thus the process $\mathfrak{Q}_{A;c}$ is parametrised by two functions and a scalar: q_A^*, δ_A^{EMD} and c. It is molded to produce realisations $\hat{q}$ which as a whole track q_A^*, with more variability between realisations at points Φ where $\delta_A^{\mathrm{EMD}}(\Phi)$ is larger (middle panel of Fig. 5**a**).

To the best of our knowledge the current literature does not provide a process satisfying all of these constraints. To remedy this situation, we propose a new *hierarchical beta (HB) process*, which we illustrate in Fig. 5. A few example realisations of $\hat{q} \sim \mathfrak{Q}_A$ are drawn as grey lines in Fig. 5**a**. The mixed PPF q_A^* (equation (26)) is drawn as a green line, while the region corresponding to $q_A^*(\Phi) \pm \sqrt{c}\delta_A^{\mathrm{EMD}}(\Phi)$ is shaded in yellow. The process is well-defined for $c \in [0, \infty]$: the $c \to 0$ limit is simply q_A^*, while the $c \to \infty$ limit is a process which samples uniformly at every step, irrespective of q_A^* and δ_A^{EMD}. In other words, as c increases, the ability of each realisation $\hat{q}_A$ to track q_A^* is limited by the constraints of monotonicity and non-accumulating increments. This translates to step-like PPFs, as seen in the lower panel of Fig. 5**a**.

Figure 5**b** shows distributions of $\hat{q}(\Phi)$ at three different values of Φ. The value of $q_A^*(\Phi)$ is indicated by the green vertical bar and agrees well with $\mathbb{E}_a\big[\hat{q}(\Phi)\big]$; the desideratum of equation (26) is therefore satisfied. The scaling of these distributions with δ_A^{EMD} (equation (27)) is however only approximate, which we can see as the yellow shading not having the same width in each panel. This is a result of the tension with the other constraints: the HB process ensures that the monotonicity and integrability constraints are satisfied exactly, but allows deviations in the statistical constraints.

A realisation of an HB process is obtained by a sequence of refinements; we illustrate three such refinements steps in Fig. 5**d**. The basic idea is to refine an increment $\Delta\hat{q}_{\Delta\Phi}(\Phi) := \hat{q}(\Phi + \Delta\Phi) - \hat{q}(\Phi)$ into two subincrements $\Delta\hat{q}_{\frac{\Delta\Phi}{2}}(\Phi)$ and $\Delta\hat{q}_{\frac{\Delta\Phi}{2}}(\Phi + \frac{\Delta\Phi}{2})$, with $\Delta\hat{q}_{\frac{\Delta\Phi}{2}}(\Phi) + \Delta\hat{q}_{\frac{\Delta\Phi}{2}}(\Phi + \frac{\Delta\Phi}{2}) = \Delta\hat{q}_{\Delta\Phi}(\Phi)$. To do this, we first determine appropriate parameters α and β, draw x_1 from the corresponding beta distribution and assign

$$\begin{aligned}\Delta\hat{q}_{\frac{\Delta\Phi}{2}}(\Phi) &\leftarrow x_1 \Delta\hat{q}_{\Delta\Phi}(\Phi)\,,\\ \Delta\hat{q}_{\frac{\Delta\Phi}{2}}(\Phi + \tfrac{\Delta\Phi}{2}) &\leftarrow x_2 \Delta\hat{q}_{\Delta\Phi}(\Phi)\,,\end{aligned} \tag{29}$$

where again $x_2 = (1 - x_1)$. Figure 5**c** shows distributions for the first (orange) and second (blue) subincrements at the fourth refinement step, where we divide an increment over an interval of length $\Delta\Phi = 2^{-3}$ to two subincrements over intervals of length 2^{-4}. Each pair of subincrements is drawn for a different distribution, which depends on the particular realisation $\hat{q}$, but we can nevertheless see the PPF reflected in the aggregate distribution: the PPF has positive curvature, so the second subincrement tends to be larger than the first. Also both increments are bounded from below by 0, to ensure monotonicity.

A complete description of the HB process, including a procedure for choosing the beta parameters α and β such that our desiderata are satisfied, is given in the Methods.

2.6.2 *Using $\mathfrak{Q}$ to compare models*

Having constructed a process $\mathfrak{Q}_{A;c}$ for a candidate model $\mathcal{M}_A$, we can use it to induce a distribution on risks. We do this by generating a sequence of PPFs $\hat{q}_{A,1}, \hat{q}_{A,2}, \ldots, \hat{q}_{A,M_A}$, where $M_A \in \mathbb{N}$ and each $\hat{q}_{A,i}$ is drawn from $\mathfrak{Q}_{A;c}$ (see Fig. 5 for examples of sampled $\hat{q}_{A,i}$, and the Methods for more details on how we evaluate B^{EMD}). As we explain in the next section, the sensitivity parameter c is a property of the experiment; it is the same for all candidate models.

For each generated PPF, we evaluate the risk functional (using the integral form of equation (20)), thus obtaining a sequence of scalars $R\big[\hat{q}_{A,1}\big], R\big[\hat{q}_{A,2}\big], \ldots, R\big[\hat{q}_{A,M_A}\big]$ which follows $p(R_A \mid \mathfrak{Q}_A)$. With M_A sufficiently large, these samples accurately characterize the distribution $p(R_A \mid \mathfrak{Q}_A)$ (we use $\leftrightarrow$ to relate equivalent descriptions):

$$\begin{aligned}R_A \sim p(R_A \mid \mathcal{D}_{\mathrm{test}}, \mathcal{M}_A; c)\\ &\leftrightarrow \left\{R\big[\hat{q}_{A,1}\big], R\big[\hat{q}_{A,2}\big], \ldots, R\big[\hat{q}_{A,M_A}\big] \;\middle|\; \hat{q}_A \sim \mathfrak{Q}_A\right\}\\ &\leftrightarrow \left\{R_{A,1}\,,\quad R_{A,2}\,,\quad \ldots, R_{A,M_A}\right\}.\end{aligned} \tag{30}$$

Repeating this procedure for a different model $\mathcal{M}_B$ yields a different distribution for the risk:

$$R_B \sim p(R_B \mid \mathcal{D}_{\mathrm{test}}, \mathcal{M}_B; c) \;\leftrightarrow\; \left\{R_{B,1}\,, R_{B,2}\,, \ldots, R_{B,M_B}\right\}. \tag{31}$$

The $B_{AB;c}^{\mathrm{EMD}}$ criterion (16) then reduces to a double sum:

$$\begin{aligned}B_{AB;c}^{\mathrm{EMD}} &:= P(R_A < R_B \mid c)\\ &\approx \frac{1}{M_A M_B} \sum_{i=1}^{M_A} \sum_{j=1}^{M_B} \mathbf{1}_{R_{A,i} < R_{B,j}}\,.\end{aligned} \tag{32}$$

In equation (32), the term within the sum is one when $R_{A,i} < R_{B,j}$ and zero otherwise. A value of $B_{AB;c}^{\mathrm{EMD}}$ greater (lesser) than 0.5 indicates evidence for (against) model $\mathcal{M}_A$.

The undetermined parameter c (which converts discrepancy into epistemic uncertainty; c.f. equation (27)) can be viewed as a way to adjust the sensitivity of the criterion: larger values of c will typically lead to broader distributions of R_A, and therefore lead to a more conservative criterion (i.e. one which is more likely to result in an equivocal outcome). We give some guidelines on choosing c in the next section.

2.7 Calibrating and validating the B^{EMD}

The process $\mathfrak{Q}$ constructed in the previous section succeeds in producing distributions of the risk. And at least with some values of the scaling constant c, the distributions look as we expect (Fig. 3) and define a B^{EMD} criterion which assigns reasonable tail probabilities (Table 1).

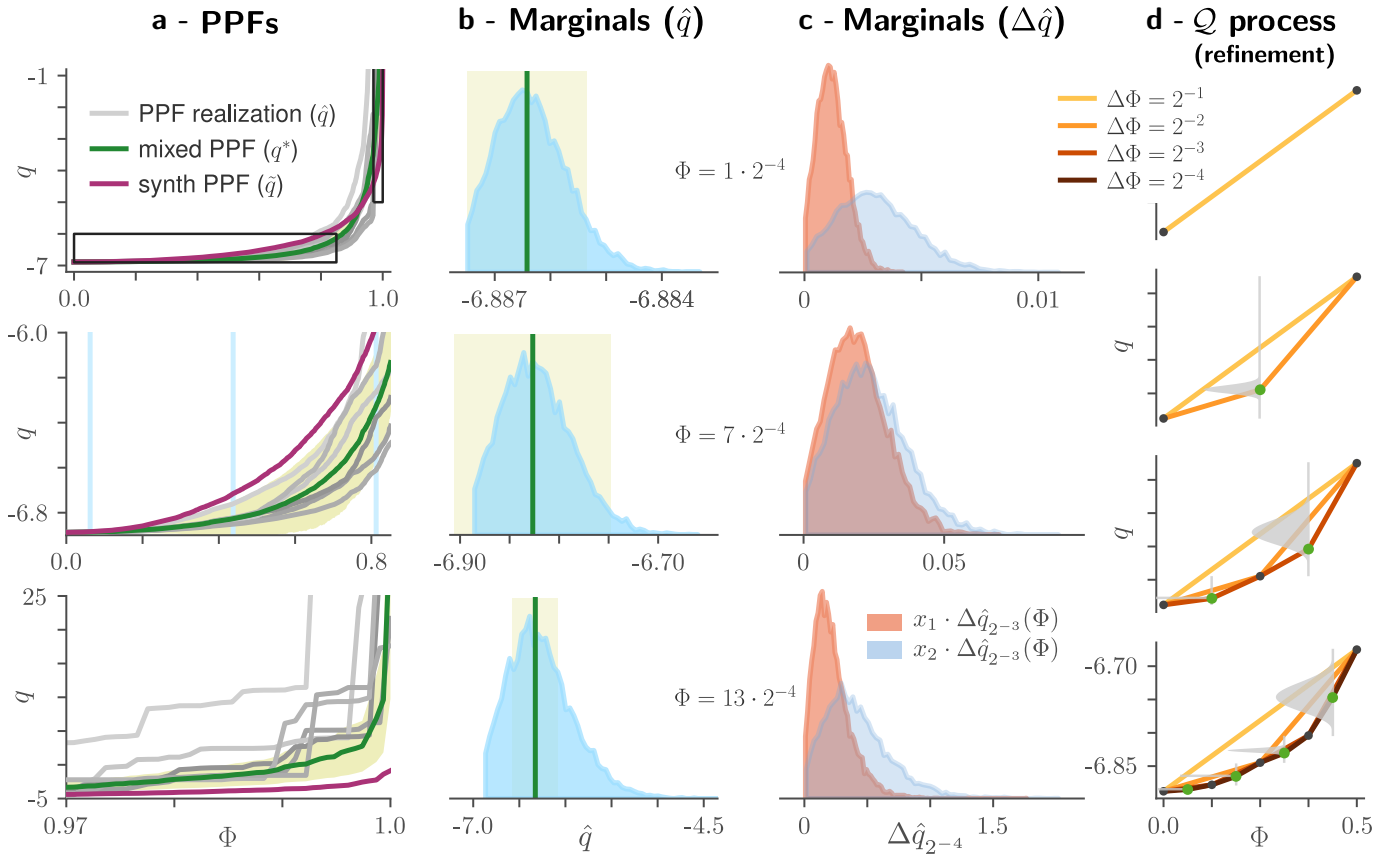


Figure 5: **Sampling a hierarchical beta (HB) process. (a)** PPF samples (grey lines) drawn from a hierarchical beta process $\mathfrak{Q}$. Mixed (green) and synthetic (red) PPFs are those for the Planck model of Fig. 7f (respectively q^* from equation (19) and $\tilde{q}$ from equation (22)). Lower panels are enlarged portions of the top panel, corresponding to the black rectangles in the latter. At each Φ, the variance between realisations is controlled by $\sqrt{c}\,\delta^{\mathrm{EMD}}$ (yellow shading), per equation (27); here we use $c = 0.5$. **(b)** Marginals of $\mathfrak{Q}$ at three values of Φ, obtained as histograms of 10,000 realisations $\hat{q}$. As in (a), the green line indicates the value of $q^*(\Phi)$ and the yellow shading describes the range $q^*(\Phi) \pm \sqrt{c}\delta^{\mathrm{EMD}}(\Phi)$. Values of Φ are given alongside the panels and drawn as cyan vertical lines in (a). **(c)** Distributions of the subincrements drawn at the same three Φ positions as in (b), obtained as histograms from the same 10,000 realisations. Subincrements are for the fourth refinement step, corresponding to the fourth panel of (d). Notation in the legend corresponds to equation (29). **(d)** Illustration of the refinement process. This $\mathfrak{Q}$ is parametrised by the same PPFs as the one in (a–c), but we used a larger c (16) to facilitate visualisation. Each refinement step halves the width $\Delta\Phi$ of an increment; here we show four refinements steps, while in most calculations we use eight. New points at each steps are coloured green; the beta distribution from which each new point is drawn is shown in grey. The domain of each of these distributions spans the vertical distance between the two neighbouring points and is shown as a grey vertical line.

To now put the B^{EMD} criterion on firmer footing, recall that it is meant to model variations between data replications. We can therefore validate the criterion by simulating such variations *in silico* (in effect simulating many replications of the experiment): since the true data-generating process $\mathcal{M}_{\text{true}}$ is then known, this gives us a ground truth of known comparison outcomes, against which we can test the probabilistic prediction made by B^{EMD}. Because this procedure will also serve to select a suitable value for the scaling constant c, we call it a *calibration experiment*.

To generate the required variations we use *epistemic distributions* of the type introduced at the top of our Results, of which we already gave an example in equation (13). With this we define an alternative tail probability (see the Methods for details),

$$B^{\text{epis}}_{AB;\Omega} \coloneqq P(R_A < R_B \mid \Omega)\,, \tag{33}$$

which uses the epistemic distribution Ω instead of the process $\mathfrak{Q}_A$ to represent epistemic uncertainty. We then look for $c > 0$ such that the criterion $B^{\text{EMD}}_{AB;c}$ a) is correlated with $B^{\text{epis}}_{AB;\Omega}$; and b) satisfies

$$\left|B^{\text{EMD}}_{AB;c} - 0.5\right| \lesssim \left|B^{\text{epis}}_{AB;\Omega} - 0.5\right|. \tag{34}$$

Equation (34) encodes a desire for a *conservative criterion*: we are most worried about incorrectly rejecting a model, i.e. of making a type I error. The consequences for a type II error (failing to reject either model) are in general less dire: it is an indication that we need better data to differentiate between models. This desire for a conservative criterion reflects a broad belief in science that it is better to overestimate errors than underestimate them. Given the ontological differences between B^{EMD} and B^{epis} however, in practice one will need to allow at least small violations, which is why we give the relation in equation (34) as $\lesssim$.

The advantage of a probability like $B^{\text{epis}}_{AB;\Omega}$ is that it makes explicit which kinds of replication variations are involved in computing the tail probability; the interpretation of equation (33) is thus clear. However it can only be computed after generating a large number of synthetic datasets, which requires a known and parametrisable data-generating process $\mathcal{M}_{\text{true}}$; on its own therefore, $B^{\text{epis}}_{AB;\Omega}$ cannot be used directly as a criterion for comparing models. *By choosing c such that the criteria are correlated, we transfer the interpretability of $B^{\text{epis}}_{AB;\Omega}$ onto $B^{\text{EMD}}_{AB;c}$*. Moreover, if we can find such a c, then we have de facto validated $B^{\text{EMD}}_{AB;c}$ for the set of simulated experiments described by Ω.

Since defining an epistemic distribution involves making many arbitrary choices, one may want to define multiple distributions $\Omega_1, \Omega_2, \dots$ to ensure that results are not sensitive to a particular choice of Ω. Equation (34) can easily be generalised to account for this, following a similar logic as equation (15), in which case it becomes

$$\left|B^{\text{EMD}}_{AB;c} - 0.5\right| \lesssim \min_{\Omega \in \{\Omega_1, \Omega_2, \dots\}} \left|B^{\text{epis}}_{AB;\Omega} - 0.5\right|. \tag{35}$$

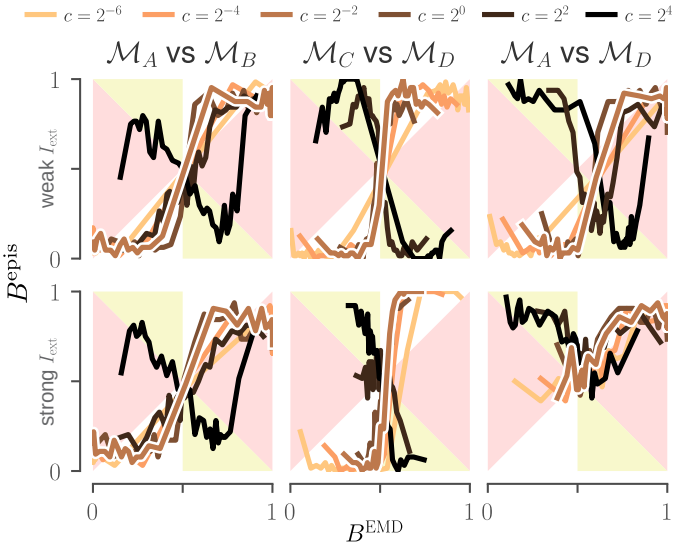


Figure 6: **Calibration curves for the LP models of Fig. 2.** Calibration curves for six epistemic distributions, computed following our proposed calibration procedure. Each curve summarizes 2048 simulated experiments with datasets of size 4000, and the six panels explore how curves depend on three parameters: the pair of models being compared, the constant c and the input strength I_{ext}. The latter is either `weak` **(top)** or `strong` **(bottom)**. Other parameters are kept fixed, namely a `short` correlation time τ and a `Gaussian` observation noise with `low` standard deviation. The regions depicted in red and yellow are those where equation (34) is violated. (For an extended version where all conditions are tested, see Supplementary Fig. 1. For example R-distributions for each of these c values, see Supplementary Fig. 2.)

We found that an effective way to verify equation (34) is by plotting $B^{\text{epis}}_{AB;\Omega}$ against $B^{\text{EMD}}_{AB;c}$, where values of $B^{\text{epis}}_{AB;\Omega}$ are obtained by averaging comparison outcomes, conditioned on the value of $B^{\text{EMD}}_{AB;c}$ being within an interval (this is explained more precisely in the Methods). We thus obtain a histogram of $B^{\text{epis}}_{AB;\Omega}$ against $B^{\text{EMD}}_{AB;c}$, which works best when the size of bins is adjusted so that they have similar statistical power. We illustrate this in Fig. 6, where histograms are shown as curves to facilitate interpretation. These curves are drawn against the "overconfident regions" (where equation (34) is violated), depicted in red or yellow: we look therefore for values of c which as much as possible stay within the white regions. The visual representation makes it easier to judge the extent to which small violations of equation (34) can be tolerated. An extended version of this figure, where all 48 conditions are tested, is provided as Supplementary Fig. 1.

Figure 6 shows calibration curves for six different epistemic distributions: three model pairs ($\mathcal{M}_A$ vs $\mathcal{M}_B$, $\mathcal{M}_C$ vs $\mathcal{M}_D$, and $\mathcal{M}_A$ vs $\mathcal{M}_D$), each with `weak` and `strong` external input I_{ext}. (For full details on the choice of epistemic distributions for these experiments, see the Methods.) The model pairs were chosen to test three different situations: one where the candidate models are similar both to each other and the observations ($\mathcal{M}_A$ vs $\mathcal{M}_B$), one where the candidate models

are similar to each other but different from the observations ($\mathcal{M}_C$ vs $\mathcal{M}_D$), and one where only one of the candidates is similar to the observations ($\mathcal{M}_A$ vs $\mathcal{M}_D$).

As one might expect, we see some violations of equation (35) in Fig. 6, which can partly be explained by our choice of a pointwise loss: although pedagogical, it tends to prefer models which produce fewer spikes, as we explain in the Methods. This is partly why, for these types of models, practitioners often prefer loss functions based on domain-specific features like the number of spikes or the presence of bursts.[24] We are also less concerned with the calibration of $\mathcal{M}_A$ vs $\mathcal{M}_D$, since those models are very different (c.f. Fig. 2). Not only is the B^{EMD} not needed to reject $\mathcal{M}_D$ in favour of $\mathcal{M}_A$, but the variational approximation expressed by equation (27) works best when the discrepancy between the model and the observed data (i.e. δ^{EMD}) is not too large.

Nevertheless, for values of c between 2^{-4} and 2^{0}, we see that calibration curves largely avoid the overconfidence (red and yellow) regions under most conditions and that B^{epis} is strongly correlated with B^{EMD}. This confirms that $B^{\text{EMD}}_{ab;c}$ can be an estimator for the probability $P(R_a < R_b)$ (recall equations (16) and (33)) and validates our choice of $c = 2^{-2}$ for the R-distributions in Fig. 2. More importantly, it shows that c does not need to be tuned to a specific value for B^{EMD} to be interpretable: it suffices for c to be within a finite range. We observe a similar region of validity for c with the model introduced in the next section (see Fig. 9 in the Methods). The fact that B^{EMD} remains valid for a range of c values is a major reason why we think it can be useful for analysing real data, where we don't know the epistemic distribution, as well as for comparing multiple models simultaneously.

Within its range of validity, the effect of c is somewhat analogous to a confidence level: just like different confidence levels will lead to different confidence intervals, different values of c can lead to different (but equally valid) values of $B^{\text{EMD}}_{ab;c}$. See the Supplementary Discussion for more details. Note also that due to the constraints of the $\mathfrak{Q}$ process, increasing c does not simply increase the spread of R-distributions, but can also affect their shape; this is illustrated in Supplementary Fig. 2.

There are limits however to the range of validity. Too small values of c will remove any overlap between R-distributions and produce an overconfident B^{EMD}. Too large values of c will exaggerate overlaps, underestimating statistical power. Large values can also lead to distortions in the $\mathfrak{Q}$ process, due to the monotonicity constraint placing an upper bound on the achievable metric variance; we see this as the curves reversing in Fig. 6 for $c \geq 2^2$. In the Supplementary Discussion we list some possible approaches to enlarge the range of validity, or otherwise improve the calibration of the B^{EMD}.

2.8 Characterizing the behaviour of R-distributions

To better anchor the interpretability of R-distributions, in this section we perform a more systematic study of the relationship between epistemic uncertainty, aleatoric uncertainty, and the shape of the R-distributions. To do this we use a different example, chosen for its illustrative simplicity, which allows us to independently adjust the ambiguity (how much two models are qualitatively similar) and the level of observation noise.

Concretely, we imagine a fictitious historical scenario where the Rayleigh-Jeans

$$\mathcal{B}_{\text{RJ}}(\lambda; T) = \frac{2ck_BT}{\lambda^4} \tag{36}$$

and Planck

$$\mathcal{B}_{\text{P}}(\lambda; T) = \frac{2hc^2}{\lambda^5} \frac{1}{\exp\left(\frac{hc}{\lambda k_B T}\right) - 1} \tag{37}$$

models for the radiance of a black body are two candidate models given equal weight in the scientific community. They stem from different theories of statistical physics, but both agree with observations at infrared or longer wavelengths, and so both are plausible if observations are limited to that window. (When it is extended to shorter wavelengths, the predictions diverge and it becomes clear that the Planck model is the correct one.)

Remark. For our purposes, these are just two models describing the relationship between an independent variable λ (the wavelength) and a dependent variable $\mathcal{B}$ (the spectral radiance), given a parameter T (the temperature) which is inferred from data; our discussion is agnostic to the underlying physics. The parameters h, c and k_B are known physical constants (the Planck constant, the speed of light and the Boltzmann constant) and can be omitted from the discussion.

We use a simple Poisson counting process to model the data-generating model $\mathcal{M}_{\text{true}}$ including the observation noise:

$$\mathcal{B} \mid \lambda, T, s \sim \frac{1}{s} \text{Poisson}\big(s\, \mathcal{B}_{\text{P}}(\lambda; T)\big) + \mathcal{B}_0\,, \tag{38}$$

where s is a parameter related to the gain of the detector (see the Methods for details). Most relevant to the subsequent discussion is that the mean and variance of $\mathcal{B}$ are

$$\mathbb{E}[\mathcal{B}] = \mathcal{B}_{\text{P}}(\lambda; T) + \mathcal{B}_0\,,$$
$$\mathbb{V}[\mathcal{B}] = \frac{\mathcal{B}_{\text{P}}(\lambda; T)}{s}\,,$$

and can therefore be independently controlled with the parameters $\mathcal{B}_0$ and s.

For the purposes of this example, both *candidate* models $\mathcal{M}_{\text{RJ}}$ and $\mathcal{M}_{\text{P}}$ make the incorrect (but common) assumption of additive Gaussian noise, such that instead of equation (38) they assume

$$\mathcal{B} \mid \lambda, T, \sigma \sim \mathcal{N}\big(\mathcal{B}_a(\lambda; T), \sigma^2\big)\,, \tag{39}$$

with $\mathcal{B}_a \in \{\mathcal{B}_{\text{RJ}}, \mathcal{B}_{\text{P}}\}$ and $\sigma > 0$. This ensures that there is always some amount of mismatch between $\mathcal{M}_{\text{true}}$ and the two candidates. That mismatch is increased when $\mathcal{B}_0 > 0$, which we interpret as a sensor bias which the candidate models neglect.

With this setup, we have four parameters which move the problem along three different "axes": The parameters λ_{min} and λ_{max} determine the spectrometer's detection window,

and thereby the **ambiguity**: the shorter the wavelength, the easier it is to distinguish the two models. The parameter s determines the **level of noise**. The parameter $\mathcal{B}_0$ determines an additional amount of **misspecification** between the candidate model and the data.

We explore these three axes in Fig. 7, and illustrate how the overlap of the R_P and R_RJ distributions changes through mainly two mechanisms: Better data can shift one R-distribution more than the other, and/or it can tighten one or both of the R-distributions. Either of these effects can increase the separability of the two distributions (and therefore the strength of the evidence for rejecting one of them).

2.9 Different criteria embody different notions of robustness

As a final step, we now wish to locate our proposed method within the wider constellation of model selection methods.

The motivation behind any model selection criterion is to make the selection in some way robust — i.e. to encourage a choice which is good not just for the given particular data and models, but also for variations thereof. Otherwise we would just select the model with the lowest empirical risk and be done with it. Criteria vary widely however in terms of what kinds of variations they account for. For the purposes of comparing with the EMD rejection rule, we consider three types:

In-distribution variations of the data where new samples are drawn from the *same* data-generating process $\mathcal{M}_\mathrm{true}$ as the data.

Out-of-distribution variations of the data where new samples are drawn from a *different* data-generating process $\mathcal{M}'_\mathrm{true}$ (than the data-generating process $\mathcal{M}_\mathrm{true}$). How much $\mathcal{M}'_\mathrm{true}$ is allowed to differ from $\mathcal{M}_\mathrm{true}$ will depend on the target application.

Variations of model parameters, for example by sampling from a Bayesian posterior, or refitting the model to new data.

Robustness to data variations, whether in- or out-of-distribution, is generally considered a positive. In contrast, robustness to model variations (what might more commonly be described as *insensitivity* to model parameters) is often attributed to excess model complexity and thus considered a negative.

Exactly which type of variation a criterion accounts for — and just as importantly, *how* it does so — defines what we might call the *paradigm* for that criterion. In Fig. 8 we compare the EMD approach to five other commonly used criteria with a variety of model selection paradigms; we expand on those differences below.

For our purposes the most important of those differences is how a criterion accounts for epistemic uncertainty. Four of the considered criteria do this by comparing parametrised *families* of models — either explicitly by averaging their performance over a prior (BIC, $\log \mathcal{E}$), or implicitly by refitting the model to each new dataset (MDL, AIC). In other words, such criteria compare the mathematical structure of different models, rather than specific parameters; they would not be suited to the neural circuit example of earlier sections, where the candidate models were solely distinguished by their parameters. The comparison we do here is therefore *only possible* when the models being compared are *structurally distinct*.

Our comparison is also tailored to highlight features of particular importance for experimental data and which differentiate our method. It is by no means exhaustive, and many aspects of model comparison are omitted — such as their consistency or how they account for informative priors. The example data were also chosen to be in a regime favourable to all models; this is partly why many panels show similar behaviour.

We can broadly classify criteria according to whether they select models based solely on their *predictive accuracy*, or whether they also penalize their *complexity* (so as to prefer simpler models).

Let us begin with the latter. The most common of these is likely the *Bayes factor*, formally defined as the ratio of model probabilities and in practise the ratio of marginal likelihoods, or *model evidence*:[40,67,19]

$$B_{AB}^\mathrm{Bayes} = \frac{p(\mathcal{M}_A \mid \mathcal{D})}{p(\mathcal{M}_B \mid \mathcal{D})} = \frac{\int p(\mathcal{D} \mid \theta, \mathcal{M}_A)\pi_A(\theta)d\theta}{\int p(\mathcal{D} \mid \theta, \mathcal{M}_B)\pi_B(\theta)d\theta} =: \frac{\mathcal{E}_A}{\mathcal{E}_B}\,. \tag{40}$$

(We previously defined the evidence $\mathcal{E}_A$ the same way in equation (6).) Here $p(\mathcal{D} \mid \theta, \mathcal{M}_A)$ and π_A are likelihood and prior distributions respectively. Going forward, instead of Bayes factors we will consider each model's log evidence: the former are easily recovered from the difference of the latter,

$$\log B_{AB}^\mathrm{Bayes} = \log \mathcal{E}_A - \log \mathcal{E}_B\,, \tag{41}$$

and this makes it easier to compare multiple models at once. It also matches the conventional format of information criteria. Numerical evaluation of the model evidence is non-trivial, but generic software solutions are available. For this comparison we used an implementation of nested sampling[59] provided by the Python package `dynesty`.[34]

The model evidence corresponds to studying *variations of the model* (represented by the prior) for a *fixed dataset* $\mathcal{D}$. A model is penalised if its prior is unnecessarily broad or high-dimensional; this is the so-called "Occam's razor" principle which favours simpler models[40,67] , which sometimes leads to undesirable results when priors are non-informative.[18]

The *Bayesian information criterion* (BIC) is an asymptotic approximation of the log evidence, either around the maximum likelihood or maximum a posteriori estimate.[57,19] It will approach the log evidence ($\log \mathcal{E}$) when the likelihood is non-singular and approximately Gaussian around its maximum.

Another approach which penalizes complexity is the *minimum description length (MDL)*. This is more accurately described as a methodological framework built around the choice of a "universal probability distribution";[25] usually however the *normalised maximum likelihood* (NML) distribution is implied, and this is what we do here. Assuming equal prior preference for each model, MDL selects the one with the highest NML probability; equivalently, it computes for each model

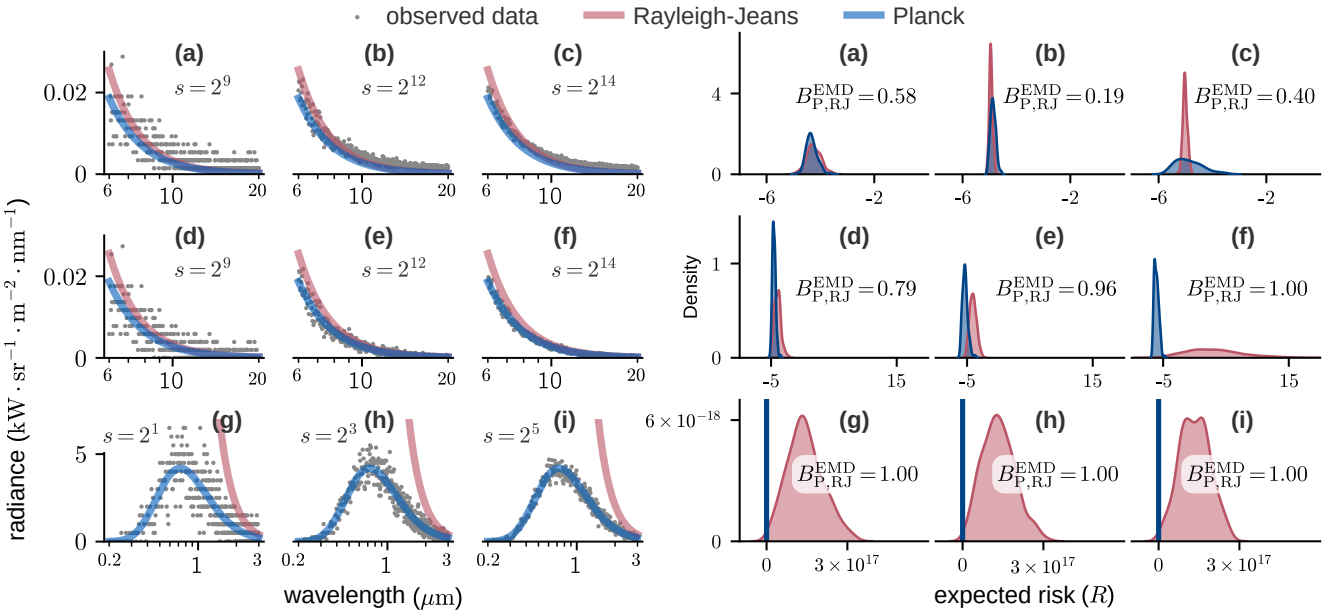


Figure 7: R-**distributions** (right) **for different simulated datasets of spectral radiance** (left). Datasets were generated using equation (38). For each model a, the R_a-distribution was obtained by sampling an HB process $\mathfrak{Q}$ parametrised by δ_a^{EMD} and a sensitivity $c = 2^{-1}$; see the respective Results sections for definitions of $\mathfrak{Q}$ and δ^{EMD}. A kernel density estimate is used to visualise the resulting samples (equation (30)) as densities. In all rows, noise (s) decreases left to right. **Top row:** Over a range of long wavelengths with positive bias $\mathcal{B}_0 = 0.0015$, both models fit the data equally well. **Middle row:** Same noise levels as the first row, but now the bias is zero. Planck model is now very close to $\mathcal{M}_{\text{true}}$, and consequently has nearly-Dirac R-distributions and lower expected risk. **Bottom row:** At visible wavelengths, the better fit of the Planck model is incontrovertible.

the quantity

$$\text{MDL}_A \coloneqq -\max_\theta p(\mathcal{D} \mid \mathcal{M}_A(\theta)) + \text{COMP}(\mathcal{M}_A)\,. \tag{42}$$

and selects the model for which MDL_A is minimal. (It is assumed here that $\mathcal{M}_A$ is parametrised, and $\mathcal{M}_A(\theta)$ denotes the model $\mathcal{M}_A$ with its parameters set to the values θ.) The first term is the maximum of the likelihood evaluated on the training data. The second term is the *MDL complexity*, defined by integrating over the entire event space:

$$\text{COMP}(\mathcal{M}_A) \coloneqq \int \max_\theta p(\mathcal{D}' \mid \mathcal{M}_A(\theta))\, d\mathcal{D}'\,. \tag{43}$$

In other words, MDL complexity quantifies the ability of a model to fit arbitrary data. Note that it does not matter here what data-generating process $\mathcal{M}_{\text{true}}$ generated the data since each possible dataset is given equal weight. Approximations are almost always required to compute equation (43) since the integral is usually intractable; for this comparison, we chose a data regime where it is just about possible to enumerate the possible datasets and estimate $\text{COMP}(\mathcal{M}_A)$ directly with Monte Carlo.

With MDL therefore we consider variations of the models, since each model is refitted to each dataset $\mathcal{D}'$. We have no in-distribution dataset variations (the true process $\mathcal{M}_{\text{true}}$ is ignored), but we do have out-of-distribution variations in the form of the integral over the event space.

Moving on now to criteria which compare predictive accuracy, we start with the *expected log pointwise predictive density (elpd)*.[71] This is defined as the expectation of the loss Q for L samples drawn from the data-generating process:

$$\begin{aligned}\text{elpd}_A &\coloneqq \sum_{i=1}^{L} \mathbb{E}_{(x_i,y_i)\sim\mathcal{M}_{\text{true}}}\left[Q(x_i, y_i; \mathcal{M}_A)\right] \\ &= L\,R_A\,.\end{aligned} \tag{44}$$

The most common form of the elpd uses the (negative) log likelihood as the loss function Q. In this form where each sample is drawn from the same distribution, the elpd is proportional to the risk R_A defined in equation (4), and therefore conceptually analogous: it describes robustness to in-distribution variations of the dataset. The models are kept fixed in these comparisons; they are not refitted to the new data. The uncertainty on elpd comes from the finite number of samples used to compute the expectation.

Since in current practice the elpd is mostly used with Bayesian models, this is also what we do in Fig. 8. Efficient estimators exist for computing the elpd from training data, using either leave-one-out cross-validation[71] or the widely applicable information criterion (WAIC).[72] This avoids the need to reserve a separate test dataset. We used `ArviZ`'s[35] implementation of WAIC for this comparison.

In contrast to the elpd, the *Akaike information criterion (AIC)* does not assume a fixed model. Although it also ranks models according to their expected loss, this time the expectation is

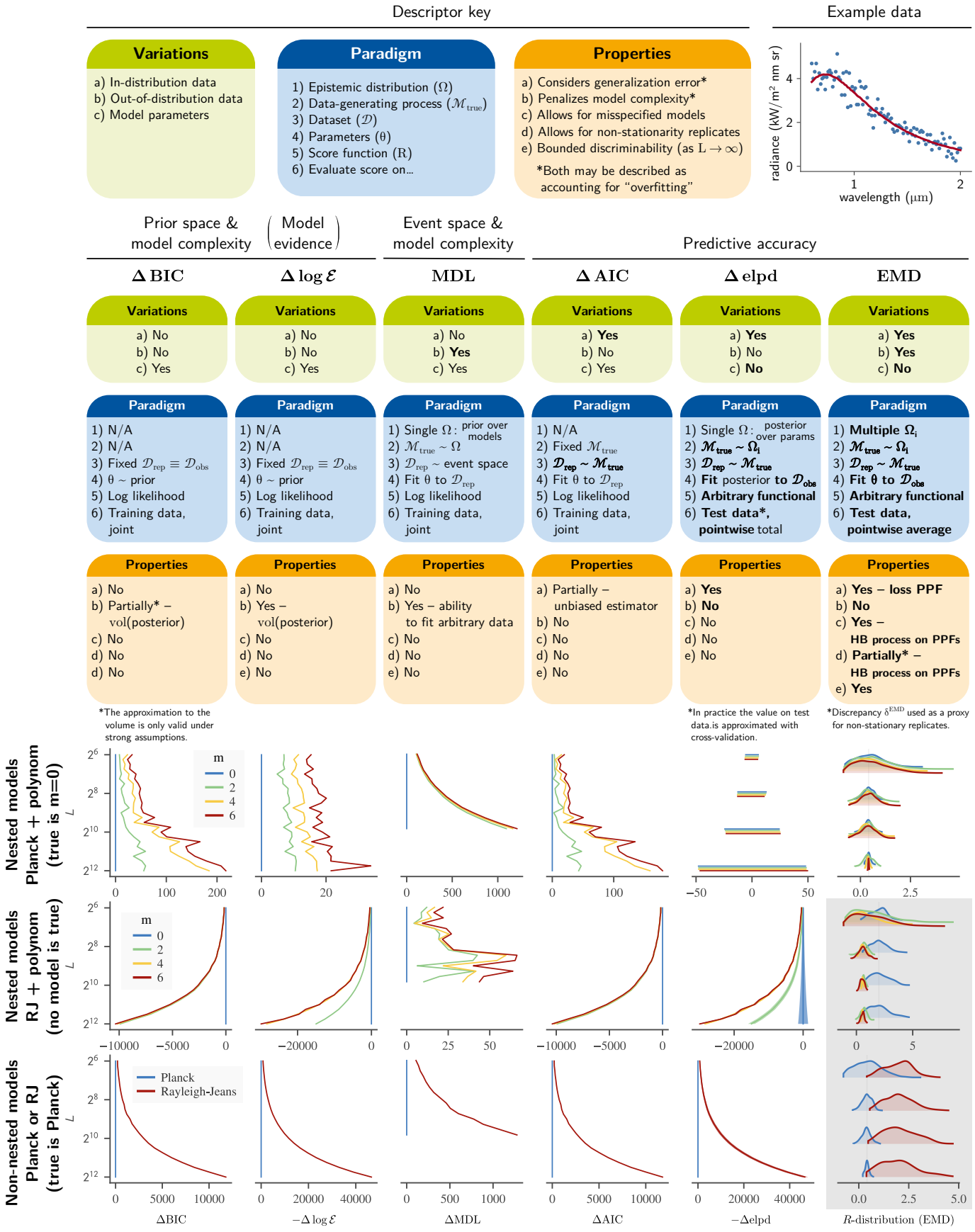


Figure 8: **Behaviour of model selection criteria as a function of dataset size** Datasets were generated using equation (38); they differ only in their number of samples and random seed; an example is shown in the top right. The comparison outcome between two models is given by the difference between their criteria; these are plotted as differences w.r.t. an arbitrary reference model (blue curve). The log evidence ($\log \mathcal{E}$) and elpd provide uncertainties, shown either as shading or as error bars. MDL curves are shorter because computations were aborted after 100 hours.

(conceptually) over both the test set $\mathcal{D}_{\text{test}}$ and training data $\mathcal{D}_{\text{train}}$, both drawn independently from the data-generating process:[2,1]

$$\text{elpd}_A^{\text{AIC}} = \mathbb{E}_{\mathcal{D}_{\text{test}}\sim\mathcal{M}_{\text{true}}}\mathbb{E}_{\mathcal{D}_{\text{train}}\sim\mathcal{M}_{\text{true}}}\left[Q\left(\mathcal{D}_{\text{test}}; \mathcal{M}_A\left(\hat{\theta}_{\mathcal{D}_{\text{train}}}\right)\right)\right], \tag{45}$$

where $\hat{\theta}_{\mathcal{D}_{\text{train}}}$ is the maximum likelihood estimate given the training data. Concretely however the AIC is a simple statistic, similar to the BIC and other information criteria:

$$\text{AIC}_A \coloneqq 2\,Q\left(\mathcal{D}_{\text{train}}; \mathcal{M}_A\left(\hat{\theta}_{\mathcal{D}_{\text{train}}}\right)\right) + 2k_A\,, \tag{46}$$

where k_A is the number of parameters of model $\mathcal{M}_A$. The derivation of the AIC relies on Q being the log likelihood, and assumes that the likelihood is approximately Gaussian and non-singular around $\hat{\theta}_{\mathcal{D}_{\text{train}}}$. When these conditions are verified, then $\Delta\,\text{AIC}_{AB} \coloneqq \text{AIC}_A - \text{AIC}_B$ is an unbiased estimator of the difference $\text{elpd}_A^{\text{AIC}} - \text{elpd}_B^{\text{AIC}}$. In this case, the model with the lowest AIC will — in expectation — also be the one with the lowest expected risk.

So how does the B^{EMD} compare to the criteria listed above? First, it computes R-distributions for *fixed models*. This is a key feature, since otherwise it is impossible to compare different parameter vectors for otherwise structurally identical models. (Note that EMD also allows models to have different structure — it simply does not require it.) This also means that the B^{EMD} does not penalise model complexity, in contrast to the evidence, MDL and AIC.

Second, like the elpd, the B^{EMD} accounts for *in-distribution variability* by estimating the risk using held-out test data $\mathcal{D}_{\text{test}}$.

Third, EMD R-distributions also account for *out-of-distribution variability*. Conceptually this is done via the epistemic distributions Ω, concretely via the HB process on PPFs — the variance of the latter stemming ultimately from the discrepancy function δ^{EMD}. Better models therefore result in less variability, reflecting our better grasp of the process being modelled. In contrast, only the MDL allows for some form of out-of-distribution variability (in the form of a complexity penalty), and it is essentially rigid: always an integral over the entire event space.

The biggest difference between the EMD R-distributions and other criteria however is their behaviour as the number of samples L increases. This we illustrate in the lower part of Fig. 8, which plots the value of a criterion for different models as a function of L. We do this for three collections of models, where the data-generating process is always the Planck model $\mathcal{B}_{\text{P}}$ described in the previous section. In all panels, scores are defined so that models with lower scores are preferred.

In the first row we compare four candidate models of the form

$$\mathcal{B} \mid \lambda, T, \sigma, \{b_j\} \sim \mathcal{N}\left(\mathcal{B}_{\text{P}}(\lambda; T) + \sum_{j=0}^{m} b_j \lambda^j,\ \sigma^2 \mathcal{B}_{\text{P}}(\lambda; T)\right). \tag{47}$$

The candidates are therefore *nested*: any model within the class $\mathcal{M}_{m=2}$ can be recovered from the class $\mathcal{M}_{m=4}$ by fixing $b_3 = b_4 = 0$. Moreover, all models effectively contain $\mathcal{M}_{\text{true}}$, since it corresponds to $m = 0$. (The data are in a regime where the Poisson distribution of $\mathcal{M}_{\text{true}}$ is close to Gaussian.)

The second row is similar, but expands around the Rayleigh-Jeans model instead:

$$\mathcal{B} \mid \lambda, T, \sigma, \{b_j\} \sim \mathcal{N}\left(\mathcal{B}_{\text{RJ}}(\lambda; T) + \sum_{j=0}^{m} b_j \lambda^j,\ \sigma^2 \mathcal{B}_{\text{RJ}}(\lambda; T)\right). \tag{48}$$

In this case all models are misspecified, but those with higher degree polynomials are better able to compensate for this and thus achieve lower prediction errors.

The third row compares the two $m = 0$ models from the first two rows.

If we interpret the score difference between two models (measured as the horizontal distance between two curves) as the strength of the evidence supporting a particular model choice, then it is clear from Fig. 8 that for all criteria except EMD, strength grows unbounded as we increase the number of samples L. A bounded strength however is desirable when analyzing experimental data, since no experiment has infinite accuracy.

In contrast, we see that the R-distributions in the grey shaded area stabilize as a function of L. Therefore also the tail probabilities $B^{\text{EMD}}_{ab;c} = P(R_A < R_B : c)$ (equation (16)), which define the EMD rejection rule, converge in general to finite, non-zero values. In some cases the R-distributions may approach a Dirac delta as L increases, as we see in the first row of Fig. 8. This occurs when the candidate model is able to match $\mathcal{M}_{\text{true}}$ exactly.

We provide a different set of comparisons in the Supplementary Results, listing numerical values of criteria differences in Supplementary Table 1 to complement the graphical presentation of Fig. 8. These alternative comparisons place greater emphasis on how misspecification can affect a criterion's value, and how that value can (mis)represent the strength of the evidence.

3 Discussion

Model selection in the scientific context poses unique challenges. Ranking models according to their empirical risk captures the scientific principle that models should be able to predict new data, but in order for results to be *reproducible* across experiments, that risk should also be *robust to variations in the replication process*. We propose that the empirical model discrepancy between quantile functions (δ^{EMD}) — measured from a single dataset — can serve as a baseline for the uncertainty across replications. We represent this uncertainty by the stochastic process $\mathfrak{Q}$; due to intrinsic constraints of quantile functions, this process is relatively stereotyped, allowing us to automate the generation of *R(isk)-distributions* — visually summarized in Fig. 1 — from empirical data (see Code availability). Moreover, we can calibrate the process against simulated replications — by adjusting the proportionality factor c which converts model discrepancy into epistemic uncertainty — so that it approximates the epistemic variability expected in experiments. In the end, the decision to reject a model in favour of another reduces to a simple

tail probability $B^{\mathrm{EMD}}_{AB;c} = P(R_A < R_B \mid c)$ between two R-distributions (equation (12)). Crucially, we do not force a model choice: a model is only rejected if this probability exceeds a chosen threshold ϵ. Guidelines for choosing rejection thresholds should be similar to those for choosing significance levels.

A key feature of our approach is to compare models based on their risk. This is an uncommon choice for statistical model selection, because it tends to select overfitted models when the same data are used both to fit and test them. In more data-rich machine learning paradigms however, risk is the preferred metric, and overfitting is instead avoided by testing models on held-out data. Our method assumes this latter paradigm, in line with our motivation to compare complex, data-driven models. A selection rule based on risk is also easily made consistent (in the sense of asymptotically choosing the correct model) by choosing a proper loss function, such as the negative log likelihood. Moreover, with the specific choice of a log likelihood loss, differences in risk can be interpreted as differences in differential entropy. The risk formulation however is more general, and allows the choice of loss to be tailored to the application.

We illustrated the approach on two example problems, one describing the radiation spectrum of a black body, and the other the dynamical response of a neuron of the lobster pyloric circuit. In the case of the former, we compared the Planck and Rayleigh-Jeans models; these being two structurally different models, we could also compare the B^{EMD} with other common selection criteria (Fig. 8). We thereby highlight not only that different criteria account for different types of variations, but also that B^{EMD} probabilities are unique in having well-defined, finite limits (i.e. neither zero nor infinite) when the number of samples becomes large. This is especially relevant in the scientific context, where we want to select models which work for datasets of all sizes. The simplicity of the black body radiation models also allowed us to systematically characterize how model ambiguity, misspecification, and observation noise affect R-distributions (Fig. 7).

The neural dynamics example was inspired by the increasingly common practice of data-driven modelling. Indeed, Prinz et al.[48,49] — whose work served as the basis for this example — can be viewed as early advocates for data-driven approaches. Although their exhaustive search strategy faces important limitations,[45] more recent approaches are much more scalable. For example, methods using gradient based optimization can simultaneously learn dozens — or in the case of neural networks, millions — of parameters.[54,73] These methods however are generally used to solve non-convex problems, and therefore run against the same follow-up question: *having found (possibly many) candidate models, which ones are truly good solutions, and which ones should be discarded as merely local optima?* The B^{EMD} probabilities, which account for epistemic uncertainty on a model's risk, can address this latter question. They can do so even when candidate models are structurally identical (distinguished therefore only by the values of their parameters), which sets this method apart from most other model selection criteria.

A few other model selection methods have been proposed for structurally identical candidate models, either within the framework of approximate Bayesian computing[66] or by training a machine learning model to perform model selection.[9] Largely these approaches extend Bayesian model selection to cases where the model likelihood is intractable, and as such should behave similarly to the Bayesian methods studied in Fig. 8. Epistemic uncertainty is treated simply as a prior over models, meaning in particular that these approaches do not account for variations in the replication process.

Epistemic uncertainty itself is not a novel idea: Kiureghian & Ditlevsen[33] discuss how it should translate into scientific practice, and Hüllermeier & Waegeman[30] do the same for machine learning practice. There again however, the fact that real-world replication involves variations of the data-generating process is left largely unaddressed. Conversely, the property of a model of being robust to replication variations is in fact well ingrained in the practice of machine learning, where it is usually referred to as "generalisability", or more specifically "out-of-distribution performance". This performance however is usually only measured post hoc on specific alternative datasets. Alternatively, there have been efforts to quantify the epistemic uncertainty by way of ensemble models, including the GLUE methodology,[62,5] Bayesian calibration,[32,3] Bayesian neural networks,[31] and drop-out regularization at test time.[15] Since these methods focus on quantifying the effect of uncertainty on *model predictions*, they improve the estimate of risk, but still do not assign uncertainty to that estimate. In contrast, with the hierarchical beta process $\mathfrak{Q}$, we express epistemic uncertainty on the *statistics of the sample loss*. This has two immediate advantages: the method trivially generalises to high-dimensional systems, and we obtain *distributions* for the risk (Fig. 3). In this way, uncertainty in the model translates to uncertainty in the risk.

A few authors have argued for making robustness against replication uncertainty a logical keystone for translating fitted models into scientific conclusions,[74,13] although without modelling replications explicitly. There, as here, the precise definition of what we called an epistemic distribution is left to the practitioner, since appropriate choices will depend on the application. One would of course expect a selection rule not to be too sensitive to the choice of epistemic distribution, exactly because it should be robust.

One approach which does explicitly model replications, and uses similar language as our own to frame its question, is that of Gelman *et al.*.[17] Those authors also model the distribution of discrepancies (albeit here again of model predictions rather than losses) and compute tail probabilities. Their method however is designed to assess the plausibility of a single model; it would not be suited to choosing the better of two approximations. It is also intrinsically Bayesian, using the learned model itself (i.e. the posterior) to simulate replications; we do not think it could be generalised to non-Bayesian models like our neural circuit example.

Building on similar Bayesian foundations but addressing model comparison, Moran *et al.*[44] propose the posterior predictive null check (PPN), which tests whether the predictions of two models are statistically indistinguishable. In contrast to the B^{EMD}, the PPN is purely a significance test: it can detect

if two models are distinguishable, but not which one is better, or by how much. One can see parallels between this approach and the model selection tests (MST) proposed by Golden.[22,23] While MST is framed within the same paradigm as AIC — and so makes the same approximation that inferred parameters follow a Gaussian distribution — instead of simply proposing an unbiased estimator for the difference $R_A - R_B$, that difference is mapped to a χ^2 distribution. This allows the author to propose a selection procedure similar in spirit to equation (11), but where the "no rejection" case is selected on the basis of a significance test. Although replication uncertainty is not treated — and therefore the test can be made arbitrarily significant by increasing the number of samples — the author anticipates our emphasis on misspecification and our choice to select models based on risk.

A guiding principle in designing the EMD rejection rule was to emulate how scientists interpret noisy data. This approach of using conceptually motivated principles to guide the development of a method seems to be fruitful, as it has lead to other recent developments in the field of model inference. For instance, Swigon *et al.*[63] showed that by requiring a model to be invariant under parameter transformation, one can design a prior (i.e. a regulariser) which better predicts aleatoric uncertainty. Another example is simulation-based calibration (SBC),[64,43] for which the principle is self-consistency of the Bayesian model. This is conceptually similar to the calibration procedure we proposed: in both cases a consistency equation is constructed by replacing one real dataset with many simulated ones. The main difference is that SBC checks for consistency with the Bayesian prior (and therefore the aleatoric uncertainty), while we check for consistency with one or more epistemic distributions. In both cases the consistency equation is represented as a histogram: in the case of SBC it should be flat, while in our case it should follow the identity $B^{\text{epis}} = B^{\text{EMD}}$.

The model discrepancy δ^{EMD} (equation (23)) also has some similarities with the Kolmogorov-Smirnoff (K-S) statistic; the latter is defined as the difference between two CDFs and is also used to compare models. Methodologically, the K-S statistic is obtained by taking the supremum of the difference, whereas we keep δ^{EMD} as a function over $[0, 1]$; moreover, a Kolmogorov-Smirnoff test is usually computed on the distribution of data samples ($\mathcal{D}_{\text{test}}$) rather than that of their losses ($\{Q(x, y) : (x, y) \in \mathcal{D}_{\text{test}}\}$).

One would naturally expect a comparison criterion to be symmetric: the evidence required to reject model $\mathcal{M}_A$ must be the same whether we compare $\mathcal{M}_A$ to $\mathcal{M}_B$ or $\mathcal{M}_B$ to $\mathcal{M}_A$. Moreover, it must be possible that *neither* model is rejected if the evidence is insufficient. Bayes factors and information criteria do allow for this, but with an important caveat: the relative uncertainty on those criteria is strongly tied to dataset size, so much so that with enough samples, there is *always* enough data to reject a model, as we illustrate in Fig. 8. This is in clear contradiction with scientific experience, where more data cannot compensate for bad data.

Another important consequence of a symmetric criterion is that it trivially generalises to comparing any number of models, such as we did in Table 1. One should however take care in such cases that models are not simply rejected by chance when a large number of models are simultaneously compared. By how much that chance would increase, and how best to account for this, are questions we leave for future work.

From a practical standpoint, the most important feature of the B^{EMD} is likely its ability to compare both specific model parametrisations of the same structural model, as well as structurally different models. Most other criteria listed in Fig. 8 only compare model structures, because they either assume globally optimal parameters (AIC, BIC), integrate over the entire parameter space (Bayes factor) or refit their parameters to each dataset (AIC, MDL). The calculation of the $B^{\text{EMD}}_{AB;c}$ between two models is also reasonably fast, taking less than a minute in most of our examples. Although the calibration experiments are more computationally onerous, for both of our models we found that the $B^{\text{EMD}}_{AB;c}$ is valid over a range of c values, with the high end near $c = 1$. This suggests a certain universality to the $B^{\text{EMD}}_{AB;c}$, which would allow practitioners to perform initial analyses with a conservative value like $c = 1$, only then proceeding to calibration experiments if there is indication that they are worth pursuing. Alternatively, since calibration is a function of experimental details, a laboratory might perform it once to determine a good value of c for its experimental setup, and then share that value between its projects.

This property that the same value of c can be used to compare different models in different contexts is the key reason why we expect the B^{EMD} to generalise from simulated epistemic distributions to real-world data. It is clear however that this cannot be true for arbitrary epistemic distributions: we give some generic approaches for improving the outcome of calibrations in the Supplementary Discussion, but the B^{EMD} criterion will always work best in cases where one has a solid experimental control (to minimize variability between replications) and deep domain knowledge (to accurately model both the system and the replications).

Another important feature is that computing the B^{EMD} only requires knowledge which is usually accessible in practice: a method to generate synthetic samples from both $\mathcal{M}_A$ and $\mathcal{M}_B$, a method to generate true samples, and a loss function. Some reasoned choices are used to define the stochastic process $\mathfrak{Q}$ in the PPF space, but they are limited by the hard constraints of that space: PPFs must be one-dimensional on $[0, 1]$, monotone, integrable, and non-accumulating. Although they make defining an appropriate stochastic process considerably more complicated, these constraints seem to be key for obtaining distributions which are representative of epistemic uncertainty (see Fig. 4). We proposed *hierarchical beta (HB) processes* (further detailed in the Methods) to satisfy these constraints, but it would be an interesting avenue of research to look for alternatives which also satisfy the desiderata for $\mathfrak{Q}$. In particular, if one could define processes which better preserve the proportionality of equation (27) at large values of c, or allow for a faster computational implementation, the B^{EMD} could be applied to an even wider range of problems.

There are other open questions which would be worth clarifying in future work. One is the effect of finite samples: having fewer samples increases the uncertainty on the shape of the estimated mixed PPF q^*, and should therefore increase

the discrepancy with the synthetic PPF $\tilde{q}$ (and thereby the estimated epistemic uncertainty). In this sense our method accounts for this implicitly, but whether it should do so explicitly — and how — is yet unexplored. It may also be desirable in some cases to use the same samples both to fit and test models, as it is commonly done in statistical model selection. One might explore for such cases the possibility of combining our approach with existing methods, such as those developed to estimate the elpd,[71,58] to avoid selecting overfitted models.

Looking more broadly, it is clear that the wide variety of possible epistemic distributions cannot be fully captured by the linear relationship of equation (27), which formalises as a desideratum for the stochastic process $\mathfrak{Q}$ what we referred to as the Emdprinciple 2.6. Going forward, it will be important therefore to consider the B^{EMD} criterion an imperfect solution, and to continue looking for ways to better account for epistemic uncertainty. These could include relating misspecification directly to the distribution of losses (going further than the B^{Q} considered our Supplementary Discussion), or more sophisticated self-consistency metrics than our proposed δ^{EMD}. A key strategy here may be to look for domain-specific solutions.

Already however, the B^{EMD} as proposed offers many opportunities to incorporate domain expertise: in the choice of the candidate models, loss function, sensitivity parameter c, rejection threshold ϵ, and epistemic distributions used to validate c. These are clear interpretable choices which help express scientific intent. These choices, along with the assumptions underpinning the B^{EMD}, should also become *testable*: if they lead to selecting models which continue to predict well on new data, that in itself will provide for them a form of empirical validation.

4 Methods

4.1 Poisson noise model for black body radiation observations

In equation (38), we used a Poisson counting process to simulate the observation noise for recordings of a black body's radiance. This is a more realistic model than Gaussian noise for this system, while still being simple enough to serve our illustration.

The physical motivation is as follows. We assume that data are recorded with a spectrometer which physically separates photons of different wavelengths and measures their intensity with a CCD array. We further assume for simplicity that wavelengths are integrated in bins of equal width, such that the values of λ are sampled uniformly (the case with non-uniform bins is less concise but otherwise equivalent). We also assume that the device uses a fixed time window to integrate fluxes, such that what it detects are effective photon *counts*. The average number of counts is proportional to the radiance, but also to physical parameters of the sensor (including size, integration window and sensitivity) which we collect into the factor s; the units of s are $\text{m}^2 \cdot \text{nm} \cdot \text{photons} \cdot \text{sr} \cdot \text{kW}^{-1}$, such that $s\mathcal{B}_a(\lambda; T)$ is a number of photons. Since the photons are independent, the recorded number of photons will be random and follow a Poisson distribution. This leads to the following data-generating process $\mathcal{M}_{\text{true}}$:

$$\mathcal{B} \mid \lambda, T \sim \frac{1}{s}\,\text{Poisson}\big(s\,\mathcal{B}_{\text{P}}(\lambda; T)\big) + \mathcal{B}_0\,, \qquad \text{(repeated from (39))}$$

$$\lambda \in \{\lambda_{\min}, \lambda_{\min} + \Delta\lambda, \lambda_{\min} + 2\Delta\lambda, \ldots, \lambda_{\max}\}\,. \tag{49}$$

Here $\mathcal{B}_0$ captures the effect of dark currents and random photon losses on the radiance measurement. Recall that a random variable k following a distribution Poisson(μ) has probability mass function $P(k) = \frac{\mu^k e^{-\mu}}{k!}$.

Note that we divide by s so that $\mathcal{B}$ also has dimensions of radiance and is comparable with the models $\mathcal{B}_{\text{P}}$ and $\mathcal{B}_{\text{RJ}}$ defined in equations (36) and (37).

4.2 Neuron model

The neuron model used in our Results is the Hodgkin-Huxley-type model of the lobster pyloric rhythm studied by Prinz *et al.*.[48] The specific implementation we used can be obtained from reference 51, along with a complete description of all equations and parameters. We summarize the model below and refer the reader to that reference for more details.

For each cell k, the potential across a patch of area A and capacitance C evolves according to the net ionic current through the cell membrane:

$$\frac{C}{A}\frac{dV^k}{dt} = \underbrace{-\sum_{i \in \text{ion channels}} I_i^k}_{\substack{\text{ion diffusion}\\ \text{through membrane}}} \underbrace{-\sum_{l \in \text{neurons}} I_s^l}_{\substack{\text{chemical}\\ \text{synapses}}} \underbrace{-\sum_{l \in \text{neurons}} I_e^{kl}}_{\substack{\text{electrical}\\ \text{synapses}}} - I_{\text{ext}}^k\,, \tag{50}$$

where I_{ext} is an arbitrary external current, the I_i describe ion exchanges between a cell and its environment, and I_s and I_e describe ion exchanges between different cells. The I_{ext} current can be used to represent a current applied by the experimenter, or the inputs from other cells in the network. The voltage $\tilde{V}$ which an experimenter records would then be some corrupted version of V, subject to noise sources which depend on their experiment.

$$\tilde{V}(t) \sim \text{Experimental noise}\Big(V(t)\Big)\,. \tag{51}$$

We generate the external input I_{ext} as a Gaussian coloured noise with autocorrelation:

$$\Big\langle I_{\text{ext}}(t) I_{\text{ext}}(t')\Big\rangle = \sigma_i^2 e^{-(t-t')^2/2\tau^2}\,. \tag{52}$$

We do this using an implementation[53] of the sparse convolution algorithm.[36] We found that in addition to being more realistic, coloured noise also smears the model response in time and thus reduces degeneracies when comparing models.

Each cell in the model has eight currents through ion channels, indexed by i: one Na^+ current, two Ca^{2+} currents, four K^+ currents and one leak current. Each current is modelled

as (square brackets indicate functional dependence)

$$I_e^{kl} = g_e\left(V^k - V^l\right), \qquad \tau_{m,i}\left[V^k\right]\frac{dm}{dt} = m_{\infty,i}\left[V^k\right] - m_i^k,$$
$$I_i^k = g_i^k\left(m_i^k\right)^{p_i} h_i\left(V^k - E_i\right), \qquad \tau_{h,i}\left[V^k\right]\frac{dh}{dt} = h_{\infty,i}\left[V^k\right] - h_i^k,$$
$$I_s^{kl} = g_s^{kl}\, s^l\left(V^k - E_s^k\right), \qquad \tau_s^l\left[V^l\right]\frac{ds}{dt} = s_\infty^l\left[V^l\right] - s^l.$$

These equations are understood as describing currents through permeable channels with maximum conductivity g_i^k (for membrane currents within the same cells) or g_s^{kl} (for synaptic currents between different cells). Conductivities are dynamic: they are governed by the equations for the gating variables m, h and s given above. (Some channels do not have inactivating gates; for these, h_i is set to 1.) The fixed points m_∞, h_∞ and s_∞, as well as the time constants τ_m, τ_h and τ_s, are functions of the voltage; the precise shape of these functions is specific to each channel type and can be found in either {only-html}«Prinz *et al.*[48] or René[51]»references 48 or 51.

Following Prinz et al., we treat the functions m_∞, h_∞, s_∞, τ_m, τ_h and τ_s, as well as the electrical conductance g_e and the Nernst (E_i) and synaptic ($E_{s,k}$) reversal potentials, as known fixed quantities. Thus the only free parameters in this model are the maximum conductances g_i^k and g_s^{kl}: the former determine the type of each neuron, while the latter determine the circuit connectivity.

The pyloric circuit model studied by Prinz *et al.*[49] consists of three populations of neurons with eight different ion channels. Biophysically plausible values for the channel and connectivity parameters were determined through separate exhaustive parameter searches by Prinz et al.,[49,48] the results of which were reduced to sixteen qualitatively different parameter solutions: 5 *AB/PD* cells, 5 *LP* cells and 6 *PY* cells. Importantly, these parameter solutions are distinct: interpolating between them does not yield models which reproduce experimental recordings. For purposes of illustration we study the simple two-cell circuit shown in Fig. 2**a**, where an AB cell drives an LP cell. Moreover we assume the parameters of the AB cell to be known, such that we only need to compare model candidates for the LP cell. (These assumptions are not essential to applying our method, but they avoid us contending with model-specific considerations orthogonal to our exposition.)

The AB neuron is an autonomous pacemaker and serves to drive the circuit with realistic inputs; all of our examples use the same AB model (labelled 'AB/PD 3' in Table 2 of Prinz *et al.*[48]), whose output is shown in Fig. 2**b**. Panels **c** and **d** show the corresponding response for each of the five LP models given in Table 2 of Prinz *et al.*.[48]

To generate our simulated observations, we use the output of `LP 1`, add Gaussian noise and then round the result (in millivolts) to the nearest 8-bit integer; in this way $\mathcal{M}_{\text{true}}$ includes both electrical and digitization noise. This leaves `LP 2` through `LP 5` to serve as candidate models; for these we assume only Gaussian noise, and we label them $\mathcal{M}_A$ to $\mathcal{M}_D$. We use Θ_a to denote the concatenation of all parameters for a given model a, which here consist of the vectors of conductance values g_i and g_s. Model definitions are summarized in Table 2 and Algorithm 1.

Table 2: Neuron circuit model labels

Model symbol	Model components
$\mathcal{M}_{\text{true}}$	$I_{\text{ext}} \to$ `AB 3` $\to$ **LP 1** $\to$ Gaussian noise $\to$ digitize
$\mathcal{M}_A$	$I_{\text{ext}} \to$ `AB 3` $\to$ **LP 2** $\to$ Gaussian noise
$\mathcal{M}_B$	$I_{\text{ext}} \to$ `AB 3` $\to$ **LP 3** $\to$ Gaussian noise
$\mathcal{M}_C$	$I_{\text{ext}} \to$ `AB 3` $\to$ **LP 4** $\to$ Gaussian noise
$\mathcal{M}_D$	$I_{\text{ext}} \to$ `AB 3` $\to$ **LP 5** $\to$ Gaussian noise

Algorithm 1 Neuron model

procedure GENERATE DATA($\mathcal{T}$)
  for $t \in \mathcal{T}$ **do**
    integrate equation (50) to obtain $V^{\mathsf{AB}}(t)$, $V^{\mathsf{LP}}(t;\Theta_{\text{true}})$
    draw $\xi(t) \sim \mathcal{N}(0, \sigma_o)$
    evaluate $\tilde{V}^{\mathsf{LP}}(t) = \texttt{digitize}_8\left(V^{\mathsf{LP}}(t) + \xi(t)\right)$
  end for
  return $\left\{\tilde{V}^{\mathsf{LP}}(t) : t \in \mathcal{T}\right\}$
end procedure

procedure SIMULATE CANDIDATE($\mathcal{T}$, $a \in \{A, B, C, D\}$)
  for $t \in \mathcal{T}$ **do**
    integrate equation (50) to obtain $V^{\mathsf{AB}}(t)$, $V^{\mathsf{LP}}(t;\Theta_a)$
    draw $\xi(t) \sim \mathcal{N}(0, \sigma_o)$
    evaluate $\tilde{V}^{\mathsf{LP}}(t) = V^{\mathsf{LP}}(t;\Theta_a) + \xi(t)$
  end for
  return $\left\{\tilde{V}^{\mathsf{LP}}(t) : t \in \mathcal{T}\right\}$
end procedure

procedure $\texttt{digitize}_8$(x)
Simulates the data encoding of a digital sensor by converting x to an 8-bit integer
  clip: $x \leftarrow \texttt{max}\left(-128, \texttt{min}(x, 127)\right)$
  return $\texttt{int}(x)$
end procedure

4.3 Loss function for the neuron model

To evaluate the risk of each candidate model, we use the log likelihood of the observations (equation (9)). This standard choice is convenient for exposition purposes: it is simple to explain and illustrates the generality of the method, since a likelihood function is available for any model in the form of equation (1).

However it is not a requirement to use the negative log likelihood as the loss, and in fact for time series models it can be disadvantageous. For example, the neuron models used in this work have sharp temporal responses (spikes), which makes the log likelihood sensitive to the timing of these spikes. In practice a less sensitive loss function may be preferable, although the best choice will depend on the application.

4.4 Construction of an R-distribution from an HB process $\mathfrak{Q}$

For each candidate model we use the hierarchical beta process $\mathfrak{Q}_a$ described below to generate on the order of $M_a \approx 100$ PPFs $\hat{q}_{a,1}, \ldots, \hat{q}_{a,M_a}$; the exact number of PPFs is determined automatically, by increasing the number M_a until the relative

standard error on $\frac{1}{M_a}\sum_{i=1}^{M_a} R[\hat{q}_{a,i}]$ is below 2^{-5} (six such PPFs are shown as grey traces in Fig. 5). Each curve is integrated to obtain a value for the risk (equation (20)), such that the R_a distribution can be represented by the set $\{R[\hat{q}_{a,i}]\}_{i=1}^{M_a}$. We then use a kernel density estimate to visualise these distributions in Fig. 3; specifically we use the `univariate_kde` function provided by Holoviews[56] with default parameters. The function automatically determines the bandwidth.

4.5 Calibration experiments

As described in our Results, the goal of calibration is twofold. First we want to correlate the value of $B^{\text{EMD}}_{AB;c}$ with the probability $B^{\text{epis}}_{AB;\Omega}$ that, across variations of the data-generating process $\mathcal{M}_{\text{true}}$, model A has lower risk than model B. We use an *epistemic distribution* Ω to describe those variations. Second, we want to ensure that a decision to reject either model based on B^{EMD}_{AB} is robust: it should hold for any reasonable epistemic distribution Ω of experimental conditions (and therefore hopefully also for *unanticipated* experimental variations).

The calibration procedure involves fixing two candidate models and an epistemic distribution Ω. We then sample multiple data-generating models $\mathcal{M}_{\text{true}}$ from Ω, generate a replicate dataset $\mathcal{D}^{\text{rep}}_{\text{test}}$ from each, and compute both B^{EMD} and B^{epis} on that replicate dataset.

This process can be repeated for as many epistemic distributions and as many different pairs of candidate models as desired, until we are sufficiently confident in the robustness of B^{EMD}. This means in particular that we cannot expect perfect correlation between B^{EMD} and B^{epis} across all models and conditions.

However we don't actually require perfect correlation because this is an *asymmetrical test*: a decision to reject *neither* model is largely preferred over a decision to reject the *wrong* one. Therefore what we especially want to ensure is that for any pair of models A, B, and any reasonable epistemic distribution Ω, we have

$$\left|B^{\text{EMD}}_{AB;c} - 0.5\right| \lesssim \left|B^{\text{epis}}_{AB;\Omega} - 0.5\right|. \tag{53}$$

This is given as either equation (34) or equation (35) in the Results.

Note that $B^{\text{epis}}_{AB;\Omega}$ is defined as the fraction of experiments (i.e. data-generating models $\mathcal{M}_{\text{true}}$) for which R_A is smaller than R_B. Therefore it cannot actually be calculated for a single experiment; only over sets of experiments. Since our goal is to identify a correlation between $B^{\text{epis}}_{AB;\Omega}$ and $B^{\text{EMD}}_{AB;c}$, we compute the former as a function of the latter:

$$\begin{aligned} B^{\text{epis}}_{AB;\Omega}\left(B^{\text{EMD}}_{AB;c}\right) &:= P\left(R_A < R_B \;\middle|\; \Omega, B^{\text{EMD}}_{AB;c}\right) \\ &= P_{\mathcal{M}_{\text{true}}\sim\Omega|B^{\text{EMD}}_{AB;c}}\left(R_A < R_B\right) \\ &= \mathbb{E}_{\mathcal{M}_{\text{true}}\sim\Omega|B^{\text{EMD}}_{AB;c}}\left[R_A < R_B\right]. \end{aligned} \tag{54}$$

(On the last line, [ ] is the Iverson bracket, which is 1 when the clause it contains is true, and 0 otherwise.) In other words, we partition the experiments in the domain of Ω into subsets having the same value $B^{\text{EMD}}_{AB;c}$, then compute $B^{\text{epis}}_{AB;\Omega}$ over each subset. In practice we do this by binning the experiments according to $B^{\text{EMD}}_{AB;c}$, as we explain below.

4.5.1 *Calibration and computation of B^{epis} for the black body radiation models*

For the black body models, we consider one epistemic distribution Ω with two sources of experimental variability: the true physical model ($\mathcal{B}_{\text{P}}$ or $\mathcal{B}_{\text{RJ}}$) and the bias $\mathcal{B}_0$. All other parameters are fixed to the standard values used throughout the text ($s = 10^5\,\text{m}^2\cdot\text{nm}\cdot\text{photons}\cdot\text{sr}\cdot\text{kW}^{-1}$, $T = 4000\,\text{K}$).

It makes sense to use both candidate models to generate synthetic replicate datasets, since ostensibly both are plausible candidates for the true data generating process. This is also an effective way to ensure that calibration curves cover both high and low values of B^{EMD}.

To generate each replicate dataset $\mathcal{D}^{\text{rep}}_{\text{test}}$, we select

$$\mathcal{B}_{\text{phys}} \sim \text{Unif}\left(\left\{\mathcal{B}_{\text{P}}, \mathcal{B}_{\text{RJ}}\right\}\right) \tag{55}$$

$$\mathcal{B}_0 \sim \text{Unif}\left(\left[-10^{-4}, 10^{-4}\right]\right)\cdot\text{m}^2\cdot\text{nm}\cdot\text{photons}\cdot\text{sr}\cdot\text{kW}^{-1}. \tag{56}$$

We then generate 4096 data points for each dataset using the Poisson observation model of equation (38), substituting $\mathcal{B}_{\text{phys}}$ for $\mathcal{B}_{\text{P}}$.

As in the main text, both candidate models assume Gaussian noise, so for each replicate dataset and each candidate model we fit the standard deviation of that noise to the observation data by maximum likelihood.

We then compute $B^{\text{EMD}}_{\text{P,RJ};c} \in [0,1]$ for each value of c being tested. We also compute a binary variable — represented again with the Iverson bracket — which is 1 if the true risk of $\mathcal{M}_{\text{P}}$ is lower risk than that of $\mathcal{M}_{\text{RJ}}$, and 0 otherwise:

$$\left[R_{\text{P}} < R_{\text{RJ}}\right] := \begin{cases} 1 & \text{if } R_{\text{P}} < R_{\text{RJ}}\,, \\ 0 & \text{otherwise.} \end{cases} \tag{57}$$

(In practice the true risks R_{P} and R_{RJ} are computed as the empirical risk with a very large number of data points.)

This produces a long list of $\left(B^{\text{EMD}}_{\text{P,RJ};c}, \left[R_{\text{P}} < R_{\text{RJ}}\right]\right)$ pairs, one for each dataset $\mathcal{D}^{\text{rep}}_{\text{test}}$. Formally, we can then compute $B^{\text{epis}}_{\text{P,RJ};\Omega}(B^{\text{EMD}}_{\text{P,RJ};c})$ as per equation (54): by keeping only those pairs with matching $B^{\text{EMD}}_{\text{P,RJ};c}$ value and averaging $B^{\infty}_{\text{P,RJ}}$. In practice, all $B^{\text{EMD}}_{\text{P,RJ};c}$ values will be slightly different, so we bin neighbouring values together into a histogram. To make best use of our statistical power, we assign the same number of pairs to each bin, which means that bins have different widths; the procedure is as follows:

1. Sort all $(B^{\text{EMD}}_{\text{P,RJ};c}, [R_{\text{P}} < R_{\text{RJ}}])$ pairs according to $B^{\text{EMD}}_{\text{P,RJ};c}$.
2. Bin the experiments by combining those with similar values of $B^{\text{EMD}}_{\text{P,RJ};c}$. Each bin should contain a similar num-

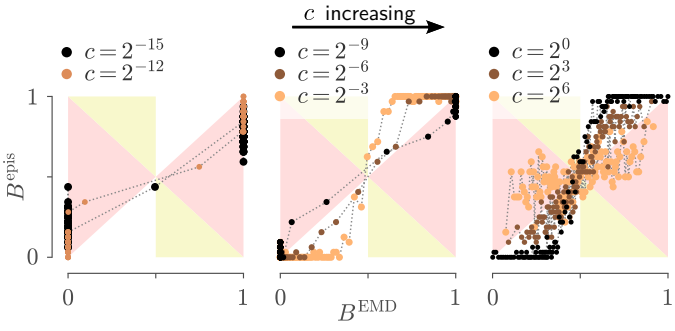


Figure 9: **Calibration plots for the $B^{\text{EMD}}_{\text{P,RJ}}$ criterion comparing Planck and Rayleigh-Jeans models.** The value of c increases from left to right and dark to bright. Each curve consists of 4096 experiments aggregated into 128 bins. Wide gaps between points indicate that few experiments produced similar values of B^{EMD}; this is especially salient in the left panel, where almost all experiments have either $B^{\text{EMD}} = 0$ or $B^{\text{EMD}} = 1$.

ber of experiments, so that they have similar statistical power.

3. Within each bin, average the values of $B^{\text{EMD}}_{\text{P,RJ};c}$ and $[R_{\text{P}} < R_{\text{RJ}}]$, to obtain a single pair $(\bar{B}^{\text{EMD}}_{\text{P,RJ};c}, \bar{B}^{\text{epis}}_{\text{P,RJ}})$. These are the values we plot as calibration curves.

The unequal bin widths are clearly visible in Fig. 9, where points indicate the $(\bar{B}^{\text{EMD}}_{\text{P,RJ};c}, \bar{B}^{\text{epis}}_{\text{P,RJ}})$ pair corresponding to each bin. Larger gaps between points are indicative of larger bins, which result from fewer experiments having B^{EMD} values close to $\bar{B}^{\text{EMD}}$.

Figure 9 is also a particularly clear illustration of the effect of the sensitivity factor c. When c is too small ($c \leq 2^{-9}$), R-distributions almost never overlap, and the $B^{\text{EMD}}_{\text{P,RJ};c}$ is mostly either 0 or 1: the criterion is *overconfident*, with many points located in the red regions. As c increases ($2^{-6} \leq c \leq 2^{3}$), $B^{\text{EMD}}_{\text{P,RJ};c}$ values become more evenly distributed over the entire $[0, 1]$ interval and either don't encroach at all into the overconfident regions, or do so sparingly. These curves show a nearly monotone relation between $B^{\text{EMD}}_{\text{P,RJ};c}$ and $B^{\text{epis}}_{\text{P,RJ};\Omega}$, indicating that the former is strongly predictive of the latter; the corresponding c values would likely be suitable for model comparison. In contrast, when c is too large ($c \geq 2^{6}$), B^{EMD} becomes decorrelated from B^{epis}: all points are near $B^{\text{epis}} \approx 0.5$, indicating that the $B^{\text{EMD}}_{\text{P,RJ};c}$ for such large values of c would not be a useful comparison criterion.

4.5.2 *Calibration for the neural response models*

For the neuron model, we consider four sources of experimental variability: variations in the distribution used to model observation noise ξ, variations in the strength (σ_o) of observation noise, as well as variations in the strength (σ_i) and correlation time (τ) of the external input I_{ext}. Each epistemic distribution Ω is therefore described by four distributions over hyperparameters:

Observation noise model One of *Gaussian* or *Cauchy*. The distributions are centered and σ_o is drawn from the distribution defined below. Note that this is the model used to *generate* a dataset $\mathcal{D}^{\text{rep}}_{\text{test}}$. The candidate models always evaluate their loss assuming a Gaussian observation model.

$$\texttt{Gaussian} \qquad p(\xi) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\xi^2}{2\sigma_o^2}\right) \qquad\qquad \texttt{Cauchy} \qquad p(\xi) = \frac{2}{\pi\sigma\left[1+\left(\frac{\xi^2}{\sigma_o/2}\right)\right]}$$

Observation noise strength σ_o

$$\texttt{Low noise}: \quad \log \sigma_o \sim \mathcal{N}(0.0\,\text{mV}, (0.5\,\text{mV})^2)$$
$$\texttt{High noise}: \quad \log \sigma_o \sim \mathcal{N}(1.0\,\text{mV}, (0.5\,\text{mV})^2)$$

External input strength σ_i The parameter σ_i sets the strength of the input noise such that $\langle I^2_{\text{ext}} \rangle = \sigma_i^2$.

$$\texttt{Weak input}: \quad \log \sigma_i \sim \mathcal{N}(-15.0\,\text{mV}, (0.5\,\text{mV})^2)$$
$$\texttt{Strong input}: \quad \log \sigma_i \sim \mathcal{N}(-10.0\,\text{mV}, (0.5\,\text{mV})^2)$$

External input correlation time τ The parameter τ sets the correlation time of the input noise such that $\left\langle I_{\text{ext}}(t) I_{\text{ext}}(t+s) \right\rangle = \sigma_i^2 e^{-s^2/2\tau^2}$.

$$\texttt{Short correlation}: \quad \log_{10} \tau \sim \text{Unif}([0.1\,\text{ms}, 0.2\,\text{ms}])$$
$$\texttt{Long correlation}: \quad \log_{10} \tau \sim \text{Unif}([1.0\,\text{ms}, 2.0\,\text{ms}])$$

We thus defined 2 statistical distributions for ξ, 2 distributions for σ_o, 2 distributions for τ, and 2 distributions for σ_i. Combined with three model pairs (see Fig. 6), this makes a total of $48 = 3 \times 2 \times 2 \times 2 \times 2$ possible epistemic distributions Ω, each of which can be identified by a tuple such as $(\mathcal{M}_A \text{ vs } \mathcal{M}_B, \texttt{Cauchy}, \texttt{Low noise}, \texttt{Strong input}, \texttt{Long correlation})$. Calibration results for each of these conditions are given in Supplementary Fig. 1.

During calibration against an epistemic distribution Ω, drawing a dataset $\mathcal{D}^{\text{rep}}_{\text{test}}$ happens in two steps. First, we randomly draw a vector ω of epistemic parameters (i.e. hyperparameters) from Ω; for example $\omega = (\underbrace{-0.27\text{mV}}_{\sigma_o}, \underbrace{47\text{ms}}_{\tau}, \underbrace{0.003\text{mV}}_{\sigma_i})$. Second, we use those parameters to generate the dataset $\mathcal{D}^{\text{rep}}_{\text{test}}$, composed in this case of data points $(t, \tilde{V}^{\text{LP}}(t))$. We can then evaluate the loss Q (given by equation (9)) on those data points. Note that the loss does not depend on the epistemic parameters ω directly, but in general will involve parameters which are fitted to the simulated data. In short, the vector ω describes the parameters of a simulated experiment, while $\mathcal{D}^{\text{rep}}_{\text{test}}$ describes a particular outcome of that experiment. In theory we could generate multiple datasets with the same vector ω, but in practice it is more statistically efficient to draw a new experiment for each new dataset.

When choosing epistemic distributions, it is worth remembering that the goal of calibration is to empirically approximate a probability over experimental conditions. Thus choosing a distribution which can generate a large number of conditions — ideally an infinite number — will lead to better estimates. Here the use of continuous distributions for σ_o, σ_i and τ, and

the fact that I_{ext} is a continuous process, helps us achieve this goal. Also important is to generate data where the model selection problem is ambiguous, so as to resolve calibration curves around $B^{\text{EMD}} = 0.5$.

This calibration is an imperfect procedure, and how well it works depends on the quality of the candidate models and the choice of loss function. Here for example, the choice of a pointwise loss makes it sensitive to the timing of spikes; this tends to favour models which produce fewer spikes, since the penalty on a mistimed spike is high. This is why in Fig. 6, in the $\mathcal{M}_A$ vs $\mathcal{M}_D$ comparison, we see a floor on the values of B^{epis}_{AD}. In short, some models have consistently lower loss even on random data, and so their risk — which is the expectation of their loss — is a priori lower. The bias we see in the $\mathcal{M}_A$ vs $\mathcal{M}_B$ comparison is likely due to a similar effect. (In this case because $\mathcal{M}_B$ is slightly better than $\mathcal{M}_A$ on average.)

4.6 The hierarchical beta process

In this work we identify the epistemic uncertainty of a model $\mathcal{M}_A$ with the variability of a stochastic process $\mathfrak{Q}_A$: realizations of $\mathfrak{Q}_A$ approximate the PPF of the model loss. In our Results we yperref[alg_HBP]. The rest of this section explains the theoretical justifications for each step of this generative algorithm for $\mathfrak{Q}_A$.

Algorithm 2 Hierarchical beta process

Given

- q_A^*, c and δ_A^{EMD}, ▷ computed from data
- $N \in \mathbb{N}^+$, ▷ number of refinements
- $p(\hat{q}(0), \hat{q}(1))$, ▷ 2-d distribution over end points

generate a discretized realization $\hat{q}_A$.

Procedure:

▷ Initialize the procedure by drawing end point
repeat
  draw $\big(\hat{q}(0), \hat{q}(1)\big) \sim p\big(\hat{q}(0), \hat{q}(1)\big)$
until $\hat{q}(0) < \hat{q}(1)$ ▷ PPFs must be increasing
▷ Successively refine the interval
for $n \in 1, 2, \dots, N$ **do** ▷ refinement levels
  for $\Phi \in \{k \cdot 2^{-n+1}\}_{k=0}^{2^{n-1}-1}$ **do** ▷ intermediate incrs.
    compute r, v according to equations (73)
    solve equations (75) to obtain α and β
    draw $x_1 \sim \text{Beta}(\alpha, \beta)$
    $\hat{q}(\Phi + 2^{-n}) \leftarrow \hat{q}(\Phi) + \Delta\hat{q}_{\Delta\Phi}(\Phi) \cdot x_1$
  end for
end for
return $\big(\hat{q}(\Phi) : \Phi \in \{\mathcal{I}_\Phi\}^{(N)}\big)$

The quantities r and v computed in Algorithm 2 conceptually represent the **r**atio between two sucessive increments and the **v**ariance of those increments.

4.6.1 *Relevant concepts of Wiener processes*

Before introducing the HB process, let us first review a few key properties which stochastic processes must satisfy and which are covered in most standard introductions.[16,55,28] We use for this the well-known Wiener process, and also introduce notation which will become useful when we define the HB process. Since our goal is to define a process for PPFs, we use Φ to denote the independent "domain" variable and restrict ourselves to 1-d processes for which $\Phi \in [0, 1]$.

For the Wiener process $\mathcal{W}$, each realization is a continuous function $W\colon [0,1] \to \mathbb{R}$. One way to approximate a realization of $\mathcal{W}$ is to first partition the interval into subintervals $[0, \Phi_1), [\Phi_1, \Phi_2), \dots, [\Phi_n, 1)$ with $0 < \Phi_1 < \Phi_2 < \cdots < \Phi_n < 1$; for simplicity we will only consider equal-sized subintervals, so that $\Phi_k = k\,\Delta\Phi$ for some $\Delta\Phi \in \mathbb{R}$. We then generate a sequence of independent random increments (one for each subinterval) $\big\{\Delta W_{\Delta\Phi}(0), \Delta W_{\Delta\Phi}(\Delta\Phi), \Delta W_{\Delta\Phi}(2\Delta\Phi), \dots, \Delta W_{\Delta\Phi}(1-\Delta\Phi)\big\}$ and define the corresponding realization as

$$W(k\,\Delta\Phi) = W(0) + \sum_{l=0}^{k-1} \Delta W_{\Delta\Phi}(l\,\Delta\Phi) \tag{58}$$

(Within each interval the function may be linearly interpolated, so that W is continuous.)

A *refinement* of a partition is obtained by taking each subinterval and further dividing it into smaller subintervals. For instance we can refine the unit interval $[0, 1)$ into a set of two subintervals, $[0, 2^{-1})$ and $[2^{-1}, 2^{-0})$. Let us denote these partitions $\{\mathcal{I}_\Phi\}^{(0)}$ and $\{\mathcal{I}_\Phi\}^{(1)}$ respectively. Repeating the process on each subinterval yields a sequence of ever finer refinements:

$$\begin{aligned}
\{\mathcal{I}_\Phi\}^{(0)} &\coloneqq \Big\{ \big[0,\ 2^{-0}\big) \Big\} \\
\{\mathcal{I}_\Phi\}^{(1)} &\coloneqq \Big\{ \big[0,\ 2^{-1}\big),\ \big[2^{-1},\ 2^{-0}\big) \Big\} \\
\{\mathcal{I}_\Phi\}^{(2)} &\coloneqq \Big\{ \big[0,\ 2^{-2}\big),\ \big[2^{-2},\ 2^{-1}\big),\ \big[2^{-1},\ 3\cdot 2^{-2}\big),\ \big[3\cdot 2^{-2},\ 2^{-0}\big) \Big\} \\
&\ \ \vdots \\
\{\mathcal{I}_\Phi\}^{(n)} &\coloneqq \Big\{ \big[k\cdot 2^{-n},\ (k+1)\cdot 2^{-n}\big) \Big\}_{k=0}^{2^n-1}.
\end{aligned} \tag{59}$$

With these definitions, for any $m \geq n$, $\{\mathcal{I}_\Phi\}^{(m)}$ is a refinement of $\{\mathcal{I}_\Phi\}^{(n)}$.

Later we will need to refer to the vector of new end points introduced at the n-th refinemement step. These are exactly the odd multiples of 2^{-n} between 0 and 1, which we denote $\{\Phi\}^{(n)}$:

$$\{\Phi\}^{(n)} \coloneqq \Big(2^{-n},\ 3\cdot 2^{-n},\ \dots,\ (2^n-3)\cdot 2^{-n},\ (2^n-1)\cdot 2^{-n}\Big). \tag{60}$$

Any random process must be *self-consistent*:[20] for small enough $\Delta\Phi$, the probability distribution at a point Φ must not depend on the level of refinement. For example, the Wiener process is defined such that the increments $\Delta W_{\Delta\Phi} \sim \mathcal{N}(0, \Delta\Phi)$ are independent; therefore

$$\begin{aligned}
&W(\Phi+2\Delta\Phi) \\
&= W(\Phi) + \Delta W_{2\Delta\Phi}(\Phi) = W(\Phi) + \Delta W_{\Delta\Phi}(\Phi) + \Delta W_{\Delta\Phi}(\Phi+\Delta\Phi) \\
&= W(\Phi) + \mathcal{N}(0, 2\Delta\Phi) = W(\Phi) + \mathcal{N}(0, \Delta\Phi) + \mathcal{N}(0, \Delta\Phi)\,.
\end{aligned} \tag{61}$$

It turns out that the combination of the Markovian and self-consistent properties set quite strong requirements on the

stochastic increments, since they impose the square root scaling of the Wiener increment: $O(\Delta W_{\Delta\Phi}) = O(\sqrt{\Delta\Phi})$. (§II.C of reference 20.)

While the Wiener process underlies much of stochastic theory, it is not suitable for defining a process $\mathfrak{Q}$ over PPFs. Indeed, it is not monotone by design, which violates one of our desiderata. Moreover, it has a built-in directionality in the form of accumulated increments. A clear symptom of this is that as Φ increases, the variance of $W(\Phi)$ also increases. (This follows immediately from equation (58) and the independence of increments.) Directionality makes sense if we think of W as modelling the diffusion of particles in space or time, but empirical PPFs are obtained by first sorting data samples according to their loss (see the definition of δ^{EMD} in the Results). Since samples of the loss arrive in no particular order, a process which samples PPFs should likewise have no intrinsic directionality in Φ.

4.6.2 *A hierarchical beta distribution is monotone, non-accumulating and self-consistent*

Constructing a stochastic process $\mathfrak{Q}$ which is *monotone* is relatively simple: one only needs to ensure that the random increments $\Delta\hat{q}_{\Delta\Phi}(\Phi)$ are non-negative.

Ensuring that those increments are *non-accumulating* requires more care, because that requirement invalidates most common definitions of stochastic processes. As described in our Results, we achieve this by defining $\mathfrak{Q}$ as a sequence of refinements, starting from a single increment for the entire interval, then doubling the number of increments (and halving their width) at each refinement step. In the rest of this subsection we give an explicit construction of this process and show that it is also *self-consistent*. (Altough in this work we consider only pairs of increments sampled from a beta distribution, in general one could consider other compositional distributions. Higher-dimensional distributions may allow to sample all increments simultaneously, if one can determine the conditions which ensure self-consistency.)

For an interval $\mathcal{I} = [\Phi, \Phi + \Delta\Phi)$, we suppose that the points $\hat{q}_A(\Phi)$ and $\hat{q}_A(\Phi + \Delta\Phi)$ are given. We define

$$\Delta\hat{q}(\mathcal{I}) \;=\; \Delta\hat{q}_{\Delta\Phi}(\Phi) \;\coloneqq\; \hat{q}_A(\Phi + \Delta\Phi) - \hat{q}_A(\Phi)\,, \tag{62}$$

then we draw a subincrement $\Delta\hat{q}_{\frac{\Delta\Phi}{2}}(\Phi)$, associated to the subinterval $[\Phi, \Phi + \Delta\Phi)$, from a scaled beta distribution:

$$\frac{1}{\Delta\hat{q}_{\Delta\Phi}(\Phi)}\Delta\hat{q}_{\frac{\Delta\Phi}{2}}(\Phi) \coloneqq x_1 \sim \mathrm{Beta}(\alpha, \beta)\,. \tag{63}$$

(Refer to the subsection below for the definition of $\mathrm{Beta}(\alpha, \beta)$.) The scaling is chosen so that

$$0 \le \underbrace{\Delta\hat{q}_{\frac{\Delta\Phi}{2}}(\Phi)}_{= x_1 \Delta\hat{q}_{\Delta\Phi}(\Phi)} \le \Delta\hat{q}_{\Delta\Phi}(\Phi)\,. \tag{64}$$

The value of $\Delta\hat{q}_{\frac{\Delta\Phi}{2}}(\Phi)$ then determines the intermediate point:

$$\hat{q}_A\big(\Phi + \tfrac{\Delta\Phi}{2}\big) \leftarrow \hat{q}_A(\Phi) + \Delta\hat{q}_{\frac{\Delta\Phi}{2}}(\Phi)\,. \tag{65}$$

If desired, the complementary increment $\Delta\hat{q}_{\frac{\Delta\Phi}{2}}\big(\Phi + \frac{\Delta\Phi}{2}\big)$ can be obtained as

$$\Delta\hat{q}_{\frac{\Delta\Phi}{2}}\big(\Phi + \tfrac{\Delta\Phi}{2}\big) = (1 - x_1)\Delta\hat{q}_{\Delta\Phi}(\Phi)\,. \tag{66}$$

Generalizing the notation to the entire $[0, 1)$ interval, we start from a sequence of increments associated to subintervals at refinement step n (recall equation (59)):

$$\begin{aligned}\{\Delta\hat{q}_{2^{-n}}\} &\coloneqq \big\{\Delta\hat{q}(\mathcal{I}) \quad : \quad \mathcal{I} \in \{\mathcal{I}_\Phi\}^{(n)}\big\} \\ &\overset{\text{def}}{=} \big\{\Delta\hat{q}_{2^{-n}}(\Phi) : \big[\Phi, \Phi + 2^{-n}\big) \in \{\mathcal{I}_\Phi\}^{(n)}\big\}\,.\end{aligned}$$

Applying the procedure just described, for each subinterval we draw x_1 and split the corresponding increment $\Delta\hat{q}_{2^{-n}}(\Phi) \in \{\Delta\hat{q}_{2^{-n}}\}$ into a pair $\big(\Delta\hat{q}_{2^{-n-1}}(\Phi), \Delta\hat{q}_{2^{-n-1}}(\Phi + 2^{-n-1})\big)$ of subincrements such that

$$\Delta\hat{q}_{2^{-n}}(\Phi) = \Delta\hat{q}_{2^{-n-1}}(\Phi) + \Delta\hat{q}_{2^{-n-1}}(\Phi + 2^{-n-1})\,. \tag{67}$$

The union of subincrements is then the next refinement step:

$$\big\{\Delta\hat{q}(\mathcal{I}) : \mathcal{I} \in \{\mathcal{I}_\Phi\}^{(n+1)}\big\} = \{\Delta\hat{q}_{2^{-n-1}}\}\,. \tag{68}$$

After n refinement steps, we thus obtain a function $\hat{q}(\Phi)$ defined at discrete points:

$$\hat{q}^{(n)}(k \cdot 2^{-n}) \coloneqq \hat{q}(0) + \sum_{l<k} \Delta\hat{q}_{2^{-n}}(l \cdot 2^{-n})\,, \tag{69}$$

which we extend to the entire interval $[0, 1)$ by linear interpolation; see Fig. 5**d** for an illustration. In practice we found that computations (specifically the risk computed by integrating $\hat{q}^{(n)}(\Phi)$) converge after about eight refinement steps.

This procedure has the important property that once a point is sampled, it does not change on further refinements:

$$\hat{q}^{(n)}(k \cdot 2^{-n}) = \hat{q}^{(m)}(k \cdot 2^{-n})\,, \quad \forall m \ge n\,, \tag{70}$$

which follows from equation (67). Recall now that, as stated above, a process is self-consistent if "for small enough $\Delta\Phi$, the probability distribution at a point Φ [does] not depend on the level of refinement". Since equation (70) clearly satisfies that requirement, we see that the process obtained after infinitely many refinement steps is indeed self-consistent. We thus define the *hierarchical beta (HB) process* $\mathfrak{Q}_A$ as

$$p(\hat{q} \mid \mathfrak{Q}_A) \coloneqq p\Big(\hat{q} = \lim_{n\to\infty} \hat{q}^{(n)} \;\Big|\; c, q^*_A, \delta^{\mathrm{EMD}}_A\Big)\,. \tag{71}$$

To complete the definition of $\mathfrak{Q}_A$, we need to specify how we choose the initial end points $\hat{q}_A(0)$ and $\hat{q}_A(1)$. In our implementation, they are drawn from normal distributions $\mathcal{N}\big(q^*(\Phi), \sqrt{c}\,\delta^{\mathrm{EMD}}_A(\Phi)^2\big)$ with $\Phi \in \{0, 1\}$, where again c is determined via our proposed calibration procedure; this is simple and convenient, but otherwise arbitrary. We also need to explain how we choose the beta parameters α and β, which is the topic of the next subsection.

4.6.3 *Choosing beta distribution parameters*

All HB processes are monotone, continuous and self-consistent, but within this class there is still a lot of flexibility:

since α and β are chosen independently for each subinterval, we can mold $\mathfrak{Q}_A$ into a wide variety of statistical shapes. We use this flexibility to satisfy the two remaining desiderata: a) that realisations $\hat{q}_A(\Phi)$ track $q_A^*(\Phi)$ over $\Phi \in [0,1]$; and b) that the variability of $\hat{q}_A(\Phi)$ be proportional to $\delta_A^{\mathrm{EMD}}(\Phi)$. It is the goal of this subsection to give a precise mathematical meaning to those requirements.

Let $x_1 \sim \mathrm{Beta}(\alpha, \beta)$ and $x_2 = 1 - x_1$. (The density function of a beta distribution is given in equation (24).) The mean and variance of x_1 are

$$\mathbb{E}[x_1] = \frac{\alpha}{\alpha + \beta}\,, \tag{72a}$$

$$\mathbb{V}[x_1] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}\,. \tag{72b}$$

For a given Φ, it may seem natural to select α and β by matching $\mathbb{E}[x_1]$ to $q_A^*(\Phi)$ and $\mathbb{V}[x_1]$ to $c\cdot\big(\delta_A^{\mathrm{EMD}}(\Phi)\big)^2$. However both equations are tightly coupled, and we found that numerical solutions were unstable and unsatisfactory; in particular, it is not possible to make the variance large when $\mathbb{E}[x_1]$ approaches either 0 or 1 (otherwise the distribution of x_1 would exceed $[0,1]$).

Much more practical is to consider moments with respect to the Aitchison measure; as mentioned in the Results, the Aitchison measure first maps the bounded interval $[0,1]$ to the unbounded space $\mathbb{R}$ with a logistic transformation, then computes moments in the unbounded space. The first two such moments are called the *centre* and the *metric variance*;[42,47] for the beta distribution, they are given by (reproduced from 25)

$$\mathbb{E}_a[(x_1,x_2)] = \frac{1}{e^{\psi(\alpha)} + e^{\psi(\beta)}}\big(e^{\psi(\alpha)}, e^{\psi(\beta)}\big)\,, \tag{25a}$$

$$\mathrm{Mvar}[(x_1,x_2)] = \frac{1}{2}\big(\psi_1(\alpha) + \psi_1(\beta)\big)\,, \tag{25b}$$

where ψ and ψ_1 are the digamma and trigamma functions respectively. The centre and metric variance are known to be more natural statistics for compositional distributions, and this is what we found in practice. Therefore we will relate q_A^* to the centre, and $\sqrt{c}\,\delta_A^{\mathrm{EMD}}$ to the square root of the metric variance.

To be precise, suppose that we have already selected a set of increments $\{\Delta\hat{q}_{2^{-n}}\}$ over the domain $\{k\cdot 2^{-n}\}_{k=0}^{n-1}$, and wish to produce the refinement $\{\Delta\hat{q}_{2^{-n-1}}\}$. For each Φ in $\{k\cdot 2^{-n}\}_{k=0}^{n-1}$ we define

$$r \coloneqq \frac{q_A^*(\Phi + 2^{-n-1}) - q_A^*(\Phi)}{q_A^*(\Phi + 2^{-n}) - q_A^*(\Phi + 2^{-n-1})}\,, \tag{73a}$$

$$v \coloneqq 2\,c\left(\delta_A^{\mathrm{EMD}}\big(\Phi + 2^{-n-1}\big)\right)^2. \tag{73b}$$

The value r is the ratio of subincrements of $\Delta q^*_{2^{-n}}(\Phi)$. Since we want $\hat{q}_A$ to track q_A^*, it makes sense to expect r to also approximate the ratio of subincrements of $\Delta\hat{q}_{2^{-n}}(\Phi)$. Identifying

$$r \leftrightarrow \frac{\mathbb{E}_a[x_1]}{\mathbb{E}_a[x_2]} \quad \text{and} \quad v \leftrightarrow 2\cdot\mathrm{Mvar}[(x_1,x_2)]\,, \tag{74}$$

and substituting into equations (25) then leads to the following system of equations:

$$\psi(\alpha) - \psi(\beta) = \ln r \tag{75a}$$

$$\ln\big[\psi_1(\alpha) + \psi_1(\beta)\big] = \ln v\,, \tag{75b}$$

which we can solve to yield the desired parameters α and β for each subinterval in $\{\mathcal{I}_\Phi\}^{(n)}$.

In summary, although the justification is somewhat technical, the actual procedure for obtaining α and β is quite simple: first compute r and v following equations (73), then solve equations (75).

Data availability

All figure data can be regenerated from code (see Code availability). Results of especially long computations are included as precomputed data files alongside the accompanying computational notebooks.[50]

Code availability

All source code used to produce the figures in this paper is available as a collection of Jupyter notebooks.[50] Additional Python code for simulating the neuron circuit model[51] and generating colored noise[53] is also available.

We also provide the Python package `emdcmp`,[52] available on the Python Packaging Index (PyPI), which automates many of the computation steps in our approach: given a set of observed data $\mathcal{D}_{\text{test}}$, a generative model $\mathcal{M}_A$, a loss function Q and sensitivity parameter $c \in \mathbb{R}$, `emdcmp` will automatically compute $\tilde{q}_A$, q_A^* and δ_A^{EMD}, then draw samples from the corresponding R_A-distribution. In other words, through `emdcmp` we provide a standard implementation for all the steps represented by downward facing arrows on the right of Fig. 1.

The package also provides utilities to help execute calibration experiments.

References

1. Akaike, H. *Information Theory and an Extension of the Maximum Likelihood Principle* in *2nd International Symposium on Information Theory* (eds Petrov, B.N. & Csaki, F.) (Akadémiai Kiadó, Budapest, Hungary, 1973), 267–281.
2. Akaike, H. & de Leeuw, J. in *Breakthroughs in Statistics: Foundations and Basic Theory* (eds Kotz, S. & Johnson, N. L.) 610–624 (Springer, New York, NY, 1992). ISBN: 978-1-4612-0919-5. doi: 10.1007/978-1-4612-0919-5.
3. Arendt, P. D., Apley, D. W. & Chen, W. Quantification of Model Uncertainty: Calibration, Model Discrepancy, and Identifiability. *Journal of Mechanical Design* **134.** ISSN: 1050-0472. doi: 10.1115/1.4007390 (Sept. 2012).

4. Baets, B. D. & Meyer, H. D. in *Perception-Based Data Mining and Decision Making in Economics and Finance* (eds Batyrshin, I., Kacprzyk, J., Sheremetov, L. & Zadeh, L. A.) 261–274 (Springer, Berlin, Heidelberg, 2007). ISBN: 978-3-540-36247-0. doi: 10.1007/978-3-540-36247-0_10.
5. Beven, K. & Binley, A. GLUE: 20 Years On. *Hydrological Processes* **28,** 5897–5918. ISSN: 0885-6087, 1099-1085. doi: 10.1002/hyp.10082 (Nov. 2014).
6. Beven, K. & Binley, A. The Future of Distributed Models: Model Calibration and Uncertainty Prediction. *Hydrological Processes* **6,** 279–298. ISSN: 1099-1085. doi: 10.1002/hyp.3360060305 (1992).
7. Beven, K. & Freer, J. Equifinality, Data Assimilation, and Uncertainty Estimation in Mechanistic Modelling of Complex Environmental Systems Using the GLUE Methodology. *Journal of Hydrology* **249,** 11–29. ISSN: 0022-1694. doi: 10.1016/S0022-1694(01)00421-8 (Aug. 2001).
8. Beven, K. & Smith, P. Concepts of Information Content and Likelihood in Parameter Calibration for Hydrological Simulation Models. *Journal of Hydrologic Engineering* **20,** A4014010. ISSN: 1943-5584. doi: 10.1061/(ASCE)HE.1943-5584.0000991 (Jan. 2015).
9. Chen, S., Mira, A. & Onnela, J.-P. Flexible Model Selection for Mechanistic Network Models. *Journal of Complex Networks* **8,** cnz024. ISSN: 2051-1329. doi: 10.1093/comnet/cnz024 (Apr. 2020).
10. Chmiela, S. *et al.* Machine Learning of Accurate Energy-Conserving Molecular Force Fields. *Science Advances* **3,** e1603015. doi: 10.1126/sciadv.1603015 (May 2017).
11. De Blasi, P. & Walker, S. G. Bayesian Asymptotics with Misspecified Models. *Statistica Sinica*. ISSN: 10170405. doi: 10.5705/ss.2010.239 (2013).
12. De Schuymer, B., De Meyer, H. & De Baets, B. Cycle-Transitive Comparison of Independent Random Variables. *Journal of Multivariate Analysis* **96,** 352–373. ISSN: 0047-259X. doi: 10.1016/j.jmva.2004.10.011 (Oct. 2005).
13. de Silva, B. M., Higdon, D. M., Brunton, S. L. & Kutz, J. N. Discovery of Physics From Data: Universal Laws and Discrepancies. *Frontiers in Artificial Intelligence* **3.** ISSN: 2624-8212 (2020).
14. Findley, D. F. Counterexamples to Parsimony and BIC. *Annals of the Institute of Statistical Mathematics* **43,** 505–514. ISSN: 1572-9052. doi: 10.1007/BF00053369 (Sept. 1991).
15. Gal, Y., Hron, J. & Kendall, A. *Concrete Dropout* in *Advances in Neural Information Processing Systems* (eds Guyon, I. *et al.*) **30** (Curran Associates, Inc., 2017).
16. Gardiner, C. W. *Handbook of Stochastic Methods: For Physics, Chemistry and the Natural Sciences ; with 29 Figures Springer Series in Synergetics* **13.** ISBN: 978-3-540-11357-7 (Springer, Berlin, 1983).
17. Gelman, A., Meng, X.-L. & Stern, H. Posterior Predictive Assessment of Model Fitness Via Realized Discrepancies. *Statistica Sinica* **6,** 733–760. ISSN: 1017-0405 (1996).
18. Gelman, A. & Yao, Y. Holes in Bayesian Statistics. *Journal of Physics G: Nuclear and Particle Physics* **48,** 014002. ISSN: 0954-3899, 1361-6471. doi: 10.1088/1361-6471/abc3a5. arXiv: 2002.06467 (Jan. 2021).
19. Gelman, A. *et al. Bayesian Data Analysis* ISBN: 978-1-4398-9822-2 (CRC Press, Boca Raton, 2014).
20. Gillespie, D. T. The Mathematics of Brownian Motion and Johnson Noise. *American Journal of Physics* **64,** 225–240. ISSN: 0002-9505, 1943-2909. doi: 10.1119/1.18210 (Mar. 1996).
21. Gneiting, T. & Raftery, A. E. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association* **102,** 359–378. ISSN: 0162-1459. JSTOR: 27639845 (2007).
22. Golden, R. M. Discrepancy Risk Model Selection Test Theory for Comparing Possibly Misspecified or Nonnested Models. *Psychometrika* **68,** 229–249. ISSN: 0033-3123, 1860-0980. doi: 10.1007/BF02294799 (June 2003).
23. Golden, R. M. Statistical Tests for Comparing Possibly Misspecified and Nonnested Models. *Journal of Mathematical Psychology* **44,** 153–170. ISSN: 0022-2496. doi: 10.1006/jmps.1999.1281 (Mar. 2000).
24. Gonçalves, P. J. *et al.* Training Deep Neural Density Estimators to Identify Mechanistic Models of Neural Dynamics. *eLife* **9** (eds Huguenard, J. R., O'Leary, T. & Goldman, M. S.) e56261. ISSN: 2050-084X. doi: 10.7554/eLife.56261 (Sept. 2020).
25. Grünwald, P. & Roos, T. Minimum Description Length Revisited. *International Journal of Mathematics for Industry*. doi: 10.1142/S2661335219300018 (Mar. 2020).
26. Grünwald, P. & van Ommen, T. Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It. *Bayesian Analysis* **12,** 1069–1103. ISSN: 1936-0975, 1931-6690. doi: 10.1214/17-BA1085 (Dec. 2017).
27. Gu, M., Fang, X. & Luo, Y. Data-Driven Model Construction for Anisotropic Dynamics of Active Matter. *PRX Life* **1,** 013009. doi: 10.1103/PRXLife.1.013009 (Aug. 2023).
28. Horsthemke, W. & Lefever, R. *Noise-Induced Transitions: Theory and Applications in Physics, Chemistry, and Biology* 2. print. *Springer Series in Synergetics* **15.** ISBN: 978-3-540-11359-1 (Springer, Berlin, 2006).
29. Hsu, H.-L., Ing, C.-K. & Tong, H. On Model Selection from a Finite Family of Possibly Misspecified Time Series Models. *The Annals of Statistics* **47.** ISSN: 0090-5364. doi: 10.1214/18-AOS1706 (Apr. 2019).
30. Hüllermeier, E. & Waegeman, W. Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods. *Machine Learning* **110,** 457–506. ISSN: 1573-0565. doi: 10.1007/s10994-021-05946-3 (Mar. 2021).
31. Kahle, L. & Zipoli, F. Quality of Uncertainty Estimates from Neural Network Potential Ensembles. *Physical Review E* **105,** 015311. doi: 10.1103/PhysRevE.105.015311 (Jan. 2022).

32. Kennedy, M. C. & O'Hagan, A. Bayesian Calibration of Computer Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63,** 425–464. ISSN: 1467-9868. doi: 10.1111/1467-9868.00294 (2001).
33. Kiureghian, A. D. & Ditlevsen, O. Aleatory or Epistemic? Does It Matter? *Structural Safety. Risk Acceptance and Risk Communication* **31,** 105–112. ISSN: 0167-4730. doi: 10.1016/j.strusafe.2008.06.020 (Mar. 2009).
34. Koposov, S. *et al. Joshspeagle/Dynesty: V2.1.4* Zenodo. June 2024. doi: 10.5281/zenodo.12537467.
35. Kumar, R., Carroll, C., Hartikainen, A. & Martin, O. ArviZ a Unified Library for Exploratory Analysis of Bayesian Models in Python. *Journal of Open Source Software* **4,** 1143. ISSN: 2475-9066. doi: 10.21105/joss.01143 (Jan. 2019).
36. Lewis, J. P. Algorithms for Solid Noise Synthesis. *SIGGRAPH Comput. Graph.* **23,** 263–270. ISSN: 0097-8930. doi: 10.1145/74334.74360 (July 1989).
37. List, F., Rodd, N. L., Lewis, G. F. & Bhat, I. Galactic Center Excess in a New Light: Disentangling the γ-Ray Sky with Bayesian Graph Convolutional Neural Networks. *Physical Review Letters* **125,** 241102. doi: 10.1103/PhysRevLett.125.241102 (Dec. 2020).
38. Luengo, D., Martino, L., Bugallo, M., Elvira, V. & Särkkä, S. A Survey of Monte Carlo Methods for Parameter Estimation. *EURASIP Journal on Advances in Signal Processing* **2020,** 25. ISSN: 1687-6180. doi: 10.1186/s13634-020-00675-6 (May 2020).
39. Lv, J. & Liu, J. S. Model Selection Principles in Misspecified Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **76,** 141–167. ISSN: 1369-7412. doi: 10.1111/rssb.12023 (Jan. 2014).
40. MacKay, D. J. C. *Information Theory, Inference, and Learning Algorithms* ISBN: 978-0-521-64298-9 (Cambridge University Press, Cambridge, UK ; New York, 2003).
41. Mateu-Figueras, G. & Pawlowsky-Glahn, V. *The Dirichlet Distribution with Respect to the Aitchison Measure on the Simplex - a First Approach* in *Proceedings of the 2nd International Workshop on Compositional Data Analysis* (eds Mateu-Figueras, G. & Barceló-Vidal, C.) (Girona, Oct. 2005). ISBN: 84-8458-222-1.
42. Mateu-Figueras, G., Monti, G. S. & Egozcue, J. J. in *Advances in Compositional Data Analysis: Festschrift in Honour of Vera Pawlowsky-Glahn* (eds Filzmoser, P., Hron, K., Martín-Fernández, J. A. & Palarea-Albaladejo, J.) 61–82 (Springer International Publishing, Cham, 2021). ISBN: 978-3-030-71175-7. doi: 10.1007/978-3-030-71175-7_4.
43. Modrák, M. *et al.* Simulation-Based Calibration Checking for Bayesian Computation: The Choice of Test Quantities Shapes Sensitivity. *Bayesian Analysis* **Advance publication,** 28. ISSN: 1936-0975, 1931-6690. doi: 10.1214/23-BA1404 (2023).
44. Moran, G. E., Cunningham, J. P. & Blei, D. M. The Posterior Predictive Null. *Bayesian Analysis* **18,** 1071–1097. ISSN: 1936-0975, 1931-6690. doi: 10.1214/22-BA1313 (Dec. 2023).
45. Nowotny, T., Szücs, A., Levi, R. & Selverston, A. I. Models Wagging the Dog: Are Circuits Constructed with Disparate Parameters? *Neural Computation* **19,** 1985–2003. ISSN: 0899-7667. doi: 10.1162/neco.2007.19.8.1985 (Aug. 2007).
46. Parry, M., Dawid, A. P. & Lauritzen, S. Proper Local Scoring Rules. *The Annals of Statistics* **40,** 561–592. ISSN: 0090-5364, 2168-8966. doi: 10.1214/12-AOS971 (Feb. 2012).
47. Pawlowsky-Glahn, V. & Egozcue, J. J. Geometric Approach to Statistical Analysis on the Simplex. *Stochastic Environmental Research and Risk Assessment* **15,** 384–398. ISSN: 1436-3259. doi: 10.1007/s004770100077 (Oct. 2001).
48. Prinz, A. A., Bucher, D. & Marder, E. Similar Network Activity from Disparate Circuit Parameters. *Nature Neuroscience* **7,** 1345–1352. ISSN: 1097-6256, 1546-1726. doi: 10.1038/nn1352 (Dec. 2004).
49. Prinz, A. A., Billimoria, C. P. & Marder, E. Alternative to Hand-Tuning Conductance-Based Models: Construction and Analysis of Databases of Model Neurons. *Journal of Neurophysiology* **90,** 3998–4015. ISSN: 0022-3077. doi: 10.1152/jn.00641.2003 (Dec. 2003).
50. René, A. *Code for Article "Selecting Fitted Models under Epistemic Uncertainty Using a Stochastic Process on Quantile Functions"* Zenodo. Mar. 2025. doi: 10.5281/zenodo.15032858.
51. René, A. *Efficient and Flexible Simulator of the Lobster Pyloric Circuit* Zenodo. July 2024. doi: 10.5281/zenodo.12797200.
52. René, A. *EMD Model Comparison Library* Zenodo. Mar. 2025. doi: 10.5281/zenodo.15033190.
53. René, A. *Solid Colored Noise Using Sparse Convolutions* Zenodo. July 2024. doi: 10.5281/zenodo.12805191.
54. René, A., Longtin, A. & Macke, J. H. Inference of a Mesoscopic Population Model from Population Spike Trains. *Neural Computation* **32,** 1448–1498. ISSN: 0899-7667. doi: 10.1162/neco_a_01292 (June 2020).
55. Risken, H. *The Fokker-Planck Equation: Methods of Solution and Applications* 2nd ed. *Springer Series in Synergetics* **v. 18.** ISBN: 978-0-387-50498-8 (Springer-Verlag, Berlin ; New York, 1989).
56. Rudiger, P. *et al. Holoviz/Holoviews: Version 1.18.1* Zenodo. Nov. 2023. doi: 10.5281/zenodo.10089811.
57. Schwarz, G. Estimating the Dimension of a Model. *The Annals of Statistics* **6,** 461–464. ISSN: 0090-5364. JSTOR: 2958889 (1978).
58. Sivula, T., Magnusson, M., Matamoros, A. A. & Vehtari, A. *Uncertainty in Bayesian Leave-One-Out Cross-Validation Based Model Comparison* Oct. 2023. doi: 10.48550/arXiv.2008.10296. arXiv: 2008.10296.
59. Skilling, J. Nested Sampling for General Bayesian Computation. *Bayesian analysis* **1,** 833–859 (2006).
60. Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Linde, A. The Deviance Information Criterion: 12 Years On. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **76,** 485–493. ISSN: 1369-7412. doi: 10.1111/rssb.12062 (June 2014).

61. Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Van Der Linde, A. Bayesian Measures of Model Complexity and Fit. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **64,** 583–639. ISSN: 1369-7412. doi: 10.1111/1467-9868.00353 (Oct. 2002).
62. Stedinger, J. R., Vogel, R. M., Lee, S. U. & Batchelder, R. Appraisal of the Generalized Likelihood Uncertainty Estimation (GLUE) Method. *Water Resources Research* **44.** ISSN: 1944-7973. doi: 10.1029/2008WR006822 (2008).
63. Swigon, D., Stanhope, S. R., Zenker, S. & Rubin, J. E. On the Importance of the Jacobian Determinant in Parameter Inference for Random Parameter and Random Measurement Error Models. *SIAM/ASA Journal on Uncertainty Quantification* **7,** 975–1006. doi: 10.1137/17M1114405 (Jan. 2019).
64. Talts, S., Betancourt, M., Simpson, D., Vehtari, A. & Gelman, A. Validating Bayesian Inference Algorithms with Simulation-Based Calibration. *arXiv:1804.06788 [stat]*. doi: 10.48550/arXiv.1804.06788. arXiv: 1804.06788 [stat] (Apr. 2018).
65. Tarantola, A. Popper, Bayes and the Inverse Problem. *Nature Physics* **2,** 492–494. ISSN: 1745-2481. doi: 10.1038/nphys375 (Aug. 2006).
66. Toni, T., Welch, D., Strelkowa, N., Ipsen, A. & Stumpf, M. P. Approximate Bayesian Computation Scheme for Parameter Inference and Model Selection in Dynamical Systems. *Journal of The Royal Society Interface* **6,** 187–202. doi: 10.1098/rsif.2008.0172 (July 2008).
67. Trotta, R. Bayes in the Sky: Bayesian Inference and Model Selection in Cosmology. *Contemporary Physics* **49,** 71–104. ISSN: 0010-7514. doi: 10.1080/00107510802066753 (Mar. 2008).
68. Van Bellegem, S. & Dahlhaus, R. Semiparametric Estimation by Model Selection for Locally Stationary Processes. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **68,** 721–746. ISSN: 1369-7412. doi: 10.1111/j.1467-9868.2006.00564.x (Nov. 2006).
69. Vapnik, V. *Principles of Risk Minimization for Learning Theory* in *Advances in Neural Information Processing Systems* (eds Moody, J., Hanson, S. & Lippmann, R.) **4** (Morgan-Kaufmann, 1991).
70. Vapnik, V. N. *The Nature of Statistical Learning Theory* ISBN: 978-1-4757-3264-1 (Springer New York, New York, NY, 2000).
71. Vehtari, A., Gelman, A. & Gabry, J. Practical Bayesian Model Evaluation Using Leave-One-out Cross-Validation and WAIC. *Statistics and Computing* **27,** 1413–1432. ISSN: 1573-1375. doi: 10.1007/s11222-016-9696-4 (Sept. 2017).
72. Watanabe, S. A Widely Applicable Bayesian Information Criterion. *Journal of Machine Learning Research* **14,** 867–897 (2013).
73. Xu, K. & Darve, E. Physics Constrained Learning for Data-Driven Inverse Modeling from Sparse Observations. *Journal of Computational Physics* **453,** 110938. ISSN: 0021-9991. doi: 10.1016/j.jcp.2021.110938 (Mar. 2022).
74. Yu, B. Stability. *Bernoulli* **19.** ISSN: 1350-7265. doi: 10.3150/13-BEJSP14 (Sept. 2013).

Acknowledgements

We thank Anno Kürth, Aitor Morales-Gregorio, Günther Palm, Moritz Helias, Jan Bölts, Abel Jansma, Thomas Nowotny and Manfred Opper for helpful comments and discussions.

This work was partly supported by the German Federal Ministry for Education and Research (BMBF grant no. 01IS19077B; AR), the Natural Sciences and Engineering Research Council of Canada (NSERC; AL), the government of Ontario (OGS; AR), and the European Union (ERC grant "HIGH-HOPeS", no. 101039827; AR).

Author contributions

A.R developed the theory, wrote the software implementations and the first draft of the manuscript. A.R and A.L. discussed the results and revised the manuscript.

Competing interests:

The authors declare no competing interests.

Additional information

An online version of this article is available at the following URL: https://alcrene.github.io/emd-paper

Correspondence should be addressed to Alexandre René.

Selecting fitted models under epistemic uncertainty using a stochastic process on quantile functions

– Supplementary Information –

Supplementary Methods

Transitivity of B^{EMD} comparisons

Given three independent continuous random variables R_A, R_B and R_C, define the probabilities

$$\underbrace{P(R_A < R_B)}_{=:B_{AB}} \qquad \underbrace{P(R_B < R_C)}_{=:B_{BC}} \qquad \underbrace{P(R_C < R_A)}_{=:B_{CA}} .$$

For some threshold ϵ, we would like these to satisfy a transitivity relation of the form

$$\left.\begin{aligned} B_{AB} > \epsilon \\ B_{BC} > \epsilon \end{aligned}\right\} \Rightarrow B_{CA} < \epsilon \,, \tag{S1}$$

as this would reduce the required number of pairwise comparisons between models (see section Model discrepancy as a baseline for non-stationary replications, Table 1 and equation (17) in the main text).

It is known that equation (S1) does not hold for $\epsilon = \frac{1}{2}$; classic counterexamples in this case are non-transitive dice.[2] However, the set of probabilities $\mathcal{S} \coloneqq \{B_{AB}, B_{BC}, B_{CA}\}$ *does* satisfy a property known as *dice-transitivity*,[4] from which one can derive that equation (S1) holds when $\epsilon = \varphi^{-1}$, where φ is the golden ratio. This result appears as a comment below Theorem 3 in Baets & Meyer,[1] but to our knowledge has otherwise remained unknown. We provide a short self-contained derivation below, for the convenience of the reader.

The definition of dice-transitivity is obtained by substituting equation (9) of De Schuymer *et al.*[4] into equation (6) of the same reference. For our purposes we are interested in the resulting upper bound

$$\alpha - 1 \leq -\beta\gamma \,, \tag{S2}$$

where α, β, and γ are respectively the lowest, middle and highest value of $\mathcal{S}$. In other words, $\{\alpha, \beta, \gamma\} = \mathcal{S}$ and

$$\alpha \leq \beta \leq \gamma \,. \tag{S3}$$

Suppose that, as given in equation (17), we have

$$B_{AB} > \varphi^{-1} \quad \text{and} \quad B_{BC} > \varphi^{-1} \,. \tag{S4}$$

We wish to use equation (S2) to establish an upper bound on B_{CA}. We do not know a priori how the probabilities are ordered, so we consider the six possible cases:

B_{AB}	B_{BC}	B_{CA}
α	β	γ
β	α	γ
α	γ	β
γ	α	β
β	γ	α
γ	β	α

(S5)

Cases $\alpha\beta\gamma$ and $\beta\alpha\gamma$. The assumptions of equation (S4) translate to $\alpha > \varphi^{-1}$ and $\beta > \varphi^{-1}$. We seek a bound on γ. Rearranging equation (S2), we then have

$$\gamma \leq \frac{1-\alpha}{\beta} < \frac{1-\varphi^{-1}}{\varphi^{-1}} = \frac{-1+\sqrt{5}}{2} = \varphi^{-1} \,, \tag{S6}$$

which contradicts equation (S3).

Cases $\alpha\gamma\beta$ and $\gamma\alpha\beta$. The argument is exactly analogous, except that we seek a bound on β. We get

$$\beta \leq \frac{1-\alpha}{\gamma} < \frac{1-\varphi^{-1}}{\varphi^{-1}} = \varphi^{-1} \,, \tag{S7}$$

which again contradicts equation (S3).

Therefore the only possible cases are $\beta\gamma\alpha$ and $\gamma\beta\alpha$, which means that B_{CA} must be the smallest of the three probabilities. These two final cases provide the upper bound on B_{CA}:

Cases $\beta\gamma\alpha$ and $\gamma\beta\alpha$. We seek a bound on α. Rearranging equation (S2) one more time yields

$$\alpha \leq 1 - \beta\gamma < 1 - \varphi^{-2} = \varphi^{-1} \,. \tag{S8}$$

Thus equation (S1) holds for $\epsilon = \varphi^{-1}$. More generally, we have $\alpha = B_{CA}$ whenever both B_{AB} and B_{BC} are greater than φ^{-1}. In this case equation (S8) implies

$$B_{AC} = 1 - B_{CA} > 1 - (1 - B_{AB}B_{BC}) = B_{AB}B_{BC} \,, \tag{S9}$$

and therefore

$$\left.\begin{aligned} B_{AB} > \varphi^{-1} \\ B_{BC} > \varphi^{-1} \end{aligned}\right\} \Rightarrow B_{AC} > B_{AB}B_{BC} \,. \tag{S10}$$

Equation (17) given in the main text is a special case of equation (S10).

Supplementary Results

Additional calibration curves for the neuron model

In Calibrating and validating the B^{EMD}, we considered only the effect of varying the external input strength (σ_i) on the calibration curves for the Prinz model. Fig. 1 is an extended version of Fig. 6, where all 48 proposed epistemic distributions are plotted.

Sensitivity factor c can shift R-distributions

The primary effect of the sensitivity parameter c in equations (16) and (27) is to increase the spread of R-distributions, and thus their overlap. However, in contrast to the case where we simply add additive noise (e.g. the noise η in equation (S26)), the parameter c can also affect the shape of R-distributions, and in particular can change the relative distance between their centres. This is illustrated in Supplementary Fig. 2.

This phenomenon is one reason why the calibration curves actually change shape when we change c (c.f. Supplementary Fig. 1 and Figure 4), and therefore that it is worthwhile to pin down an appropriate value for c. It is best understood as an effect of the additional constraints on $\mathfrak{Q}$: that realisations be monotone, integrable, and non-accumulating. These can change the shape of the $\mathfrak{Q}$ distribution, beyond the simple linear spread around q^* defined by equation (27).

This underscores also the importance of considering those constraints to correctly sample the space of experimental variations.

Comparison of selection criteria with inconclusive data

A key motivation underlying the B^{EMD} was to define a selection criterion which provides reliable estimates of its uncertainty, in particular for large test datasets (thousands of samples or more): these are commonly produced by experiments and yet produce pathologies with many of the common statistical methods of model selection. In short, a criterion can both over and underestimate the strength of the evidence in favour of one model.

We illustrate this phenomenon in Supplementary Table 1, with a set of comparisons meant to complement those of Fig. 8 in the main text and which explore the effect of misspecification in greater detail. We use 18 different variations of data generated with equation (38): three levels of noise s (ranging from high, 2^{12}, to low 2^{20} [$\mathrm{m}^2 \cdot \mathrm{nm} \cdot \mathrm{photons} \cdot \mathrm{sr} \cdot \mathrm{kW}^{-1}$]), three dataset sizes L (2^9, 2^{12}, and 2^{15}) and two wavelength ranges (20–1000 μm and 15–30 μm). Datasets with high noise and long wavelengths provide almost no discriminatory information: the two model predictions are almost the same, and any discrepancy between them is dwarfed by the amount of noise. At the other end, the low noise, short wavelength datasets are at the threshold between moderate and strong evidence in favour of $\mathcal{M}_{\mathrm{P}}$ — in this regime the better model can already be identified by eye.

If Fig. 8 in the main text compares criteria on their terms, on a type of problem they were designed to tackle, Supplementary Table 1 explores what happens in a more challenging situation with misspecified models and an ambiguous selection problem. As with Fig. 8, all comparisons are done on the basis of a single test dataset, so replication uncertainty does not play a role.

In addition to the B^{EMD} and the risk (B^R), Supplementary Table 1 includes criteria based on the *likelihood ratio* (B^l), the *Bayes factor* (B^{Bayes})[5,7] and the *expected log predictive density* (elpd) (B^{elpd}).[8,5] We note that because $\mathcal{M}_{\mathrm{P}}$ and $\mathcal{M}_{\mathrm{RJ}}$ have the same number of parameters, the likelihood ratio is also equivalent to the *Akaike Information Criterion* (AIC), up to a factor of 2.[5]

To allow for comparability, for each criterion we report the log probability ratio; i.e. "model A is B^C_{AB} times more probable than model B." Conceptually, if $P(A)$ is the "probability of model A" and $P(B)$ the "probability of model B", then a criterion corresponds to

$$B^C_{AB} = \frac{P(A)}{P(B)} \quad \leftrightarrow \quad \log_{10} B^C_{AB} = \log_{10} P(A) - \log_{10} P(B)\,. \tag{S11}$$

This is the typical form for Bayes' factors, and we use Jeffreys' scale[6,7] to interpret values in Supplementary Table 1: values near 0, ±1 and ±2 respectively correspond to inconclusive, weak, and strong evidence.

Although only Bayes factors are typically reported in this form, most information criteria are defined in terms of log probabilities; the difference between the criteria of two models (the sign of which determines which model is selected) can be identified with the difference $\log_{10} P(A) - \log_{10} P(B)$ in equation (S11).

The main exception is the EMD criterion, which is defined in terms of a probability $P(R_A < R_B)$ rather than separate probabilities for A and B. To express it as a probability ratio, we define the "underbar" quantity, obtained by applying a logit transformation to $B^{\mathrm{EMD}}_{AB;c}$:

$$\log_{10} \underline{B}^{\mathrm{EMD}}_{AB;c} \coloneqq \log_{10} B^{\mathrm{EMD}}_{AB;c} - \log_{10}\left(1 - B^{\mathrm{EMD}}_{AB;c}\right). \tag{S12}$$

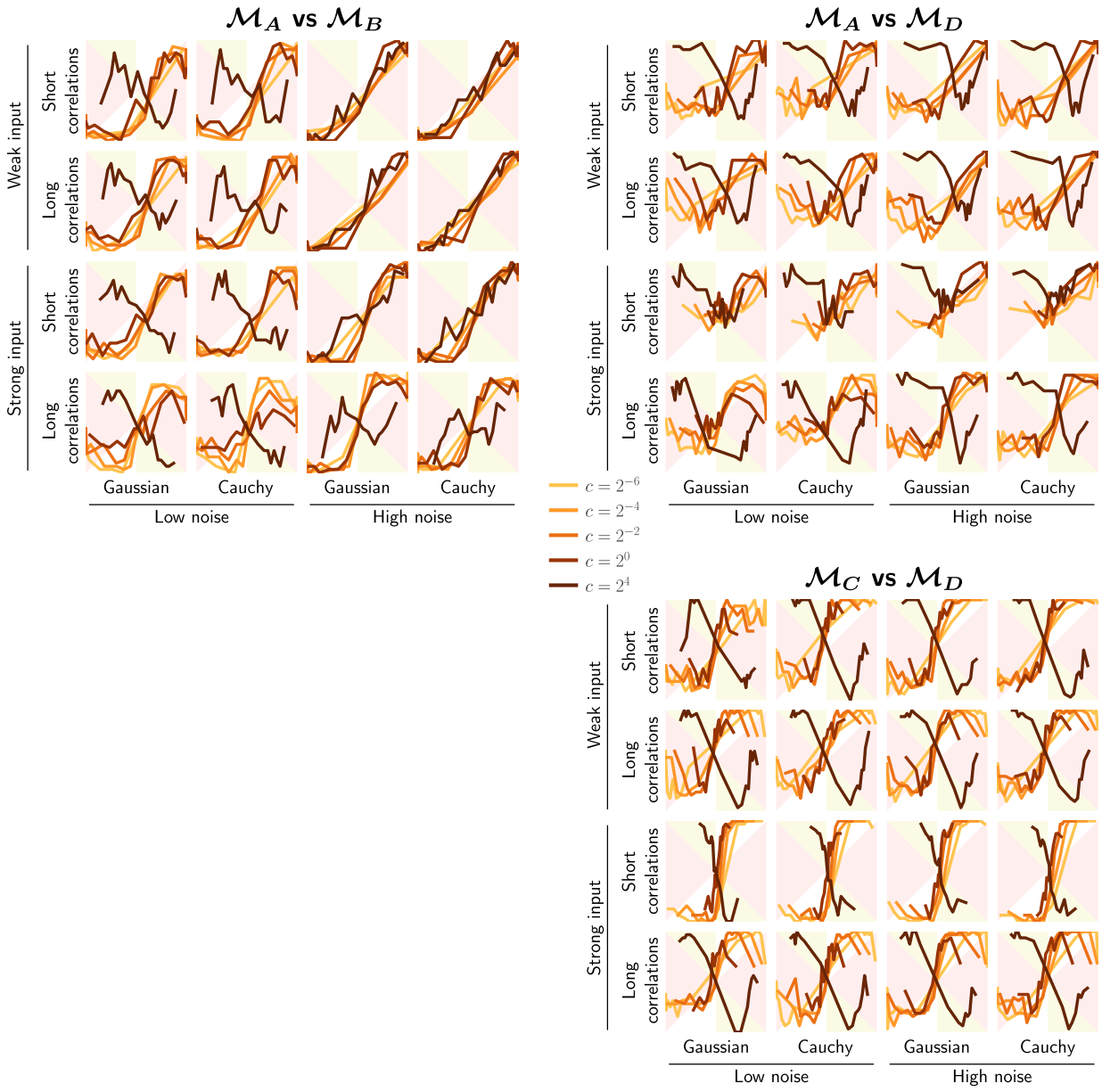
Values of $\underline{B}^{\mathrm{EMD}}_{AB;c}$ are therefore not directly comparable with those of $B^{\mathrm{EMD}}_{AB;c}$ reported elsewhere in this paper. In Supplementary Table 1, the $\log_{10} \underline{B}^{\mathrm{EMD}}_{AB;c}$ ranges from $-\infty$ to ∞, with values $\pm\infty$ indicating that the two R-distributions have zero overlap.

As a comparison, we also define $\underline{B}^R$, to show what happens if we simply compare the true risk:

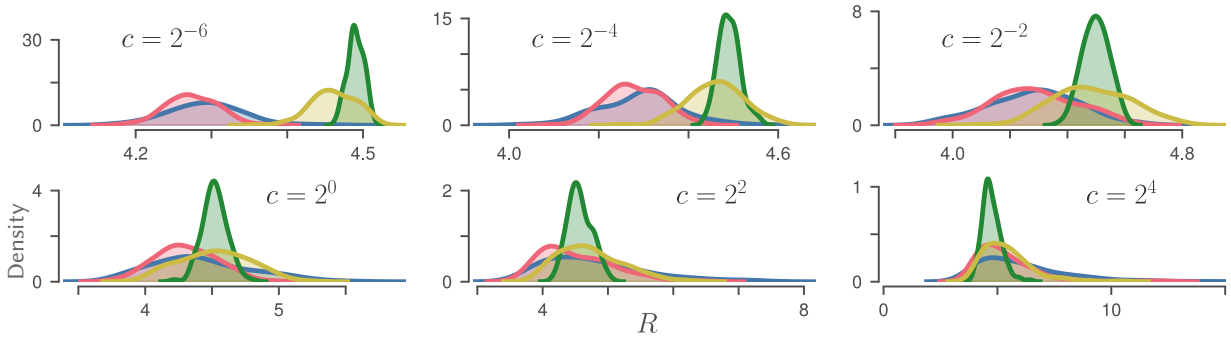
$$\log_{10} \underline{B}^R_{AB} \coloneqq (-R_A + R_B) \;/\; \log 10\,. \tag{S13}$$

The division by $\log 10$ converts the basis of the logarithms from e to 10.

To interpret the results summarised in Supplementary Table 1, we highlight two types of undesirable behaviour:



Supplementary Figure 1: **Additional calibration curves for the neuron model**, computed following the procedure described in section Calibrating and validating the B^{EMD} of the main text. Each curve summarizes 512 simulated experiments with datasets of size 4000. The regions depicted in red and yellow are those where equation (34) is violated.



Supplementary Figure 2: R-distributions for the four candidate neuron models for different values of c.

Lack of saturation There should always come a point where enough data have been collected, and simply enlarging the dataset with samples from the same distribution does not provide more information. In the table therefore we want values to converge to a finite value as L increases.

Non-robustness The magnitude of a criterion should be an indicator of its robustness. If small changes in a dataset cause a criterion to flip from strongly negative to strongly positive, this suggests that its magnitude is not a good indication for the strength of the evidence.

If we are to interpret the values of criteria as relative probabilities, the global pattern we would expect to see in Supplementary Table 1 is that cells become progressively more blue as we go to the right (less noise, less model mismatch) and as we go down (more data, less ambiguous data).

Expressions used to compute selection criteria Conventions:

- $\mathcal{D}$: Training dataset used to fit the model.
- L: Number of data samples in $\mathcal{D}$.
- $(\lambda_i, \mathcal{B}_i)$: Sample in $\mathcal{D}$.
- $(\lambda'_j, \mathcal{B}'_j)$: Additional sample not in $\mathcal{D}$ (i.e. a test sample).
- $\hat{\sigma}, \hat{T}$: Estimates of σ and T obtained by maximizing the likelihood on $\mathcal{D}$.

EMD criterion

$$\log_{10} \underline{B}^{\text{EMD}}_{AB;c} \coloneqq \log_{10} B^{\text{EMD}}_{AB;c} - \log_{10}\bigl(1 - B^{\text{EMD}}_{AB;c}\bigr)\,,$$

where $A = \mathcal{M}_{\text{P}}$ and $B = \mathcal{M}_{\text{RJ}}$ (repeated from equation (S12)).

Loss function As we did for the neuron model, we define the loss of a point $(\lambda_i, \mathcal{B}_i)$ as the negative log likelihood given that data point. Since the candidate models assume Gaussian noise (equation (39)), this is simply

$$\begin{aligned} Q_a(\mathcal{B}_i \mid \lambda_i, \sigma, T) &= -\log p\Bigl(\mathcal{B} \;\Big|\; \mathcal{B}_a(\lambda; T), \sigma\Bigr) \\ &= \log\sqrt{2\pi}\sigma + \frac{(\mathcal{B} - \mathcal{B}_a(\lambda;T))^2}{2\sigma^2}\,. \end{aligned} \tag{S14}$$

Here the subscript a is used to indicate whether we use predictions from the Rayleigh-Jeans (equation (36)) or Planck (equation (37)) model.

Log likelihood function Since we defined the loss as the negative log likelihood, one way to express the log likelihood for the whole dataset is simply as the negative total loss:

$$\ell_a(\sigma, T) \coloneqq \sum_{i=1}^{L} -Q_a(\mathcal{B}_i \mid \lambda_i, \sigma, T)\,. \tag{S15}$$

Risk The risk is the expectation of the loss under the true model, so we have

$$R_a \coloneqq \Bigl\langle Q_a\Bigl(\mathcal{B}'_j \mid \lambda'_j, \hat{\sigma}_a, \hat{T}_a\Bigr)\Bigr\rangle_{\lambda'_j, \mathcal{B}'_j \sim \mathcal{M}_{\text{true}}}\,, \tag{S16}$$

$$\log_{10} \underline{B}^{R} \coloneqq \bigl(-R_{\text{P}} + R_{\text{RJ}}\bigr) \;/\; \log 10\,, \quad \text{(repeated from (S13))}$$

where $\lambda'_j, \mathcal{B}'_j \sim \mathcal{M}_{\text{true}}$ indicates that λ'_j and $\mathcal{B}'_j$ are random variables with the same probability as the data.

Since in this case we know $\mathcal{M}_{\text{true}}$, the expectation can be computed exactly. We did this by generating a very large number of data samples ($L = 2^{12}$) and computing the empirical average of the loss:

$$R_a \approx -\frac{1}{L}\ell_a(\hat{\sigma}_a, \hat{T}_a)\,. \tag{S17}$$

Bayesian prior The Bayesian calculations require a prior on the parameters σ and T. We used a simple prior which factorizes into two independent distributions:

$$\begin{aligned} \pi\Bigl(\log_2 \sigma\Bigr) &\sim \text{Unif}\Bigl(\Bigl[\log_2 2^9, \log_2 2^{14}\Bigr]\Bigr)\,, \\ \pi\Bigl(\log_2 T\Bigr) &\sim \text{Unif}\Bigl(\Bigl[\log_2 1000, \log_2 5000\Bigr]\Bigr)\,. \end{aligned} \tag{S18}$$

Here T is expressed in Kelvin and σ has units $\text{kW} \cdot \text{m}^{-2} \cdot \text{nm}^{-1} \cdot \text{sr}^{-1}$. We chose log uniform distributions because these are more appropriate for parameters which are strictly positive and which can span multiple scales: the logarithmic scaling captures the fact that the difference between 1000 K and 1001 K is more significant than the difference between 5000 K and 5001 K. Likewise for differences in the parameter σ at opposite ends of its range.

Expected log pointwise posterior predictive density (elpd)

$$\text{elpd}_a \coloneqq \Bigl\langle \log_{10} p(\lambda'_j, \mathcal{B}'_j \mid a, \mathcal{D})\Bigr\rangle_{\lambda'_j, \mathcal{B}'_j \sim \mathcal{M}_{\text{true}}} \tag{S19}$$

$$\log_{10} B^{\text{elpd}} \coloneqq \text{elpd}_{\text{P}} - \text{elpd}_{\text{RL}}\,. \tag{S20}$$

Note p in equation (S19) is a posterior density, and therefore evaluating it involves integrating over the prior:

$$\begin{aligned} &p(\lambda'_j, \mathcal{B}'_j \mid a, \mathcal{D}) \\ &= \iint d\sigma dT\; \pi_\sigma(\sigma)\pi_T(T)\; p(\lambda'_j, \mathcal{B}'_j \mid a, \mathcal{D}, \sigma, T)\,. \end{aligned} \tag{S21}$$

Relative likelihood and AIC The relative likelihood is given by

$$\log_{10} B^{l} \coloneqq \bigl(\ell_{\text{P}}(\hat{\sigma}_{\text{P}}, \hat{T}_{\text{P}}) - \ell_{\text{RL}}(\hat{\sigma}_{\text{P}}, \hat{T}_{\text{P}})\bigr) \;/\; \log 10\,, \tag{S22}$$

while the difference between the AIC criteria of both models is (we use here the fact that both models have the same number of parameters)

$$\Delta AIC \coloneqq 2\ell_{\text{P}}(\hat{\sigma}_{\text{P}}, \hat{T}_{\text{P}}) - 2\ell_{\text{RL}}(\hat{\sigma}_{\text{P}}, \hat{T}_{\text{P}})\,. \tag{S23}$$

Supplementary Table 1: **Comparison of different model selection criteria** for variations of the datasets shown in Fig. 7. Criteria compare the Planck model ($\mathcal{M}_\mathrm{P}$) against the Rayleigh-Jeans model ($\mathcal{M}_\mathrm{RJ}$) and are evaluated for different dataset sizes (L), different levels of noise (s) and different wavelength windows (λ). To allow for comparisons, all criteria have been transformed into ratios of probabilities. **Values reported are the** $\log_{10}$ **of those ratios.** Positive (blue shaded) values indicate evidence in favour of the Planck model, while negative (red shaded) values indicate the converse. For example, a value of +1 is interpreted as $\mathcal{M}_\mathrm{P}$ being 10 times more likely than $\mathcal{M}_\mathrm{RJ}$, while a value of -2 would suggest that $\mathcal{M}_\mathrm{P}$ is 100 times *less* likely than $\mathcal{M}_\mathrm{RJ}$. Noise levels are reported in two ways: s is the actual value used to generate the data, while "rel. σ" reports the resulting standard deviation as a fraction of the maximum radiance within the data window. The 20–1000 µm window for λ stretches into the far infrared range, where the two models are nearly indistinguishable; hence higher positive values are expected for zero bias, low noise, and the 15–30 µm window. As in Fig. 7, calculations were done for both positive and null bias conditions (resp. ($\mathcal{B}_0 > 0$ and $\mathcal{B}_0 = 0$); the former emulates a situation where neither model can fit the data perfectly. Expressions for all criteria are given in the supplementary text. For the $\underline{B}^{\mathrm{EMD}}_{\mathrm{P,RJ}}$ criteria, we used $c = 0.5$. Although this is the same value as we used for neuron model, it was determined through a separate calibration with different epistemic distributions.

			$\mathcal{B}_0 > 0$			$\mathcal{B}_0 = 0$		
		s	2^{12}	2^{16}	2^{20}	2^{12}	2^{16}	2^{20}
		rel. σ	35%	9%	2%	36%	9%	2%
Criterion	$\lambda(\mu\mathrm{m})$	L						
$B^{I}_{\mathrm{P,RJ}}$	20–1000	512	−0.72	−0.74	5.64	−0.72	−0.82	4.62
		4096	0.64	−0.41	14.49	0.62	−1.05	7.94
		32768	−3.02	10.33	40.41	−3.23	8.11	−8.50
	15–30	512	0.06	−0.35	3.27	0.06	−0.39	2.61
		4096	0.52	3.86	52.20	0.49	3.53	47.12
		32768	3.04	23.02	389.62	2.87	20.33	348.85
$B^{\mathrm{Bayes}}_{\mathrm{P,RJ};\pi}$	20–1000	512	0.02	−0.01	0.02	−0.06	0.04	-6×10^{-3}
		4096	0.06	−0.03	−0.02	0.04	−0.03	0.07
		32768	−0.05	−0.05	0.04	0.03	−0.03	−0.02
	15–30	512	-6×10^{-3}	−0.03	0.01	−0.02	-7×10^{-3}	−0.05
		4096	-4×10^{-3}	-5×10^{-3}	0.04	−0.04	−0.06	−0.01
		32768	0.02	0.06	−0.05	9×10^{-3}	0.05	−0.01
$B^{\mathrm{elpd}}_{\mathrm{P,RJ};\pi}$	20–1000	512	-4×10^{-3}	3×10^{-5}	3×10^{-4}	2×10^{-4}	2×10^{-3}	5×10^{-4}
		4096	-9×10^{-4}	1×10^{-3}	-4×10^{-3}	-3×10^{-3}	-3×10^{-3}	7×10^{-4}
		32768	2×10^{-4}	2×10^{-3}	-2×10^{-3}	1×10^{-5}	-7×10^{-5}	-2×10^{-3}
	15–30	512	-4×10^{-3}	-2×10^{-3}	-2×10^{-3}	6×10^{-4}	2×10^{-3}	-4×10^{-4}
		4096	-3×10^{-3}	1×10^{-3}	6×10^{-4}	6×10^{-4}	-8×10^{-4}	-2×10^{-3}
		32768	6×10^{-4}	-2×10^{-4}	-1×10^{-3}	-2×10^{-3}	-2×10^{-3}	-1×10^{-3}
$\underline{B}^{R}_{\mathrm{P,RJ}}$	20–1000	512	2×10^{-5}	-5×10^{-4}	5×10^{-3}	2×10^{-5}	-7×10^{-4}	4×10^{-3}
		4096	2×10^{-5}	-5×10^{-4}	5×10^{-3}	2×10^{-5}	-7×10^{-4}	4×10^{-3}
		32768	2×10^{-5}	-5×10^{-4}	5×10^{-3}	2×10^{-5}	-7×10^{-4}	4×10^{-3}
	15–30	512	8×10^{-5}	6×10^{-4}	0.01	8×10^{-5}	6×10^{-4}	9×10^{-3}
		4096	8×10^{-5}	6×10^{-4}	0.01	8×10^{-5}	6×10^{-4}	9×10^{-3}
		32768	8×10^{-5}	6×10^{-4}	0.01	8×10^{-5}	6×10^{-4}	9×10^{-3}
$\underline{B}^{\mathrm{EMD}}_{\mathrm{P,RJ}}$	20–1000	512	−0.04	0.11	0.33	−0.07	0.12	0.41
		4096	−0.01	0.07	0.34	−0.01	0.07	0.39
		32768	-1×10^{-3}	−0.04	0.26	-1×10^{-2}	−0.05	0.29
	15–30	512	−0.01	0.31	3.08	0.05	0.34	3.50
		4096	−0.03	0.40	1.87	−0.02	0.39	1.80
		32768	−0.05	0.24	1.74	−0.04	0.25	1.66

Since the two are equivalent up to a factor $2 \log 10$, the trends we see with the likelihood ratio therefore also occur with the AIC.

(The factor 2 in the AIC is meant to allow it to be interpreted — under certain assumptions — as a draw from a χ^2 distribution. This correspondence however is not required when interpreting Supplementary Table 1.)

Model evidence The model evidence is used to compute the Bayes factors. It is the expectation of the likelihood of the data — $\mathcal{D} = \{\lambda_i, \mathcal{B}_i\}_{i=1}^L$ — under the prior for T and σ:

$$\begin{aligned}\mathcal{E}_a &\coloneqq \iint d\sigma dT\, \pi_\sigma(\sigma)\pi_T(T)\; p\Big(\{\lambda_i, \mathcal{B}_i\}_{i=1}^L \mid a, \sigma, T\Big) \\ &= \iint d\sigma dT\, \pi_\sigma(\sigma)\pi_T(T)\ell_a(\sigma, T)\ . \end{aligned} \tag{S24}$$

The likelihood $p\Big(\{\lambda_i, \mathcal{B}_i\}_{i=1}^L \mid a, \sigma, T\Big)$ is given by the Gaussian observation model.

Bayes factor

$$\log_{10} B^B \coloneqq \log_{10} \mathcal{E}_\mathrm{P} - \log_{10} \mathcal{E}_\mathrm{RL}\,. \tag{S25}$$

Supplementary Discussion

Other forms of uncertainty

As we say in the main text, there are two main sources of epistemic uncertainty on an estimate of the risk: limited number of samples and variability in the replication process. This work focusses on the estimating replication uncertainty, and numerically studying its effect as a function of sample size. We have eschewed a formal treatment of sample size effects, since this is a well-studied problem and good estimation procedures already exist.

For example, a bootstrap procedure can be used to estimate the uncertainty on a statistic (here the risk R) from a single dataset $\mathcal{D}$, by recomputing the statistic on multiple surrogate datasets obtained by resampling $\mathcal{D}$ with replacement (Supplementary Fig. 3**b**). Alternatively, if we have access to good candidate models, we can use those models as simulators to generate multiple synthetic datasets (Supplementary Fig. 3**c**). The distribution of risks over those datasets is then a direct estimate of its uncertainty due to finite samples.

In the limit of infinite data, both of these methods produce a risk "distribution" which collapses onto a precise value, independent of any discrepancy between model and true data-generating process. In contrast, that discrepancy defines the spread of risk distributions in Fig. 3, which do not collapse when $L \to \infty$.

Aleatoric uncertainty also does not vanish in the large L limit, but manifests differently. It sets a lower bound on the spread of pointwise losses a model can achieve (Supplementary Fig. 3**a**). In terms of our formalism therefore:

- **aleatoric uncertainty** determines the shape of the PPFs q^* and $\tilde{q}$,
- **epistemic uncertainty (due to finite samples)** is the statistical uncertainty on those shapes, and
- **epistemic uncertainty (across replicates)** is the propensity of the PPF to change when the experiment is replicated.

Both forms of epistemic uncertainty can contribute uncertainty on the risk, i.e. increase the spread of the R-distribution, but in the $L \to \infty$ limit only the effect of replications remains. Changes to aleatoric uncertainty will shift R-distributions along the R axis, but do not directly contribute epistemic uncertainty.

Flexibility in selecting c

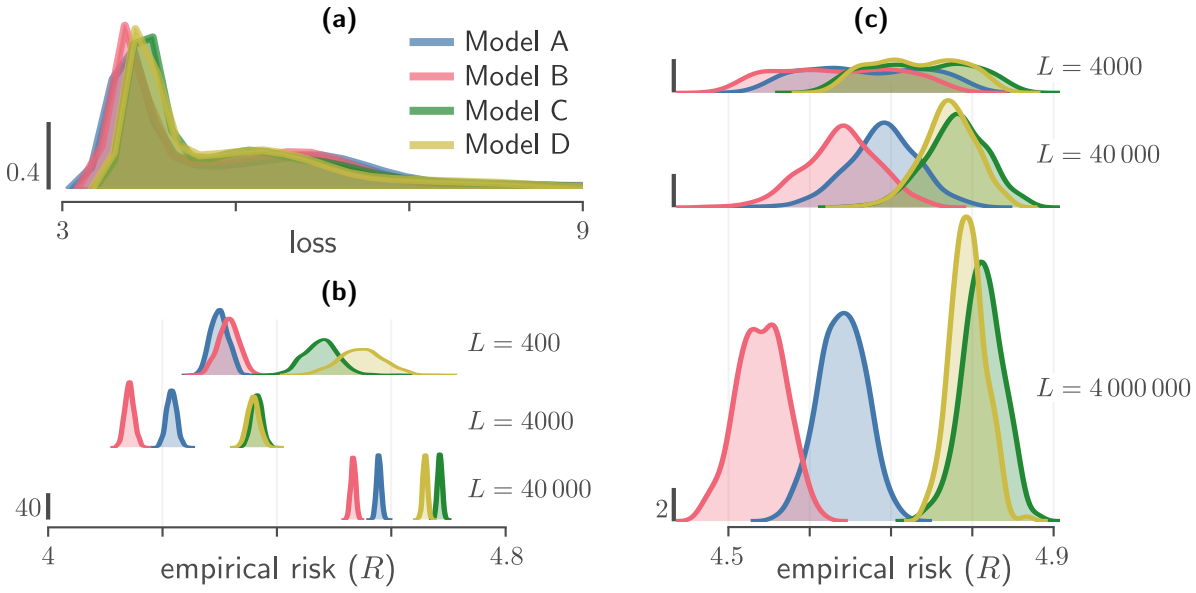
An important property of our approach is that sensitivity parameter c does not need to be tuned to a precise value, but can lie within a range; either $c \in [2^{-4}, 2^0]$ for the neuron models of Fig. 6, or $c \in [2^{-6}, 2^3]$ for the Planck and Rayleigh-Jeans models of Fig. 9. This does not mean that the value of $B^{\mathrm{EMD}}_{ab;c}$ itself is insensitive to c: for fixed data, a larger c will generally bring $B^{\mathrm{EMD}}_{ab;c}$ closer to 50%. But the probability bound given by $B^{\mathrm{EMD}}_{ab;c}$ remains correct for all c within that range.

For example, we compute $B^{\mathrm{EMD}}_{\mathrm{P,RJ};c}$ to be 80% when $c = 2^{-3}$, versus 60% when $c = 2^0$. This means that if we fix $c = 2^{-3}$ (resp. $c = 2^0$), among all experiments which yield $B^{\mathrm{EMD}}_{\mathrm{P,RJ};c} = 80\%$ (resp. $B^{\mathrm{EMD}}_{\mathrm{P,RJ};c} = 60\%$), in at least 80% (resp. 60%) of them model $\mathcal{M}_\mathrm{P}$ will have lower true risk than $\mathcal{M}_\mathrm{RJ}$.

More generally, a $B^{\mathrm{EMD}}_{ab;c} = B$ (with $0.5 < B \leq 1$) states that, of all the experiments where the calculation of $B^{\mathrm{EMD}}_{ab;c}$ is equal to B, at least a fraction B of those will have $R_a < R_b$. In this respect the B^{EMD} exhibits some similarities with a confidence interval, in that it is interpreted in terms of *replications under a fixed computational procedure*: in the former case we have a fixed procedure for calculating $B^{\mathrm{EMD}}_{ab;c}$ given c, while in the latter case we have a fixed procedure for calculating the confidence interval given a confidence level. A key difference however is that with the B^{EMD}, the interpretation requires conditioning on the outcome of the calculation.

Of course the range of valid c values will depend on the variety of epistemic distributions, and cannot be guaranteed. In general, the larger the variety, the more difficult one can expect it to be to find a c which is valid in all conditions. We can however anticipate a few strategies which might increase the range of valid c values and otherwise improve the ability of the B^{EMD} criterion to discriminate between models:

Increased rejection threshold Increasing the rejection threshold ϵ in equation (15) can be a simple way to add a safety margin to the B^{EMD} criterion, at the cost of some statistical power, to account for small violations of equation (35).



Supplementary Figure 3: **Aleatoric and finite-size uncertainty. a) Loss distribution** of individual data points — $\{Q_a(t_k, V^{\mathsf{LP}}(t_k; \mathcal{M}_{\text{true}}))\}$ — for the dataset and models shown in Fig. 2 and $a \in \{A, B, C, D\}$. For a model which predicts the data well, this is mostly determined by the aleatoric uncertainty. **b) Bootstrap estimate of finite-size uncertainty** on the risk, obtained using case resampling:[3] for each model, the set of losses was resampled 1200 times with replacement. Dataset and colours are the same as in (a); the same L data points are used for all models. **c) Synthetic estimate of finite-size uncertainty** on the risk, obtained by evaluating equation (10) on 400 different simulations of the candidate model (differing by the random seed). Here the same model is used for simulation and loss evaluation. Dataset sizes L determine the integration time, adjusted so all datasets contain the same number of spikes. All subpanels use the same vertical scale. **b–c)** The variance of the R-distributions, i.e. the uncertainty on R, goes to zero as L is increased. **a–c)** Colours indicate the model used for the loss. Probability densities were obtained by a kernel density estimate (KDE).

Multi-step comparisons Initially, the large number of candidate models may force the selection of a larger (i.e. more conservative) c. After using this c to reject some of the candidates, it may be possible to reduce the value of c, thus increasing the discriminatory power and possibly further reducing the remaining pool of candidates.

Improved experimental control If one can improve the reproducibility across experiments, the epistemic distributions used for calibration can correspondingly be made tighter. This makes it easier to find a c which works in all experimental conditions, since there are overall fewer conditions to satisfy.

Domain-informed loss function A loss function designed with knowledge of the domain or target application can ignore irrelevant differences between models. In addition to improving the relevance of comparisons, this tends to make the risk a smoother function of experimental parameters, which should make the B^{EMD} easier to calibrate.

Post hoc correction of B^{EMD} values As long as we select a c for which the $B^{\text{epis}}_{AB;\Omega}\left(B^{\text{EMD}}_{AB;c}\right)$ function of equation (54) is monotone, we can use the calibration curves themselves to interpret values of $B^{\text{EMD}}_{AB;c}$ by looking to the $B^{\text{epis}}_{AB;\Omega}$ to which they map. This approach could be used to improve discriminatory power of a conservative B^{EMD}, but also to correct it in regions where it is overconfident — assuming one has sufficient trust that the calibration curves are truly representative of experimental variations.

Comparing models directly with the loss distribution

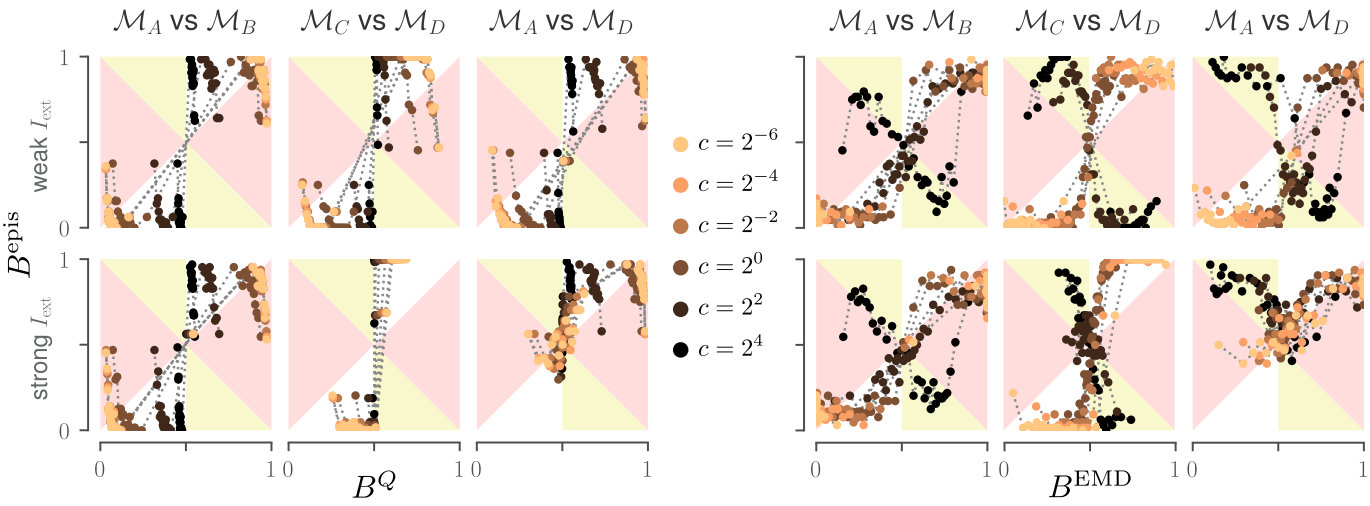
To further illustrate the previous point, we consider an alternative comparison criterion which directly uses the distribution of losses Q (i.e. the distribution in Supplementary Fig. 3**a**) instead of the more complicated process $\mathfrak{Q}$ over PPFs with which we defined the B^{EMD}. To this end, let us define a B^{Q} criterion in analogy with equation (16):

$$\begin{aligned} B^{Q}_{ab;c_Q} &:= P(Q_a < Q_b + \eta)\,, \\ \eta &\sim \mathcal{N}(0, c_Q^2)\,. \end{aligned} \tag{S26}$$

The Gaussian noise η is added to allow us to adjust the sensitivity of the criterion, analogously to how c adjusts the sensitivity of the $B^{\text{EMD}}_{ab;c}$ criterion. Both criteria thus have one free parameter, making the comparison relatively fair.

Note that in contrast to the B^{EMD}, the c_Q values in the B^{Q} are not dimensionless. Consequently, they would likely be less transferable between experimental contexts.

A possible argument in favour of using B^{Q} as a criterion is that increasing the amount of misspecification — for example by increasing the unmodelled bias $\mathcal{B}_0$ in equation (38) — will



Supplementary Figure 4: **(a) Calibration experiments for the neuron model, using B^{Q}** (equation (S26)) instead of B^{EMD} (equation (16)) as a comparison criterion. To better show the distribution of experiments, each histogram bin (see Calibration experiments in the Methods) is represented as a point. **(b)** Same curves as in Fig. 6, this time with bins presented as points to ease comparison with (a). Compared to (a), points are more uniformly distributed along both the horizontal and vertical axes. All panels use the same set of sensitivity values (either c_Q or c), with colours as indicated in the central legend.

affect the distribution of losses in some way. So although the shape of the PPFs is mostly determined by aleatoric uncertainty (as evidenced by the similarity of the distributions in Supplementary Fig. 3**a**), one can expect them to also contain some information about the amount of misspecification — and thereby also the amount of epistemic uncertainty.

Nevertheless, as we see in Figure 4 , the B^{Q} criterion is less effective for comparing models than the B^{EMD} in at least three ways (recall that the goal is not to find a criterion which accurately predicts the model with lowest risk, but one which accurately predicts the *uncertainty* on that risk, i.e. B^{epis}):

Reduced signal Only very few experiments on the main diagonal have B^{epis} values different from 0 or 1. The B^{Q} values therefore don't really inform about the epistemic uncertainty. Learning a monotone transformation from B^{EMD} or B^{Q} to B^{epis} is only possible if the curve is strictly monotone.

Increased anti-correlation between B^{Q} and B^{epis}: in five out of six comparisons between neuron models, we see strong anticorrelated tails at both ends of the curves, which dip all the way back to $B^{\text{epis}} \approx 0.5$. This makes it harder to find a c_Q for which the the B^{Q} is actually informative, since the curve is not even injective. While there is some anticorelation also with the B^{EMD} (Figure 4**b**), it is much less pronounced and limited to comparisons with the worse models ($\mathcal{M}_C$ and $\mathcal{M}_D$).

The sensitivity parameter does not really help Increasing c_Q squeezes curves horizontally, bringing all B^{Q} values closer to 0.5, but does not change their shape nor reduce the length of anticorrelated tails at both ends. All curves therefore effectively provide the same information. In contrast, the c parameter has a much stronger effect on the shape of R-distributions (c.f. Supplementary Fig. 2).

These results suggest that an approach based only on distributions of the loss Q is likely to always struggle with disentangling epistemic from aleatoric uncertainty. It also has less information to work with: whereas B^{EMD} is parameterised by two functions, q^* and δ^{EMD}, B^{Q} is parameterised by only one, the density $p(Q)$. Moreover, $p(Q)$ and q^* contain the same information (they are the PDF and PPF of the same random variable), so the discrepancy function δ^{EMD} provides B^{EMD} with strictly more information — information which is also specifically designed to disentangle the effects of misspecification from aleatoric uncertainty.

Supplementary References

1. Baets, B. D. & Meyer, H. D. in *Perception-Based Data Mining and Decision Making in Economics and Finance* (eds Batyrshin, I., Kacprzyk, J., Sheremetov, L. & Zadeh, L. A.) 261–274 (Springer, Berlin, Heidelberg, 2007). ISBN: 978-3-540-36247-0. doi: 10.1007/978-3-540-36247-0_10.
2. Conrey, B., Gabbard, J., Grant, K., Liu, A. & Morrison, K. E. Intransitive Dice. *Mathematics Magazine* **89,** 133–143. ISSN: 0025-570X, 1930-0980. doi: 10.4169/math.mag.89.2.133 (Apr. 2016).
3. Davison, A. C. & Hinkley, D. V. *Bootstrap Methods and Their Application* ISBN: 978-0-521-57471-6. doi: 10.1017/CBO9780511802843 (Cambridge University Press, Cambridge, 1997).
4. De Schuymer, B., De Meyer, H. & De Baets, B. Cycle-Transitive Comparison of Independent Random Variables. *Journal of Multivariate Analysis* **96,** 352–373. ISSN:

0047-259X. doi: 10.1016/j.jmva.2004.10.011 (Oct. 2005).

5. Gelman, A. *et al. Bayesian Data Analysis* ISBN: 978-1-4398-9822-2 (CRC Press, Boca Raton, 2014).
6. Marco Taboga. in *Lectures on Probability Theory and Mathematical Statistics* Online appendix (Kindle Direct Publishing, 2021).
7. Trotta, R. Bayes in the Sky: Bayesian Inference and Model Selection in Cosmology. *Contemporary Physics* **49,** 71–104. ISSN: 0010-7514. doi: 10.1080/00107510802066753 (Mar. 2008).
8. Vehtari, A., Gelman, A. & Gabry, J. Practical Bayesian Model Evaluation Using Leave-One-out Cross-Validation and WAIC. *Statistics and Computing* **27,** 1413–1432. ISSN: 1573-1375. doi: 10.1007/s11222-016-9696-4 (Sept. 2017).