Probabilistic Recalibration of Forecasts

Carlo Graziani[a,*], Robert Rosner[b], Jennifer M. Adams[c], Reason L. Machete[d]

[a] *Argonne National Laboratory, Lemont, IL, USA*

[b] *Department of Astronomy & Astrophysics, University of Chicago, Chicago, IL, USA*

[c] *Center for Ocean-Land-Atmosphere Studies, Fairfax, VA, USA*

[d] *Climate Change Division, Botswana Institute for Technology Research and Innovation, Gaborone, Botswana*

Abstract

We present a scheme by which a probabilistic forecasting system whose predictions have a poor probabilistic calibration may be recalibrated through the incorporation of past performance information in order to produce a new forecasting system that is demonstrably superior to the original, inasmuch as one may use it to win wagers consistently against someone who is using the original system. The scheme utilizes Gaussian process (GP) modeling to estimate a probability distribution over the probability integral transform (PIT) of a scalar predictand. The GP density estimate gives closed-form access to information entropy measures that are associated with the estimated distribution, which allows the prediction of winnings in wagers against the base forecasting system. A separate consequence of the procedure is that the recalibrated forecast has a uniform expected PIT distribution. One distinguishing feature of the procedure is that it is appropriate even if the PIT values are not i.i.d. The recalibration scheme is formulated in a framework that exploits the deep connections among information theory, forecasting, and betting. We demonstrate the effectiveness of the scheme in two case studies: a laboratory experiment with a nonlinear circuit and seasonal forecasts of the intensity of the El Niño-Southern Oscillation phenomenon.

[*] Corresponding author
Email address: `cgraziani@anl.gov` (Carlo Graziani)

1. Introduction

A forecast is necessarily a probabilistic affair, being an expression of uncertainty about the future. Probabilistic forecasts of events that fall along a continuum—such as short-term weather forecasts (Gneiting and Katzfuss, 2014; Brier, 1950), medium-term seasonal rainfall (Krzysztofowicz, 2014, 2001; Sharma, 2000), fluctuations of financial asset prices (Diebold, Gunther, and Tay, 1998; Kamstra, Kennedy, and Suan, 2001) or electrical demand (Maciejowska, Nowotarski, and Weron, 2016), rates of spread of infectious diseases (Held, Meyer, and Bracher, 2017; Moran et al., 2016), macroeconomic indicators (Casillas-Olvera and Bessler, 2006; Greenwood-Nimmo, Nguyen, and Shin, 2012), wind power availability (Pinson, 2012; Zhang, Wang, and Wang, 2014), species endangerment and extinction (Araújo and New, 2007), human population growth (Lutz, Sanderson, and Scherbov, 2001), and seismic activity (Kagan and Jackson, 2000; Marzocchi and Woo, 2007)—are of urgent interest to many kinds of decision-makers, and have occasioned much scientific literature across a broad range of fields.

Weather forecasting through numerical weather prediction (NWP) has improved its performance substantially over the past few decades due to improvements in observational data, computational models, and computational power. At present, it is capable of providing, on average, reasonably robust weather forecasts for periods in the order of 10 days (Stern and Davidson, 2015). Unfortunately, though, these physics-based forecasting schemes are known to lose their skill relative to empirical forecasting schemes for periods longer than ≈10 days. Thus, the obvious questions arise as to whether it is possible to achieve skillful forecasting for periods longer than 10 days, and if so, whether there is an upper bound on such forecasting.

Given the difficulty of achieving the computing power and model fidelity required to reduce the model errors of NWP, the best way forward for the present may well be to attempt to achieve some kind of synthetic hybrid between NWP and empirical forecast methods in an effort to use the information content of the former to enhance the predictive power of the latter. Important approaches that have been attempted include applying some kind of statistical recalibration to the NWP simulation output such as model output statistics (Glahn and Lowry, 1972) and then inferring probability distributions for predictands from the corrected simulations using some smoothing procedure (Coelho, Pezzulli, Balmaseda, Doblas-Reyes, and Stephenson, 2004; Gneiting, Raftery, Westveld, and Goldman, 2005); leaving the simulations as they are and adapting the smoothing procedure itself to validation data (Bröcker and Smith, 2008); or doing a bit of both (Raftery, Gneiting, Balabdaoui, and Polakowski, 2005; Fraley, Raftery, and Gneiting, 2010; Dutton, James, and Ross, 2013). One difficulty of such programs is that the smoothing procedure itself has statistical properties that usually are not under very good control, since they frequently take the form of highly simplified models such as Gaussian mixtures, which are not generally in well-motivated correspondence with the processes that relate the simulation output to the random predictand.

One common feature of the continuous forecast probability density functions (PDFs) that are produced from the NWP output is that more often than not they are *probabilistically miscalibrated*; that is, the long-term frequencies of observations do not match the stated probabilities of predictions (see for example Glahn et al., 2009). In such cases, the interpretation of the PDFs requires caution, and the value of having a probabilistic forecast rather than a point forecast may be questionable, particularly given the risk of underestimating the frequencies of extremes.

As was discussed by Diebold, Hahn, and Tay (1999, hereafter "DHT"), the phenomenon of probabilistic miscalibration creates another opportunity for recalibration: direct recalibration of the forecast probability distributions. This possibility arises because, regardless of the methodology adopted for producing a forecast system, using that system for long enough leads to the availability of additional performance information — through the comparison of a series of forecasts with their predictands — that can be incorporated into current forecasts in order to produce improved forecasts. Such information, which is used commonly to assess forecast system quality, was shown by DHT to permit the *correction* of future forecasts, assuming an i.i.d. restriction on predictands. More recently, Kuleshov, Fenner, and Ermon (2018, hereafter "KFE") devised a similar approach in the context of deep learning. They recast the usual machine learning activities of regression and classification as forecasting problems and regarded artificial neural network outputs as discrete or continuous predictands, respectively. Under an i.i.d. restriction on outputs, KFE obtained recalibration procedures that are equivalent to those of DHT, which are now cast as calibration procedures for classification and regression.

This study generalizes the work described by Diebold et al. (1999) and Kuleshov et al. (2018) in two respects. In the first place, we establish a mathematical framework for dealing with predictands without i.i.d. restrictions. In the process, we strongly emphasize the role of conditional information in forecast distributions, and make use of ideas from information theory for characterizing the effect of miscalibration. In addition, we replace the PDF estimation schemes suggested by

Diebold et al. (1999) and the isotonic regression adopted by Kuleshov et al. (2018) with a Gaussian-process (GP) density estimation scheme, which allows us to also estimate — with quantified uncertainties — the information entropy measures associated with the estimated pdfs. We use this technique to demonstrate that, if a series of forecasts shows evidence of poor probabilistic calibration, we can use past forecast performance information to produce new current forecasts that have well-calibrated expected distributions, and greater expected logarithmic forecast skill scores than the original forecasts, irrespective of whether the predictands are i.i.d. Furthermore, the expected performance improvement in logarithmic skill score can be computed in advance, together with an uncertainty estimate.

We demonstrate the method using two case studies: a laboratory experiment with a nonlinear circuit and seasonal forecasts of the intensity of the El Niño-Southern Oscillation (ENSO) phenomenon.

2. Probabilistic forecasts

The probabilistic forecast of a continuous scalar random variable X is simply a probability distribution $P(X|I)$ over the value of the predictand X, which is to be observed at a later date. The distribution is conditioned on prior information I, such as current and past conditions, (often approximate) deterministic and probabilistic model structures, empirically determined training parameters, and simulation outputs. Such forecasts are often generated as a time series $P(X_n|I_n, C)$, with $n \in \mathbb{Z}$ being an index that labels time t_n, so that $n_2 > n_1 \implies t_{n_2} > t_{n_1}$. Here, X_n is the random predictand at time t_n, I_n represents information that varies with n, and C represents static conditioning information that is constant for a particular forecasting system. Typically, the information I_n is stochastic and fluctuates randomly with n. As a consequence, the distribution $P(X_n|I_n, C)$ is itself a distribution-valued random variable (Gneiting et al., 2013; Gneiting and Katzfuss, 2014). Note that the notation $P(X_n|I_n, C)$ implies the assumption that the particular realization $I_n = i$ completely determines the distribution of X_n irrespective of n, so that $P(X_n|I_n = i, C) = P(X_m|I_m = i, C)$ for $m \neq n$. This will allow us to drop time subscripts from expressions such as $P(X|I, C)$ consistently in what follows.

As an example, in the case of weather prediction, I_n might represent a discrete vector of weather observations at a finite number of weather stations over the course of several previous days, while C might represent climatological data. Another example is provided by the empirical time series modeling that underlies many analyses of financial and economic data, where I_n could be the last M values of the time series X_n and C the parameters of an autoregressive moving-average (ARMA) time series model (Cressie and Wikle, 2015).

In the development our method we assume that the system is approximately stationary, so that secular drifts due to external forcings are ignored. We also overlook annual-type periodicities, which (in principle) are tractable if we add cycle phase information to I. One may show easily that a unique distribution $P(X|I, C)$ always exists in principle. This follows simply from the existence of a unique joint distribution $P(X, I|C)$, which can be ascertained empirically from a sufficiently large archive of (X_n, I_n) values. The forecast distribution $P(X|I, C)$ is then just $P(X, I|C)/P(I|C)$. This unique distribution is called the *ideal* forecast with respect to the information I (Gneiting et al., 2013; Gneiting and Katzfuss, 2014).

Above and beyond empirical observation, there is often some kind of dynamical law from which, in principle, $P(X|I, C)$ could be inferred. In such cases, however, accurate inference of $P(X|I, C)$ from first principles is often impractical, whether because the dynamical law is not known (as in the case of most time series in economics), because it is known imperfectly (as is the case with weather forecasting), or because it is not feasible to compute it even where it is well understood.

Weather forecasting furnishes an instructive example. The dynamic origin of the distribution $P(X|I, C)$ is intelligible in terms of the nonlinear physics of weather systems. However, while this forecast distribution can be described, it is in no way feasible to compute, because of limitations in both model fidelity and computational resources. Instead, limited-fidelity computational models (Fritsch and Carbone, 2004; Knutti and Sedláček, 2013) are used to filter the information in I, incorporating techniques of data assimilation (Bengtsson, Ghil, and Källén, 1981) and evolving ensembles of states that are not chosen by a fair sampling of the distribution on the observation-constrained submanifold of the chaotic attractor. In any case, it has too few ensemble members to be sufficiently informative about the distribution's structure. The postprocessing of a training set of ensembles and the corresponding validation values of X must be used to construct an approximation of $P\left(X|I, C\right)$ (Raftery et al., 2005; Roulston and Smith, 2002).

Clearly, by the time this approximation has been constructed, it is not necessarily conditioned directly on I any more, but rather is conditioned on some highly processed information $J[I]$. In ensemble NWP, $J[I]$ has both a deterministic aspect (the NWP simulations) and a stochastic aspect (the selection of the random ensemble of initial conditions to evolve). Quite generally, we can assume that some probabilistic model $J \sim P(J|I)$ describes the dependence of J on I. If that mapping happens to be deterministic, the probability distribution $P(J|I)$ would degenerate to a product of Dirac δ-distributions. In general, the dimensionality of J is not necessarily inferior to that of I: in the case of NWP, the simulations generate data over grids the data mass of which far exceeds that of the input information. Invariably, though, the *information content* of J is degraded relative to that of I, due to the very approximations described above. This is merely the observation that $P(X|J[I], C)$ is expected to be — and generally is — inferior to the computationally infeasible $P(X|I, C)$ in quality measures such as calibration and sharpness (discussed below).

Another practical concern in obtaining $P(X|I, C)$ is the fact that the static conditioning information C may not be known exactly and must be estimated from data. For example, even if a financial time series was known to be approximated well by a stationary autoregressive process, in general the process parameters would not be known, and would have to be fitted from the data. Similarly, in weather forecasting, climatological information would have to be fitted from noisy data.

We will assume that forecasts are modeled appropriately by absolutely continuous distributions, which may therefore be represented by probability densities over X. The forecasting system converts the information $J_n = J[I_n]$ and C into a *published forecast* $p(X_n; J_n, C)$, a density over X_n, at each time n. The data-generating process that is being forecast then generates a realization $X_n = x_n$. Suppose that we generate N forecasts and N corresponding observations. The series of pairs $\{\mathcal{P}_n = (x_n, p(\cdot; J_n, C))\,, n = 1, 2, \ldots N\}$ is called the *forecast-observation archive* (FOA) (Suckling and Smith, 2013; Smith, Suckling, Thompson, Maynard, and Du, 2015). The pairs $\mathcal{P}_n$ may be viewed as elements of a set called a *prediction space*, the mathematical properties of which were analyzed by Gneiting and Katzfuss (2014).

One of the most important tools for assessing the validity of published forecasts is the *probability integral transform*, or PIT (Dawid, 1984; Diebold et al., 1998; Gneiting, Balabdaoui, and Raftery, 2007). This is defined in terms of the cumulative probability distribution function $\tilde{F}(\cdot; J, C)$ associated with $p(\cdot; J, C)$,

$$\tilde{F}(x; J, C) = \int_{-\infty}^{x} dx'\, p(x'; J, C). \tag{1}$$

The PIT associated with the pair $\mathcal{P}_n = (X_n = x_n, p(\cdot; J_n, C))$ is simply the value $f_n \equiv \tilde{F}(x_n; J_n, C)$. The reason for the usefulness of the PIT is that if the density

$p(\cdot; J_n, C)$ models the stochastic behavior of $X_n | J_n, C$ correctly, then the random variable $F_n = \tilde{F}(X_n; J_n, C)$ must be distributed uniformly over the interval $[0, 1]$ irrespective of J_n. If this is the case, we say that $p(\cdot; J, C)$ is *probabilistically calibrated* (Gneiting et al., 2013; Gneiting and Katzfuss, 2014). Probabilistic calibration is a desirable feature in a published forecast because it means that the forecast is "honest" about the probabilities of its quantiles, since those probabilities correspond to long-term average frequencies. The property of being probabilistically calibrated may be checked, given a sufficiently large FOA, by displaying the values of F_n in a histogram and inspecting the histogram for evidence of non-uniformity (Dawid, 1984; Diebold et al., 1998; Gneiting et al., 2007). All ideal forecasts are probabilistically calibrated, although the reverse is not true: it is easy to construct many different calibrated forecasts, but only one is ideal with respect to the input information.

It may perhaps seem surprising that calibration, while a desirable feature of a forecast system, is not sufficient for one forecast system to be preferred over another. As was discussed by Gneiting et al. (2007) and Hamill (2001), it is quite possible for forecast systems that yield distributions which vary widely in precision to all be equally probabilistically calibrated. For example, a "climatological" forecast system that uses only long-term historical averages to make predictions and an idealized perfect NWP model that makes approximation-free use of the current information I_n to make ideal forecasts are both equally probabilistically calibrated from the point of view of PIT histogram uniformity. Clearly, the former provides forecasts that are vague compared with those of the latter, which are more informative and precise. The term *sharpness* was introduced by Bross and Bross (1953) to characterize this distinction. It refers to the degree of concentration of the published forecast density on small outcome sets, and is sometimes expressed as the distributional variance or the width of a central fixed-probability (e.g., 90%) interval (Gneiting et al., 2007). Thus, a sharper forecast is less vague in its predictions than a less-sharp one, independently of the relative degree to which their respective PIT histograms are close to uniform.

Clearly, the difference between probabilistically calibrated forecasts of different sharpness, for example the difference between the climatological and idealized NWP forecasters, lies *purely in the information on which the forecasts are conditioned.* States of more specific information lead to sharper forecasts. In the case of the idealized NWP forecaster, for example, a large increase in the number of weather stations available will necessarily lead to sharper forecasts, while an increase in the measurement uncertainty of current weather conditions will necessarily lead to less-sharp forecasts. The explicit highlighting of the relevant conditioning information is therefore essential to the discussion of probabilistic forecasting.

In the case of poorly-calibrated forecasts, the source of the misspecification of the forecast distributions must necessarily be sought in erroneous conditioning information, such as model errors that distort the information borne by the processed input data $J[I]$, model errors in constructing the published forecast distribution, or poor approximations encoded in the static conditioning information C. Forecast interpretation in the presence of misinformation is an important subject in decision support (Smith, 2016). One may be confronted with cases of sharp but uncalibrated forecasts that are (for example) biased, but that have smaller mean square errors than climatology. In these cases, the incorrect conditioning information may not be condemned entirely, because such forecasts can have better predictive skill than climatology. One naturally wonders about the extent to which this partially correct information can be exploited to produce probabilistically calibrated forecast distributions.

The next section shows that a knowledge of the PIT histogram of a sufficiently large FOA can be used to correct a current published forecast $p(\cdot; J_n, C)$ prior to the observation of the predictand X_n, in order to produce a new, updated forecast $p_1(\cdot; J_n, C)$ that outperforms $p(\cdot; J_n, C)$, in that it has a better expected logarithmic ("ignorance") score (the ignorance score is defined by Roulston and Smith, 2002, and is discussed further in Section 3.3 below). This *probabilistic recalibration* procedure allows us to exploit the correct part of the conditioning information better.

3. Probabilistic recalibration

The essence of the probabilistic recalibration procedure is that the PIT histogram of a sufficiently large FOA can be subjected to an empirical fit, so that the underlying distribution can be inferred by regression. Assuming that current forecasts suffer from the same miscalibration as those in the FOA, the fit distribution can then be used to correct a current forecast distribution in order to produce a new forecast that outperforms the original according to various objective measures, including the ignorance score. We now set out the procedure.

Since we have raised the issue of misinformation in connection with miscalibration, we adopt a notation that distinguishes between distributions that are ideal (that is, that reflect their conditioning information correctly) and distributions that may be misinformed or poorly calibrated. In what follows, therefore, we will reserve the symbol π and the notation $\pi\,(A|B)$ or $\pi(A = a|B)da$ for the probability density function of a random variable A that is conditioned correctly on information B, so that $\pi(A|B)$ is ideal. Published forecast densities, which may be "misinformed" and hence reflect the dependence on conditioning information incorrectly, we denote by simple function notation such as $p(x; J)$.

We denote the input information to the $t = t_n$ published forecast by the random variable $\mathcal{J}_n$, whose realization is J_n. We assume the existence of a unique ideal distribution relative to J, C with density $\pi\,(X|\mathcal{J} = J, C)$. Again, such a unique forecast clearly exists, based on its relationship with the empirically-ascertainable joint distribution $\pi(X, \mathcal{J}|C)$.

A published forecast $p(\cdot; J_n, C)$ and the corresponding observations x_n give rise to a PIT value $f_n = \tilde{F}(x_n, J_n, C)$, which is a realization of a random variable $F_n \equiv \tilde{F}(X_n, J_n, C)$. The variable F_n is simply a change of random variables from X_n, which implies an ideal density $\pi(F_n|\mathcal{J} = J_n, C)$ for F_n satisfying

$$\begin{aligned}\pi\,(X_n = x_n|\mathcal{J}_n = J_n, C) &= \pi\left(F_n = \tilde{F}(x_n, J_n, C)|\mathcal{J}_n = J_n, C\right)\frac{d\tilde{F}}{dx}\\ &= \pi\left(F_n = \tilde{F}(x_n, J_n, C)|\mathcal{J}_n = J_n, C\right)p(x_n; J_n, C). \quad (2)\end{aligned}$$

Eq. (2) connects the published forecast $p(\cdot; J_n, C)$ to $\pi\,(X_n|\mathcal{J}_n = J_n, C)$, the unknown ideal forecast distribution density relative to J_n, by a pointwise multiplication with another unknown density $\pi\,(F_n|\mathcal{J}_n = J_n, C)$.[1]

If the observables F_n are i.i.d., we may drop the dependence of $\pi\,(F_n|\mathcal{J}_n, C)$ on $\mathcal{J}_n$ in Eq. (2). In this case, one may proceed straightforwardly to estimate the time-independent distribution $\pi(F|C)$ (where F may be any of the identically-distributed F_n) by regression on the FOA PIT data $\mathcal{F} \equiv \{F_n : n = 1, 2, \ldots, N\}$ as described by DHT. Denoting this estimate by $\pi\,(F|\mathcal{F}, C)$ and replacing $\pi\,(F_n|\mathcal{J}_n, C)$ in Eq. (2) with $\pi\,(F_n|\mathcal{F}, C)$ results in forecasts that have improved calibration properties. Equivalently, KFE perform isotonic regression on what is, in effect, the i.i.d. CDF of the F_n (as opposed to their i.i.d. PDF), to obtain an improved

[1] For the sake of simplicity, we assume that the published forecasts $p(x; J, C)$ are always nonzero for all x. If such were not the case, an interval $[x_1, x_2]$ over which $p(x; J, C)$ is zero would be mapped to a single point F_1 by the $x \longrightarrow F$ change of variables. Should the ideal distribution density $\pi\,(X|\mathcal{J}, C)$ happen to have a nonzero probability mass over such an interval, that finite probability would be mapped to the single point F_1. It would then be necessary to represent this effect by an additive Dirac-δ distributional component in $\pi(F|\mathcal{J}, C)$. Such a generalization would be cumbersome, and we avoid having to address it by specifying $p(x; J_n, C) > 0$ for all x.

probabilistic calibration of deep learning classifiers and regressors. The theory developed by Diebold et al. (1999) and Kuleshov et al. (2018) does not address the important general case of a non-i.i.d. F_n, however, because the regression estimate $\pi(F|\mathcal{F}, C)$ from the FOA averages over all conditioning data $\mathcal{J}$, and in this sense is a “climatological” distribution that is ignorant of current conditioning information $\mathcal{J}_n = J_n$.[2] DFE address the i.i.d. restriction by considering published and ideal forecasts that belong to different “scale-location” families with the same “scale-location” parameters, and show that this case does give rise to i.i.d. F_n. However, the generality of this restriction is problematic. As we discuss below in Section 3.6, and demonstrate explicitly in Section 4.2, it is not uncommon for time series of $F_n = f_n$ realizations to exhibit strong correlations. In such cases, the i.i.d. assumption on the F_n is simply not tenable.

Nonetheless, Eq. (2) is the starting point of our probabilistic recalibration procedure. We will show that if we replace the density $\pi(F|\mathcal{J}, C)$ in Eq. (2) with the predictive distribution density $\pi(F|\mathcal{F}, C)$ estimated by a Bayesian regression fitted to the FOA PIT data $\mathcal{F}$, we will obtain a new forecast distribution that is not ideal but still improves on the logarithmic (“ignorance”) skill score of the published forecast, irrespective of whether the F data are i.i.d. or not.

3.1. Bayesian PIT-Fit

We now perform the regression fit to the data $\mathcal{F}$. Diebold et al. (1999) recommended either using a kernel estimator or simply using the empirical PIT distribution directly, whereas Kuleshov et al. (2018) recommended isotonic regression on the PIT CDF. Here, we adopt a non-parametric procedure: Gaussian process measure estimation (GPME), described in detail in Appendix A. This is a more complicated procedure than has been used for this task previously, but it has benefits that will be described presently. As used here, GPME effectively fits functions from an infinite-dimensional function space in order to estimate the predictive density $\pi(F|\mathcal{F}, C)$.

Our stationarity assumption implies that the FOA PIT data $\mathcal{F}$ may be viewed as a realization of a process that repeatedly samples a climatologically averaged distribution, which corresponds to many different random realizations J of the

[2] Note that the i.i.d. assumption of Diebold et al. (1999) and Kuleshov et al. (2018) applies to the distribution of the F_n, not to the forecast distribution of the X_n. The latter are not “climatological”, in that they are conditioned on their individual $\mathcal{J}_n$.

random variable $\mathcal{J}$ that represents the input information. The climatology gives rise to a distribution $\pi\left(\mathcal{J}|C\right)$, which we use to average $\pi\left(F|J,C\right)$, obtaining

$$\begin{aligned}\pi\left(F|C\right) &= \int dJ\,\pi\left(\mathcal{J}=J|C\right)\times\pi\left(F|\mathcal{J}=J,C\right)\\ &= E_{\mathcal{J}}\left\{\pi\left(F|\mathcal{J},C\right)\right\}. \end{aligned}\tag{3}$$

The distribution $\pi(F|C)$ is unknown and must be estimated by regression from the noisy FOA PIT data in $\mathcal{F}$. This density function estimate bears uncertainty that is represented by the Gaussian process posterior distribution over the density function $\pi(F|C)$ given $\mathcal{F}$. This uncertainty is purely epistemic, in contrast to the uncertainty that is consequent on the stochastic nature of $\mathcal{J}$, represented by the climatological distribution $\pi(\mathcal{J}|C)$.

In regard to notation, we will represent the uncertainty in the determination of the ideal density function $\pi(X|\mathcal{J},C)$ using a density function valued random variable $\Pi(X|\mathcal{J},C)$, the realizations of which are possible densities $\pi(X|\mathcal{J},C)$. We refer to density function valued random variables such as $\Pi(X|\mathcal{J},C)$ as *imperfectly known distributions*. The distribution $\Pi(X|\mathcal{J},C)$ is imperfectly known because our knowledge of it comes from a database of time series of pairs $(\mathcal{J},X)$, from which the joint distribution $\pi(X,\mathcal{J}|C)$, and hence $\pi(X|\mathcal{J},C)$, could be estimated empirically, with considerable uncertainty.

One may use Eq. (2) to define the imperfectly known distribution $\Pi(F=f|\mathcal{J}=J,C)=\Pi\left(X=x|\mathcal{J}=J,C\right)/p\left(x;J,C\right)$, where $\tilde{F}(x;J,C)=f$, the realizations of which are possible densities $\pi(F|\mathcal{J},C)$.

The uncertainty in $\pi(F|C)$ is then represented by an imperfectly known distribution $\Pi(F|C)\equiv E_{\mathcal{J}}\left[\Pi(F|\mathcal{J},C)\right]$, the realizations of which are possible density functions $\pi(F|C)$. The prior distribution over $\Pi(F|C)$ is described in GPME by a Gaussian process over $\log\Pi(F|C)$ with a chosen kernel $K(f_1,f_2)$ (here squared-exponential) and a constant mean function. The posterior distribution over $\Pi(F|C)$ given $\mathcal{F}$ is described in GPME by an updated Gaussian process over $\log\Pi(F|C)$, with a mean function $\lambda(f)$ that is given by Eq. (A.27), and a covariance $C(f_1,f_2)$ that is given by Eq. (A.28).

The *predictive distribution* $\pi(F|\mathcal{F},C)$ is the expectation of $\Pi(F|C)$ under this posterior distribution $\Pi(F|C)|\mathcal{F}$; that is

$$\pi\left(F|\mathcal{F},C\right) = E_{\Pi(F|C)|\mathcal{F}}\left\{\Pi(F|C)\right\}\tag{4}$$

$$= E_{\Pi(F|C)|\mathcal{F}}\left\{E_{\mathcal{J}}\left[\Pi(F|\mathcal{J},C)\right]\right\}.\tag{5}$$

Eq. (4) expresses the operation by which $\pi\,(F|\mathcal{F},C)$ is obtained from the GPME posterior distribution. Eq. (5) illustrates the fact that the predictive distribution incorporates both epistemic fit uncertainties and climatological averaging. This distribution, which is the key quantity that enables the probabilistic recalibration procedure, is given explicitly in Eq. (A.36) and is the principal output of the GPME procedure.

3.2. The recalibration procedure

As was adumbrated above, the recalibration procedure consists of replacing the published forecast density $p(x;J,C)$ with the recalibrated forecast density $p_1(x;J,C)$ that is given by the *recalibration equation*

$$p_1(x;J,C) = \pi\left(F = \tilde{F}(x;J,C)|\mathcal{F},C\right) \times p(x;J,C), \tag{6}$$

which is obtained from Eq. (2) simply by replacing $\pi(F|\mathcal{J},C)$ with the predictive distribution $\pi(F|\mathcal{F},C)$, estimated by the GPME procedure.

We may gauge how much has been gained through the recalibration procedure by introducing the Kullback-Leibler divergence or relative entropy (Kullback and Leibler, 1951) of one density $f_1(x)$ relative to another density $f_2(x)$:

$$KL[f_2\,||\,f_1] = \int dx\, f_2(x)\, \log_2 \frac{f_2(x)}{f_1(x)}, \tag{7}$$

which may be viewed usefully as a measure of the information embodied by f_2 relative to a prior state of information that is embodied by f_1. The divergence $KL[f_2\,||\,f_1]$ has the well-known property of being non-negative definite, and of being zero only if $f_1 = f_2$ almost everywhere.

By setting f_2 to the imperfectly known ideal distribution $\Pi(X|\mathcal{J} = J,C)$ and setting f_1 alternatively to the published forecast $p(x;J,C)$ and to the recalibrated forecast $p_1(x;J,C)$, we may, by taking expectations over the climatology $\mathcal{J}$ and over the posterior GPME model of the density $\Pi(F|C)\,|\,\mathcal{F}$, assess whether the ideal distribution is closer in information to the recalibrated forecast than to the original published forecast.

This leads to the following theorem.

Theorem 1. *(i) The distribution $\Pi\,(X = x|\mathcal{J} = J,C)$ is closer in expected (over $\mathcal{J}$ and $\Pi(F|C)\,|\,\mathcal{F}$) relative entropy to the recalibrated forecast $p_1(x;J,C)$ than it is to the published forecast $p(x;J,C)$ unless $p_1 = p$ almost everywhere. (ii) The recalibrated forecast $p_1(x;J,C)$ is on average (over $\Pi(F|C)\,|\,\mathcal{F}$) probabilistically calibrated.*

We prove (i) by first defining the difference in the relative entropies

$$\begin{aligned}
\Delta s &\equiv KL\left[\Pi(X|\mathcal{J} = J, C)\,||\,p(\cdot; J, C)\right] - KL\left[\Pi(X|\mathcal{J} = J, C)\,||\,p_1(\cdot; J, C)\right] \\
&= \int dx\, \Pi(X = x|\mathcal{J} = J, C)\, \log_2\left(\frac{p_1(x; J, C)}{p(x; J, C)}\right) \\
&= \int_0^1 df\, \Pi(F = f|\mathcal{J} = J, C)\, \log_2\left[\pi(F = f|\mathcal{F}, C)\right]. \qquad (8)
\end{aligned}$$

This quantity is a random variable, in consequence of the stochastic nature of $\mathcal{J}$ and the epistemic uncertainty in $\Pi(F|\mathcal{J}, C)\,|\,\mathcal{F}$. Taking the required expectations, we obtain

$$\begin{aligned}
E_{\Pi(F|C)\,|\,\mathcal{F}}\left\{E_{\mathcal{J}}\left[\Delta s\right]\right\} &= E_{\Pi(F|C)\,|\,\mathcal{F}}\left\{\int_0^1 df\, \Pi\,(F = f|C)\, \log_2\left[\pi(F = f|\mathcal{F}, C\right]\right\} \\
&= \int_0^1 df\, \pi\,(F = f|\mathcal{F}, C)\, \log_2\left[\pi\,(F = f|\mathcal{F}, C)\right] \\
&\geq 0, \qquad (9)
\end{aligned}$$

with equality holding only when $\pi(F = f|\mathcal{F}) = 1$ almost everywhere, which is to say, when the original distribution was probabilistically calibrated. In this case, $p_1 = p$.

To see (ii), observe that the cumulative distribution of $p_1(x; J, C)$ defines a change in variables from the random variable F to a new random variable G through the function $\tilde{G}(F; C)$, given by

$$\begin{aligned}
\tilde{G}(f; C) &\equiv \int_{-\infty}^{\tilde{F}^{-1}(f;J,C)} dx'\, p(x'; J, C)\, \pi(F = \tilde{F}(x'; J, C)|\mathcal{F}, C) \\
&= \int_0^f df'\, \pi(F = f'|\mathcal{F}, C). \qquad (10)
\end{aligned}$$

The function $\tilde{G}(\tilde{F}(x; J, C); C)$ is the PIT function of the recalibrated forecast distribution $p_1(x; J, C)$. In terms of the imperfectly known distribution $\Pi(F = f|\mathcal{J}, C)$, the imperfectly known distribution $\Pi(G = g|\mathcal{J}, C)$ is

$$\begin{aligned}
\Pi(G = g|\mathcal{J}, C) &= \frac{\Pi(F = f|\mathcal{J}, C)}{|d\tilde{G}/df|} \\
&= \frac{\Pi(F = f|\mathcal{J}, C)}{\pi(F = f|\mathcal{F}, C)}, \qquad (11)
\end{aligned}$$

where $g = \tilde{G}(f; C)$. The imperfectly known PIT distribution for p_1 is obtained by averaging over $\mathcal{J}$:

$$\begin{aligned}\Pi(G = g|C) &= E_{\mathcal{J}}\left\{\Pi(G = g|\mathcal{J}, C)\right\} \\ &= \frac{\Pi(F = f|C)}{\pi(F = f|\mathcal{F}, C)}.\end{aligned} \tag{12}$$

Averaging over the imperfectly known distribution $\Pi(F|C) \,|\, \mathcal{F}$ and using Eq. (4), we find

$$\begin{aligned}E_{\Pi(F|C)\,|\,\mathcal{F}}\left\{\Pi(G = g|C)\right\} &= \frac{\pi(F = f|\mathcal{F}, C)}{\pi(F = f|\mathcal{F}, C)} \\ &= 1.\end{aligned} \tag{13}$$

Hence, on average over $\Pi(F|C) \,|\, \mathcal{F}$, the PIT of the recalibrated forecast p_1 is uniform. □

Note a curious feature of this theorem: it does not use any fact about the GPME procedure, other than that it allows averaging over the posterior distribution $\Pi(F|C) \,|\, \mathcal{F}$. The statements of the theorem would be true given *any* such distribution, even one that is wildly incorrect about the probable shape of the true $\pi(F|C)$. However, if the posterior distribution over $\Pi(F|C) \,|\, \mathcal{F}$ did happen to be wildly wrong, the theorem, while still true, would no longer furnish the basis for a working recalibration procedure. The reason for this is that the actual data-generating process produces an observable value of $E_{\mathcal{J}}\left\{\Delta s_{True}\right\}$, given by

$$\Delta S_{True} \equiv E_{\mathcal{J}}\left\{\Delta s_{True}\right\} = \int_0^1 df\, \pi(F = f|C) \log_2 \pi(F = f|\mathcal{F}, C), \tag{14}$$

where the subscript indicates that the true (unknown) distribution $\pi(F|C)$ is used in the average, and where the distinction between Δs and ΔS is that the latter is averaged over $\mathcal{J}$. This expression differs from the expression in Eq. (9) and is under no obligation to be non-negative definite. This expression can be expected to be strongly positive only when $\pi(F|\mathcal{F}, C)$ approaches $\pi(F|C)$, which is to say when the fitting procedure results in a posterior distribution that is well concentrated near density functions that look a lot like $\pi(F|C)$. Thus, the success of the density-fitting regression procedure is essential to the success of the recalibration procedure.

This observation applies equally to other styles of regression estimates for $\pi(F|\mathcal{F}, C)$, including the kernel density and histogram estimators of DHT (Diebold et al., 1999) and the isotonic regression of KFE (Kuleshov et al., 2018). As long as those procedures succeed in furnishing an estimate of $\pi(F|\mathcal{F}, C)$ that is reasonably close to $\pi(F|C)$ and not too uncertain, they too should produce recalibrated forecasts that are informationally closer to the ideal forecast, on average (over the uncertainty in those estimates), than is the published forecast. The novel thing here is that we can now see that this ought to happen *irrespective of whether or not the F_n are i.i.d.* This is an important observation, since it was not clear previously to what extent correlations among the F_n could be expected to damage or even vitiate recalibration. As a result, the work of KFE and DHT can be seen to have a broader applicability than might otherwise have been believed to be the case.

The recalibrated forecast $p_1(x; J, C)$, while probabilistically calibrated on average, is not in general the same distribution as the unique ideal distribution $\pi(X = x|\mathcal{J} = J, C)$, even when the size of the FOA is large and the uncertainty in $\Pi(F|C)$ is very small. With a sufficiently large database of pairs (x, J), one could estimate $\pi(X = x|\mathcal{J} = J, C)$ empirically and discern its differences from $p_1(x; J, C)$. What the theorem establishes is that p_1 is a better approximation to the ideal forecast than is p on average, in the sense that it is closer in information to the ideal forecast.

3.3. Scoring, betting, and information

A natural connection exists between the relative entropy difference ΔS and the ignorance score (Roulston and Smith, 2002). For continuous predictands X and using the climatological density $\pi(X|C)$ as a reference distribution, the ignorance score $\mathrm{Ign}[p]$ of a forecast distribution $p(x; J, C)$ is defined by the expression

$$\begin{aligned}
\mathrm{Ign}[p] &= -E_{\mathcal{J}}\left\{E_{X|\mathcal{J},C}\left[\log_2 \frac{p(X;\mathcal{J},C)}{\pi(X|C)}\right]\right\} \\
&= -\int dJdx\, \pi(X = x, \mathcal{J} = J|C)\, \log_2 \frac{p(x; J, C)}{\pi(X = x|C)}. \qquad (15)
\end{aligned}$$

Because this final average is over the joint distribution on $X, \mathcal{J}$, the ignorance score may be estimated empirically from an FOA simply by the average

$$\mathrm{Ign[p]} \approx -\frac{1}{N}\sum_{n=1}^{N} \log_2 \frac{p(x_n; J_n, C)}{\pi(X = x_n|C)}, \qquad (16)$$

the expected value of which is the expression in Eq. (15). Note that for i.i.d. predictands x, the expression on the RHS of Eq. (16) is proportional to the log-likelihood for the model represented by $p(x; J, C)$, up to an additive constant. However, the resemblance is purely formal for non-i.i.d. predictands.

The difference in the ignorance scores of the published and recalibrated forecasts is

$$\begin{aligned}\Delta\mathrm{Ign}[p_1, p] &\equiv -E_{\mathcal{J}}\left\{\int dx\, \pi(X = x|\mathcal{J}, C)\, \log_2 \frac{p_1(x; \mathcal{J}, C)}{p(x; \mathcal{J}, C)}\right\} \\ &= -\int_0^1 df\, \pi(F = f|C)\, \log_2 \pi(F = f|\mathcal{F}, C) \\ &\approx -E_{\Pi(F|C)\,|\,\mathcal{F}}\left\{E_{\mathcal{J}}\left[\Delta s\right]\right\} \\ &\leq 0, \end{aligned} \tag{17}$$

where the third line approximates $\pi(F|C)$ by $\pi(F|\mathcal{F}, C)$ and the last line appeals to Theorem 1. We therefore expect that the recalibrated forecast p_1 will have a lower (i.e., better) ignorance score than the original published forecast. By the standards of the ignorance score, then, p_1 is an improvement on p.

It was pointed out by Roulston and Smith (2002) and Hagedorn and Smith (2009) that the ignorance score has an interpretation as a tool for practical decision-making under uncertainty. This interpretation is couched in terms of a horse race, in which a bettor and an oddsmaker make optimal decisions about their choices with respect to the discrete possible outcomes of the race. Kelly (1956) described a strategy for a player who was allocating wealth to bets on N outcomes $i = 1, \ldots, N$ that offered wealth multiplier odds o_i, assuming that the player works with a forecast distribution h_i, $\sum_{i=1}^N h_i = 1$. Kelly showed that the optimal strategy for the player is to allocate wealth W_i to bet on outcome i according to the rule $W_i = Wh_i$. In contrast, the optimal strategy for the pari-mutuel bookmaker with a forecast distribution g_i is to set odds $o_i = 1/g_i$. With these strategies for wagering and odds-setting, supposing that the ideal forecast is p_i, the player's wealth grows at the expected rate $2^{\Delta I}$, where

$$\begin{aligned}\Delta I &= \sum_{i=1}^N p_i \log_2 \frac{h_i}{g_i} \\ &= KL\left[p||g\right] - KL\left[p||f\right] \\ &= -\Delta\mathrm{Ign}[f, g] \end{aligned} \tag{18}$$

is the difference between the entropy divergences of the two forecast distributions relative to the true distribution p_i, and hence also the negative ignorance score difference. This can be regarded as a symmetric game between two forecasters who alternate the roles of bookmaker and bettor: as long as the current bookmaker with forecast g_i sets the odds in reciprocal proportion to g_i and the bettor with forecast h_i sets bets in proportion to h_i, Eq. (18) shows that it makes no difference which is which. A player with a forecast system that is consistently closer in information to the ideal distribution p_i (and hence has a lower ignorance score) than the other player's system has a positive expected wealth growth rate in this game.

This connection is attractive because it furnishes an example of decision-making under uncertainty that is improved by using forecasts with lower ignorance scores, and in particular by using the recalibration procedure described above. The view that underpins this work is the idea that there ought to be *some* form of symmetric-rule game in which decisions made according to a higher-scoring forecasting system should consistently dominate those made under a lower-scoring one. This view of forecast quality differs from the standard "proper scoring" outlook (Gneiting and Katzfuss, 2014; Gneiting and Raftery, 2007; Roulston and Smith, 2002; Good, 1952; Brier, 1950), in which forecasters vie with one another for high scores using scoring rules that are designed to encourage forecaster honesty and furnish usable utility functions for estimation. It corresponds better to the Bayesian decision theory outlook on proper scoring (Dawid and Sebastiani, 1999; Gneiting and Raftery, 2007).

The Kelly horse race is not ideal for furnishing a fiducial symmetric-rule game for continuous probabilistic forecast assessment, since it concerns itself with discrete outcomes and contains an appearance of asymmetry in the bettor–odds-maker model that it is based on. It can be adapted to continuous forecast distributions, for example by using fine quantiles (such as percentiles) as outcomes to be wagered on. However, a more natural symmetric-rule game can be described by observing that we may recast the second line of Eq. (17) as

$$\begin{aligned} -\Delta\mathrm{Ign}[p_1, p] &= \Delta S_{True} \\ &= \int dJdx\, \pi(X = x, \mathcal{J} = J|C)\, \log_2 \frac{p_1(x; J, C)}{p(x; J, C)}. \end{aligned} \tag{19}$$

Observe that the right-hand side of Eq. (19) may be interpreted as the expected winnings in a symmetric-rule game, which we call the *entropy game*. The rules of the game are as follows: two players compete, one using the published forecast distribution $p(\cdot; J, C)$ and the other using the recalibrated distribution $p_1(\cdot; J, C)$. At the nth turn of the game, values J_n, x_n are observed from the data-generating distribution $\pi(\mathcal{J}, X|C)$. The players compute the log-ratio of their respective densities at x_n, $w = \log_2 \left[p_1 \left(x_n; J_n, C \right) / p(x_n; J_n, C) \right]$. If $w > 0$, then the player using the published forecast p pays the amount w to the player using the recalibrated distribution p_1. Otherwise, the recalibrated player pays $|w|$ to the player using p. Eq. (19) states that ΔS_{True} is the expected winnings per turn of the recalibrated player.

The entropy game may be viewed as the logarithm of the Kelly horse race, since the expected per-turn winnings of the entropy game are in fact equal to the log (base two) of the expected wealth increase rate per turn of a bettor at a horse racing track. Kelly-style bets on (say) percentiles of the original forecast distribution have an expected wealth growth rate with a logarithm that is approximately equal to ΔS_{True}. The entropy game is a useful alternative to the Kelly horse race because it is better adapted to continuous distributions and its rules present a more symmetric appearance than does the bettor-and-bookie model of the horse race. It is our fiducial game for assessing the improvement in recalibrated forecasts, and we will make use of it extensively in what follows.

3.4. Predicting the recalibrated forecast performance

The entropy game winnings ΔS_{True} in Eq. (19) are expressed in terms of the distribution $\pi(X, J|C) \propto \pi(X|J, C)$, which is known only approximately through the GPME fit. An important and useful feature of the GPME procedure is that it allows us to compute the expected entropy game winnings per turn in advance of the game.

We first simplify ΔS_{True} by re-expressing it in PIT-space, using the fact that $\pi(X = x|J, C)dx = \pi(F = f|J, C)df$:

$$\begin{aligned}
\Delta S_{True} &= \int dJ\, \pi(\mathcal{J} = J|C) \int dx\, \pi(X = x|\mathcal{J} = J, C) \log_2 \frac{p_1(x; J, C)}{p(x; J, C)} \\
&= \int_0^1 df\, \pi(F = f|C) \log_2 \pi\left(F = f|\mathcal{F}, C\right). \qquad (20)
\end{aligned}$$

Note that, unlike the expression in Eq. (9), ΔS_{True} is not non-negative definite, but instead may (in principle) attain negative values if $\pi(F|\mathcal{F}, C)$ is a poor estimate of $\pi(F|C)$.

We define the quantity

$$\Delta S \equiv E_{\mathcal{J}} \left\{ \Delta s \right\} = \int_0^1 df\, \Pi(F = f|C) \log_2 \pi\left(F = f|\mathcal{F}, C\right), \qquad (21)$$

which is a random variable (unlike ΔS_{True}), in consequence of the use of the distribution-valued random variable $\Pi(F|C)$ for the average. This randomness expresses our epistemic uncertainty about the value of ΔS_{True} that is a result of our uncertainty about the shape of the distribution $\pi(F|C)$. The probability distribution of ΔS cannot be computed directly by using GPME. However, we can calculate

$$\overline{\Delta S} \equiv E_{\Pi(F|C)\,|\,\mathcal{F}}\left\{\Delta S\right\} \tag{22}$$

and

$$\mathrm{Var}\left(\Delta S\right) \equiv E_{\Pi(F|C)\,|\,\mathcal{F}}\left\{\Delta S^2\right\} - \left(\overline{\Delta S}\right)^2 . \tag{23}$$

These quantities represent the expectation and variance of ΔS with respect to the epistemic uncertainty contained in the GPME posterior distribution over the density $\Pi(F|C)\,|\,\mathcal{F}$. They constitute predictions of the outcomes of rounds of the entropy game that were to be conducted out-of-sample with respect to the FOA training data that furnish $\mathcal{F}$. Thus, not only can we verify the improvement in the recalibrated forecast using many rounds of the entropy game out-of-sample, but we can also predict in advance, with uncertainty bounds, what the average outcome of those games will be. The degree to which the predictions match the outcomes can be a useful gauge of model validity, as we will see below.

The expression for $\overline{\Delta S}$ can be obtained readily (even without appealing to the GPME theory):

$$\begin{aligned}
\overline{\Delta S} &= E_{\Pi(F|C)\,|\,\mathcal{F}}\left\{\int_0^1 df\, \Pi(F=f|C)\log_2 \pi\left(F=f|\mathcal{F},C\right)\right\} \\
&= \int_0^1 df\, \pi(F=f|\mathcal{F},C)\log_2 \pi(F=f|\mathcal{F},C).
\end{aligned} \tag{24}$$

That is, $\overline{\Delta S}$ is just $KL[\pi(F|\mathcal{F},C)\,||\,U(F)]$, where $U(F)$ is the uniform distribution on $[0,1]$. Unsurprisingly, we have $\overline{\Delta S} \geq 0$. Note that it is not the case in general that $\overline{\Delta S} = \Delta S_{True}$, since the weighted averages in Eqs. (20) and (24) are different. The two quantities are "close" only to the extent that the distribution $\pi(F|\mathcal{F},C)$ approaches $\pi(F|C)$. The GPME model fit makes certain approximations (such as statistical independence of the data in $\mathcal{F}$, and the Laplace approximation for the log-density) and assumptions (such as kernel and mean function choice) that in principle may produce model errors that would make $\overline{\Delta S}$ a flawed estimator for ΔS_{True}. As we will see in the case studies below, GPME does a good job of modeling in general, and in practical cases such modeling errors can be acceptably small.

Note also that, despite the fact that $\overline{\Delta S} \geq 0$, it is not the case that the random variable $\Delta S \geq 0$, as can be seen from Eq. (21). The random variable ΔS itself may certainly attain negative values for some realizations $\pi(F|C)$ of $\Pi(F|C)$, just as ΔS_{True} could turn out in principle to be negative if $\pi(F|\mathcal{F}, C)$ badly mis-estimates $\pi(F|C)$. It would be reassuring to have some measure of how unlikely this is for $\Delta S < 0$. This can be obtained from the variance of ΔS.

Here, we reproduce the GPME expression for $\mathrm{Var}(\Delta S)$, given in Eq. (A.58):

$$\begin{aligned}\mathrm{Var}(\Delta S) &= \int_0^1 df_1 \int_0^1 df_2 \left[\pi(F=f_1|\mathcal{F},C)\log_2 \pi(F=f_1|\mathcal{F},C)\right] \\ &\qquad \times \left[\pi(F=f_2|\mathcal{F},C)\log_2 \pi(F=f_2|\mathcal{F},C)\right] \\ &\qquad \times \left[e^{C(f_1,f_2)} - 1\right], \qquad (25)\end{aligned}$$

where $C(f_1, f_2)$ is the GPME posterior covariance over $\ln \Pi(F|C)$. Using this formula, we may compute the uncertainty in the estimate $\overline{\Delta S}$ in terms of a straightforward two-dimensional quadrature on $[0,1]\times[0,1]$. The appendix shows that in the asymptotic limit $N \to \infty$ of unlimited training data, $\mathrm{Var}(\Delta S) \sim N^{-1}$. Thus, the uncertainty in the estimate $\overline{\Delta S}$ scales asymptotically as $O(N^{-1/2})$.

We may use this fact to introduce a *forecast advantage measure* (FAM):

$$\mathrm{FAM} = \frac{\overline{\Delta S}}{\sqrt{\mathrm{Var}\,(\Delta S)}}, \qquad (26)$$

which scales asymptotically as $O(N^{1/2})$. The FAM gives us a measure of a priori confidence in the positivity of the out-of-sample entropy game winnings of the recalibrated forecast.

To summarize the results so far: given an FOA and its associated PIT data $\mathcal{F}$, we have a recalibration procedure that allows us to improve current published forecasts $p(X; J, C)$, replacing them with recalibrated forecasts $p_1(X; J, C)$. The recalibrated forecasts are expected to outperform the original published forecasts in ignorance scores, and, equivalently, in entropy game performance. The extent of this superiority of performance—the average per-round entropy game winnings ΔS_{True}—may be estimated in advance using only the data in $\mathcal{F}$, and that estimate is attended by an uncertainty that likewise can be computed using only the data in $\mathcal{F}$. Our confidence in the positivity of ΔS_{True}—that is, in the superiority of the recalibrated forecast—can be expressed by the expression in Eq. (26) for the FAM, which in the GPME theory grows with the training dataset size N as $N^{1/2}$. Hence, we can reassure ourselves as to the superior performance of the recalibrated forecast by accumulating a sufficiently large training set.

3.5. Fit quality

As was discussed at the end of Section 3.2, success of the density estimation procedure is crucial to the success of the recalibration procedure, since the latter's success depends on the estimated distribution $\pi(F|\mathcal{F}, C)$ not departing too much from the true distribution $\pi(F|C)$. The GPME theory furnishes a quantitative measure of the departure $\pi(F|\mathcal{F}, C)$ from $\pi(F|C)$, through the quantity

$$EI\left[\pi(F|\mathcal{F}, C)\right] \equiv E_{\Pi(F|C)\,|\,\mathcal{F}}\left\{\int_0^1 df\, \Pi(F = f|C) \log_2 \frac{\Pi(F = f|C)}{\pi(F = f|\mathcal{F}, C)}\right\}. \qquad (27)$$

The quantity inside the expectation is the K-L divergence of the imperfectly known distribution $\Pi(F|C)$ from the estimated distribution $\pi(F|\mathcal{F}, C)$. The GPME theory yields a closed-form expression for this fit quality measure, which is given in Eq. (A.50) and reproduced here:

$$EI\left[\pi(F|\mathcal{F}, C)\right] = \frac{1}{2\ln 2}\int_0^1 df\, \pi(F = f|\mathcal{F}, C) \times C(f, f), \qquad (28)$$

where $C(f_1, f_2)$ is the GPME posterior covariance over $\ln \Pi(F|C)$.

Eq. (28) expresses a very sensible result: the expected divergence between the estimated probability density and the true density is proportional to the average of the posterior variance weighted by the effective posterior probability density. As the quality of the fit improves, the variance decreases and takes $EI[\pi(F|\mathcal{F}, C)]$ down with it. Thus, in some sense, $EI[\pi(F|\mathcal{F}, C)]$ expresses the fit quality. However, one must be cautious in interpretation, since no probabilistic interpretation (such as a "p-value") attaches to $EI[\cdot]$, so it is difficult to say in an absolute sense how small a value of $EI[\cdot]$ will be adequate. Moreover, relying on $EI[\cdot]$ for the fit quality is, in effect, asking the model to report on its own success. The result will be conditioned by model assumptions (such as covariance kernel choice), and cannot detect the effects of the model error on the fit quality directly.

It is more useful to employ the fact, derived in the appendix, that, asymptotically, $EI[\pi(F|\mathcal{F}, C)] \rightarrow B/2N$, where B is the number of bins used in the GPME fit and N is the number of datapoints in $\mathcal{F}$. The N^{-1} scaling can be checked by varying the size of the training set $\mathcal{F}$. This behavior can be a useful diagnostic of model adequacy, as we will see below, since departures from this scaling at large values of N may indicate model errors that are masked by noise at smaller values of N.

3.6. Thinning data to remove correlations

The GPME theory has the considerable benefit that, in addition to providing an efficient method for obtaining the predictive distribution $\pi(F|\mathcal{F})$, it also provides closed-form expressions for the desired entropy-related quantities EI, $\overline{\Delta S}$ and Var(ΔS). However, it has one serious defect for our application: it assumes that a point-like Poisson process governs the generation of the PIT values in $\mathcal{F}$. This assumption is more often false than true. In the case of weather, forecast cadences are generally more rapid than the characteristic time in which the dynamic system loses memory (hence the adage that the best predictor of tomorrow's weather is today's weather). This means that successive members of an FOA, both forecasts and observations, typically resemble each other more than do well-separated members. This effect manifests itself in the correlations of successive PIT values in $\mathcal{F}$; examples are displayed in the right panels of Figure 4 in Section 4.2. These correlations technically invalidate the point-like Poisson process assumption that underlies the GPME theory. Two negative consequences of this are that (1) the fit may be skewed, especially if the training set is not large; and (2) the a priori estimates of the recalibrated forecast improvement over the base forecast are not reliable, since they are based on an inaccurate statistical model.

To recover the utility of the GPME theory in such cases, one must *thin* the training dataset by a factor that may be inferred from the autocorrelation function of the PIT values in $\mathcal{F}$. This process is analogous to the thinning of the samples produced by an MCMC chain (Gamerman and Lopes, 2006, p. 149), and is necessary for the same reason: the samples obtained after appropriate thinning have good independence properties, and thus, they may be modeled appropriately by a point Poisson process. The downside is that the thinning of the training set may be harmful to predictive performance if data are not abundant.

One may ask, with some justice, what was the point of emphasizing the non-i.i.d. nature of the recalibration procedure if an i.i.d. restriction is then re-introduced through the GPME fitting procedure. The answer is that GPME only imposes an i.i.d. restriction on the *model training*. The results in Section 3 on the performance improvements of recalibrated forecasts are still valid for non-i.i.d. *forecasts* of future events, given an acceptable, statistically-consistent regression estimate of $\pi(F|\mathcal{F}, C)$ from the GPME procedure. It is quite possible that there may exist a generalization of GPME that takes proper account of non-i.i.d. behavior in $\mathcal{F}$. Locating such a procedure would be a promising avenue for future research, since the result would be a recalibration procedure that is entirely free of the i.i.d. restriction.

Note, however, that the GPME i.i.d. restriction on the training data is necessary *only* for preserving the predictive performance of our recalibration procedure; that is, in order to be able to state in advance the expected improvement in logarithmic forecasting skill. The restriction is *not* necessary for improving the forecast skill by some (possibly difficult to predict) amount. As was demonstrated by Diebold et al. (1999) and Kuleshov et al. (2018), several different styles of regression on the data $\mathcal{F}$ are capable of furnishing estimates of $\pi(F|\mathcal{F}, C)$ that improve the calibration. We can say that the present work advances the state of the art from the work of Diebold et al. (1999) and Kuleshov et al. (2018), in that it is now clear that recalibration may be expected to work even in the case where the F_n are not i.i.d., as long as some reasonable regression model for $\pi(F|\mathcal{F}, C)$ is produced.

4. Verification

We now exhibit practical examples of the forecast recalibration procedure in two separate applications: a laboratory experiment with predictions of the output from a nonlinear circuit and a seasonal meteorology example using ensemble forecasts of El Niño Southern Oscillation (ENSO) temperature fluctuations.

4.1. A nonlinear circuit

Our first application of the forecasting recalibration scheme is a laboratory experiment that includes predictions of the output from a nonlinear circuit. The circuit was first introduced by Machete (2007) and later discussed in more detail by Machete (2013). Machete (2013) and Machete and Smith (2016) discuss different aspects of its predictability properties. The circuit is designed to produce output voltages that mimic the Moore-Spiegel (Moore and Spiegel, 1966) three-dimensional system of ordinary differential equations:

$$\begin{aligned}
\dot{x} &= y \\
\dot{y} &= -y + Rx - \Gamma(x + z) - Rxz^2 \\
\dot{z} &= x.
\end{aligned} \tag{29}$$

This system is a simplified model of a parcel of fluid that is moving vertically in a stratified fluid, with which it exchanges heat, while tethered by a harmonic force to a point (Moore and Spiegel, 1966). The variable z represents the height of the fluid element.

The circuit is set to operate at parameter values $R = 10, \Gamma = 3.6$, at which values the system exhibits chaotic behavior. The voltages (V_1, V_2, V_3) that correspond to the variables (x, y, z) are measured at three points on the circuit. The ODE system of equations in Eq. (29) is scaled to endow its variables with the dimensions of voltage. The voltage V_3, which corresponds to z, is our predictand. As was noted by Machete (2013), the system of equations in Eq. (29) predicts the behavior of the circuit poorly due to model imperfection. An alternative prediction model is constructed using radial basis functions.

Probe voltages for the three voltage probes were collected over a period of 14 hours at a sampling rate of 10 kH. A sample of 2,000 points that corresponds to the z voltage probe was used for the empirical estimation of the climatological distribution $\rho(z)$ of V_3 (corresponding to z) (top-left panel of Figure 1). Then, 2,048 uncorrelated voltage states were sampled in order to furnish initial conditions from which forecasts could be initialized. Each of these states was used to create an ensemble of 127 forecasts by small Gaussian additive perturbations about the observed state and evolving the resulting states using a radial basis function model up to eight time steps ahead, that is, up to a forecast lead time of 0.8 ms. These 0.8 ms lead time forecasts were then converted to probabilistic forecasts for V_3 by kernel dressing and blending with climatology (Bröcker and Smith, 2008), wherein the forecast distribution density is expressed as a sum of kernels, each of which is centered at the value of one of the 127 simulation values, and the result is blended linearly with the climatology $\rho(z)$. The kernels were chosen to be Gaussians, with equal widths that were chosen to minimize the ignorance score, and the linear blending parameter was also chosen to minimize the ignorance score (Machete and Smith, 2016).

The continuous ensemble forecasts thus generated were compared with the corresponding observed values of the z voltage in order to obtain the PIT distribution shown in the top-right panel of Figure 1.

The 2,048 available observations and forecasts were divided into training and test sets, with training set sizes $N_t \in \{200, 283, 400, 566, 800, 1131, 1600\}$ (each about a factor of $\sqrt{2}$ larger than the previous value). In each case, the test set comprised all of the remaining data. We carried out the recalibration procedure with each training set in order to compute the corresponding PIT posterior predictive density $\pi(F|\mathcal{F}, C)$, then carried out entropy games over the corresponding test sets, recording the performance predictors $(EI[\pi(F|\mathcal{F}, C)], \overline{\Delta S}, \text{Var}(\Delta S), \text{FAM})$ and the game outcomes.

The lower-left panel of Figure 1 shows the PIT fit from the training data (red

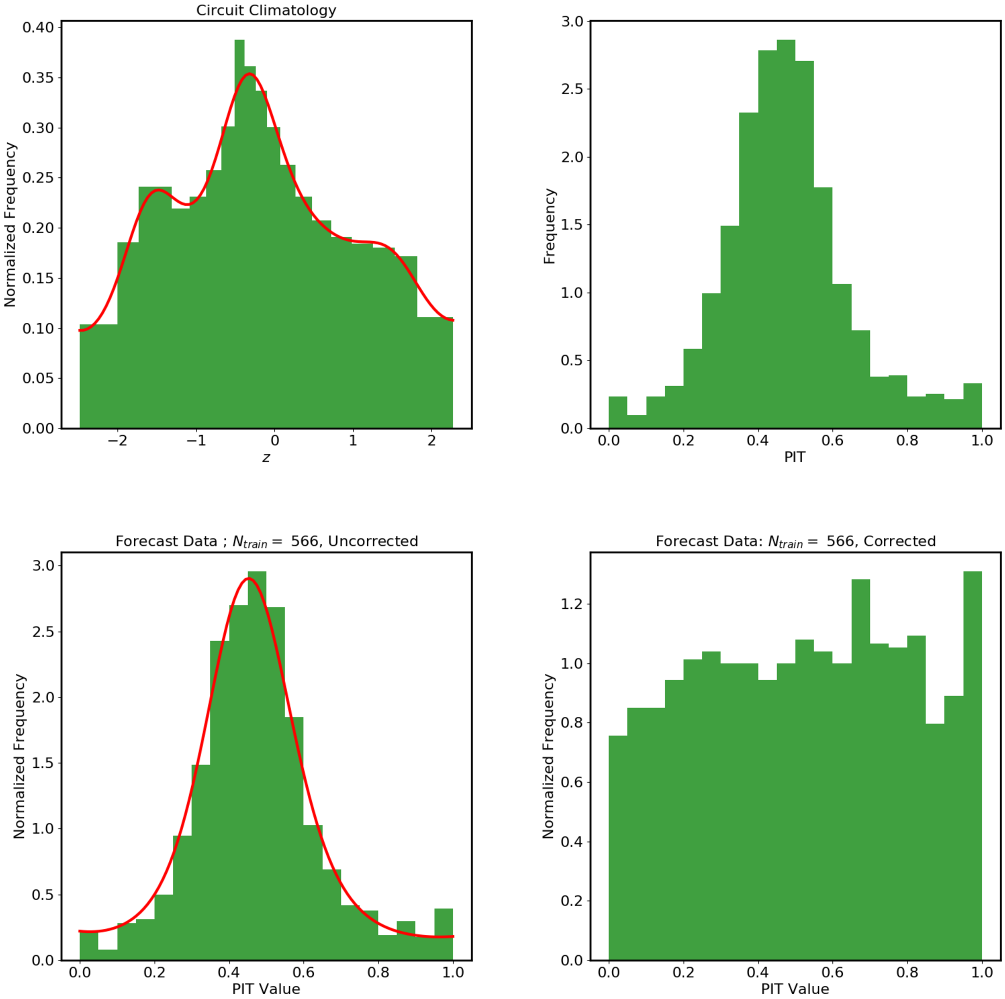

Figure 1: Recalibration of nonlinear circuit ensemble forecasts.

Top left: Climatology histogram. The red line shows the probability density estimate blended into the ensemble forecast. Top right: PIT histogram of all 2,048 ensemble forecasts. The forecasts are clearly overdispersed. Bottom left: PIT histogram of forecasts, with training data excluded, for the case N_t = 566. The red line shows the density inferred from the training data. Bottom right: PIT histogram of recalibrated forecasts for the case N_t = 566. The probabilistic calibration of the recalibrated forecast is excellent, particularly compared with that of the original forecast.

line) superposed on the PIT histogram from the test data for the case $N_t = 566$. The lower-right panel shows the PIT distribution of the recalibrated forecast for the same case. Comparing this figure with the one to its left, we can see that the recalibration procedure was successful in producing updated forecasts that are probabilistically calibrated.

The top-left panel of Figure 2 shows the run of FAM with N_t, displaying the expected $N_t^{1/2}$ trend. The top-right panel of Figure 2 displays the run of $EI[\pi(F|\mathcal{F}, C)]$ with N_t. The initial expected drop appears to level off at the highest values of N_t, possibly indicating some model inadequacy (for example, a poor choice of GP kernel) that reveals itself as the noise in the training histogram is suppressed by larger values of N_t.

The lower-left panel of Figure 2 displays entropy game winnings (red dots) together with the predicted winnings $\overline{\Delta S}$ (blue dots) and predicted uncertainty $\mathrm{Var}(\Delta S)^{1/2}$ (error bars). Here again we see a tendency for the average winnings to depart from the predictions at the highest values of N_t: the actual winnings seem somewhat higher than predicted. Again, it may be possible to explain this discrepancy in terms of inadequacies of the GP model that is used to estimate $\pi(f|\mathcal{F}, C)$, which are perceptible only when the histogram noise abates at higher values of N_t. Nonetheless, the success of the model in predicting the entropy game winnings is gratifying, and the recalibrated model clearly wins systematically against the ensemble forecasts.

The lower-right panel of Figure 2 shows a histogram of the outcomes of 1,482 rounds of the entropy game for the case $N_t = 566$, together with the empirical mean (green dashed line), predicted $\overline{\Delta S}$ (blue solid line) and predicted $1 - \sigma$ interval (pink band), where $\sigma = \sqrt{\mathrm{Var}(\Delta S)}$ is computed using Eq. (25). The distribution of outcomes is quite dispersed, comprising a sharp positive peak with a long tail of negative "bad busts." The shape of the histogram is easy to understand in terms of the GPME fit: see the red line in the lower-left panel of Figure 1. In PIT space, this line approximates the actual PIT distribution, while the original forecast distribution is represented by a horizontal line at a unit normalized frequency. The winnings cutoff at the right of the winnings histogram corresponds to the log of the maximum ratio between these two distributions, which coincides with the mode of the "actual" distribution. This is the reason why the mode of the winnings distribution is at the cutoff. By means of a second-order Taylor series at the mode of the "actual" PIT distribution, we can also show that the expected behavior in the winnings space near the cutoff should approach $(w_{cutoff} - w)^{-1/2}$, which is (integrably) divergent. The long tail to the left is also intelligible, as it corresponds to the two tails near PIT values of 0 and 1

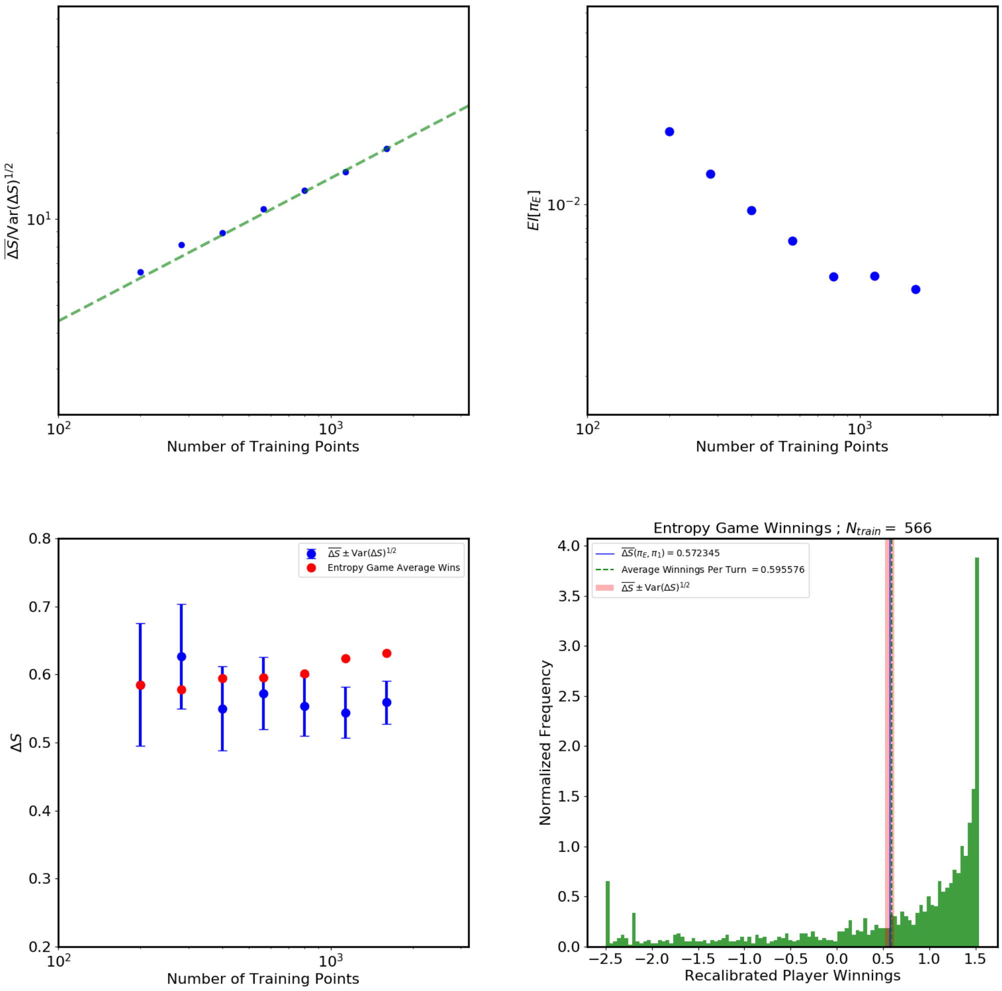

Figure 2: Top left: FAM plot showing the expected $N_t^{1/2}$ trend. Top right: Plot of $EI[\pi(F|\mathcal{F}, C)]$, showing possible evidence of model inadequacy at the largest training set sizes. Lower left: entropy game winnings and a priori predictions. Lower right: Histogram of outcomes of 1,482 rounds of the entropy game, for the case of 566 training samples. The blue line is the prediction $\overline{\Delta S}$, computed from the training data. The pink band that surrounds the blue line is the predicted $1-\sigma$ interval, with $\sigma = \sqrt{\mathrm{Var}(\Delta S)}$ computed using Eq. (25). The green dashed line is the empirical average of the distribution.

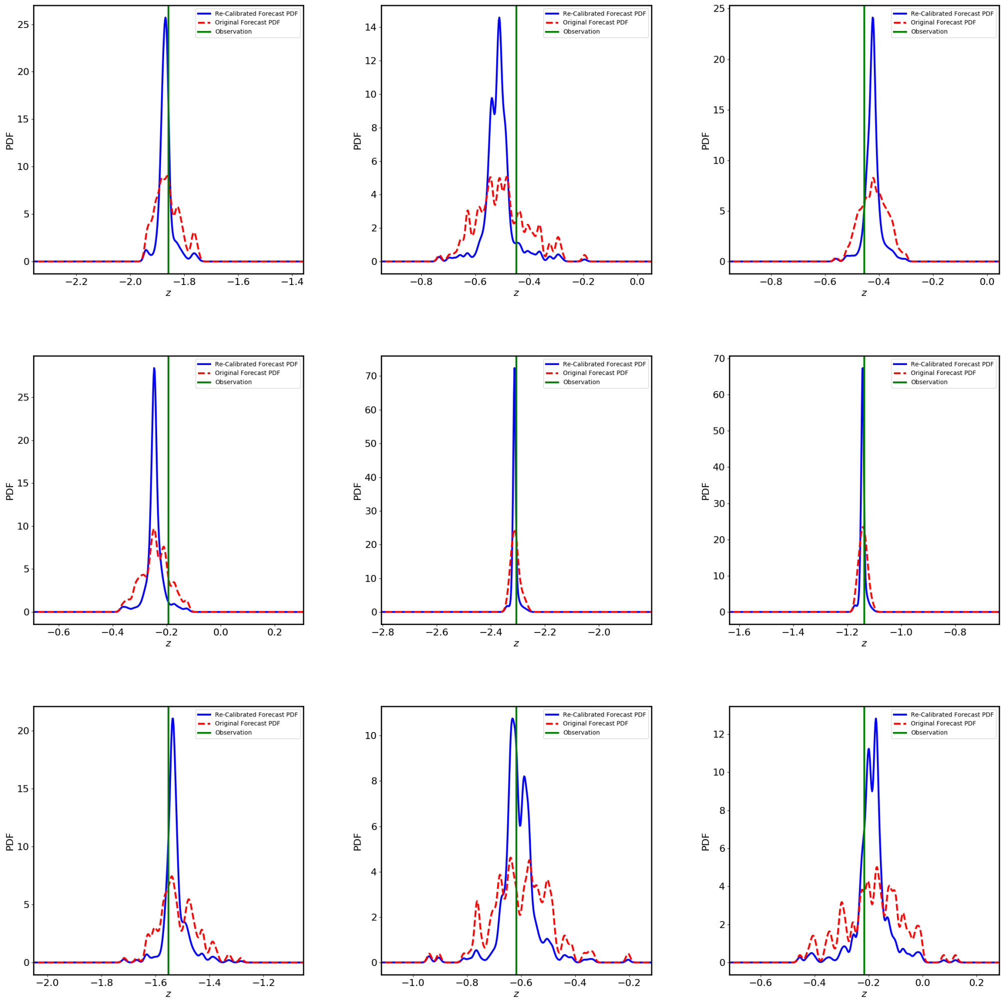

Figure 3: Nonlinear circuit.

The panels show ensemble forecasts, recalibrated forecasts, and observation for the first nine forecasts that follow the training set of 1,600 values. Dashed red curve is the ensemble forecast, solid blue curve is the recalibrated forecast, and green vertical line shows the observation.

where the “actual” distribution is smallest. At these places, the fit distribution density has a value of about 0.2, so the tail should extend to values of w near $w_{tail} = \log_2(0.2/1.0) = -2.3$. Note that these properties are to be expected of overdispersed base forecasts, because of the hump-shaped PIT distribution, but are not expected for, say, underdispersed predictions, where the PIT distribution looks like a pair of peaks near 0 and 1 with a valley in between.

Figure 3 displays, for the case $N_t = 1600$, the first nine ensemble forecasts (red dashed lines), recalibrated forecasts (blue solid lines), and observed z voltages (green vertical lines). The plots show the overdispersion of the ensemble forecasts, which is manifest in the fact that the observations are near the median of the ensemble forecast too frequently. In this case, the recalibrated forecasts are sharper than the ensemble forecasts. This would not be expected in general, but is true here because of the overdispersion of the ensemble forecasts: the relative sharpness of the recalibrated forecasts restores the missing scatter in the PIT distribution. One noteworthy feature of this plot is that the recalibrated forecasts are less noisy than the ensemble forecasts, which are more prone to showing their underlying discrete basis of ensemble-members that anchor the Gaussian mixture model of the continuous ensemble forecast. The transition from $p(x)$ to $p_1(x)$ appears to smooth out this noise somewhat.

In summary, the recalibration procedure is highly successful at improving the performances of the published ensemble forecasts of the nonlinear circuit. The average winnings of about 0.6 bits correspond to a wealth amplification factor of $2^{0.6} = 1.5$ per turn in a Kelly-style betting contest between the two forecasts—offering and wagering on odds on percentiles of the published ensemble forecast, say—which means that the ensemble forecaster would be likely to meet ruin in only a few rounds of betting.

4.2. El Niño temperature fluctuations

Our second application of the recalibration technique is to a seasonal forecasting problem. The seasonal forecast dataset used in this study is from the North American Multimodel Ensemble (NMME) project (Kirtman et al., 2014). The NMME is a collection of global ensemble forecasts from coupled atmosphere-ocean models produced by operational and research centers in the United States and Canada. The NMME forecasts are generated in real time but also include a 30-year set of retrospective monthly forecasts (hindcasts) for assessing systematic biases in the models.

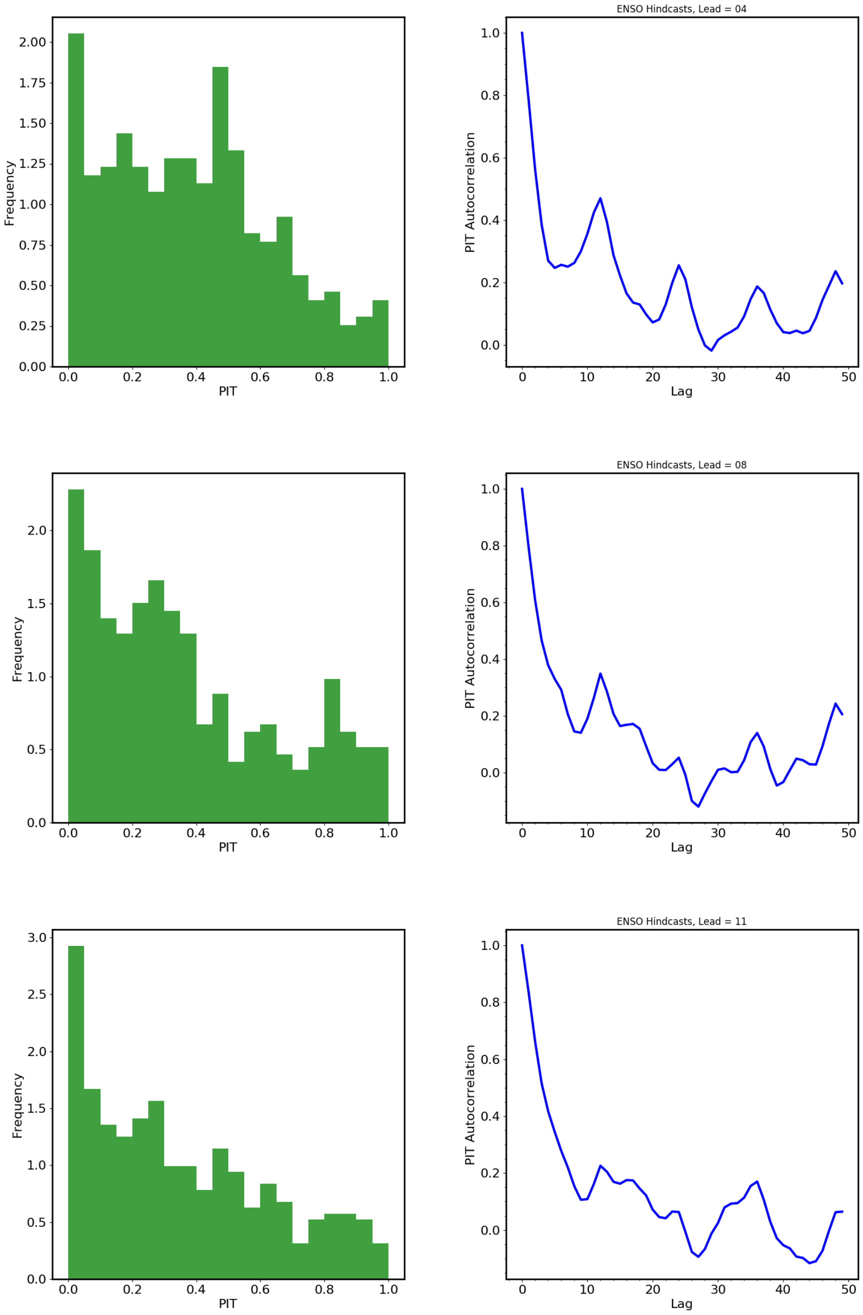

Figure 4: BMA forecasts of the NINO 3.4 index, based on NMME hindcast data.
The panels from top to bottom correspond to forecast lead times of four, eight, and 11 months. Left column: PIT histograms that result from comparisons of BMA forecasts with observations. The BMA forecasts are biased to high values of the NINO 3.4 index. Right column: Time autocorrelation functions of PIT values. Substantial temporal correlations exist out to and past 15-month lags.

Table 1: The NMME models selected for this study, and their respective ensemble sizes.

Model	Ensemble size
COLA-RSMAS-CCSM3	6
COLA-RSMAS-CCSM4	10
GFDL-CM2p1-aer04	10
GFDL-CM2p5-FLOR-A06	12
GFDL-CM2p5-FLOR-B01	12

We examine the NMME model predictions of the intensity of the El Niño Southern Oscillation (ENSO) phenomenon. The ENSO state is often characterized by the NINO 3.4 index, which is the monthly mean sea surface temperature (SST) averaged over the equatorial Pacific region: 5S to 5N, and 170W to 120W. Tippett, Ranganathan, L'Heureux, Barnston, and DelSole (2017) also use the NINO 3.4 index to assess the skill of the NMME models; they provide a useful description of all of the models and their particular configurations for the hindcasts, as well as a discussion of the errors and known problems in the model forecasts. For this study, no data corrections have been made and model climatologies have not been removed: the index values are based on real temperatures instead of temperature anomalies.

The NMME hindcast dataset is available at `http://iridl.ldeo.columbia.edu/SOURCES/.Models/.NMME/`. The NMME hindcasts are created monthly, have lead times that range from 1 to 12 months, and are validated with the observed NINO 3.4 index for the period January 1982 to October 2017. The observed index values are derived from NOAA's optimum interpolation sea surface temperature data (OISST, version 2; Reynolds, Rayner, Smith, Stokes, and Wang, 2002), which are available from `http://www.cpc.ncep.noaa.gov/data/indices/sstoi.indices`.

Our objective in this study is to work with as long a stretch of data as possible, and for that stretch of data to represent a model output that is temporally as homogeneous as possible, because the recalibration procedure could not be expected to be effective if the model composition were to fluctuate during the study, or to be substantially different between the training and test data sets. Of the 15 NMME models, only six were run daily during the entire period of the project, while others were retired at various stages of the project. Of those six, five ran with the same ensemble size throughout, while the remaining model had an ensemble size that varied with sufficient frequency to create concern for the homogeneity of the sample. As a consequence, we split the hindcast data into subsets, choosing only the five models that were run consistently on a monthly basis for 35 years of hindcasts. These models and their respective ensemble sizes are displayed in Table 1.

We converted the forecast simulation ensembles to continuous forecast distributions using the method of Bayesian model averaging (BMA; see Raftery et al., 2005), as adapted to multi-model ensembles with exchangeable members by Fraley et al. (2010). Briefly, BMA models the forecast PDFs that arise from the ensemble predictions by using a mixture model, with the mixture weights ascribed to different components. Ensemble members from a single model are "exchangeable" (Fraley et al., 2010), in that none of them may be regarded as containing better or worse information than their ensemble partners. Thus, they are all assigned equal weights under the scheme. Ensemble members from different models are "nonexchangeable", and so have different weights. Following Fraley et al. (2010) and Raftery et al. (2005), we use Gaussians for the mixture component PDFs, with Gaussian widths that are the same within each model ensemble but may differ from model to model. We first bias-correct the forecasts model by model using a linear regression of the observations on the forecasts in the training set. Then, we center the Gaussians on the bias-corrected ensemble forecast values and optimize the likelihood of the training data iteratively using the EM algorithm (Dempster, Laird, and Rubin, 1977; Tanner, 1993), updating weights and Gaussian widths at each EM iteration (Raftery et al., 2005; Fraley et al., 2010). The converged weights and widths are used to create forecasts in the test set.

We have a total of 430 monthly hindcast simulations available, and consider lead times of 1 to 11 months for each hindcast. We train the BMA forecasting machinery on the first 36 hindcasts and use the machinery to create forecasts from the remaining 394 hindcasts, one for each lead time.

The ensemble forecast results are summarized in Figure 4. The three rows in the figure correspond to lead times of four, eight, and 11 months. The left column depicts the PIT histograms, which show clear evidence that the BMA forecasts are biased to high values of the NINO 3.4 index, despite the preliminary bias corrections. It is clear that there is potential leverage here for the recalibration procedure to do its work. However, there is a fly in the ointment: the right column of Figure 4 displays the temporal autocorrelation functions of the PIT values, which are clearly correlated significantly out to 15-month lags and beyond. While this is not entirely surprising, it is a serious potential restraint on the effectiveness of the method, since the fact that we have only 394 forecasts to work with means that thinning by a factor of 15 leaves hardly enough forecasts to form

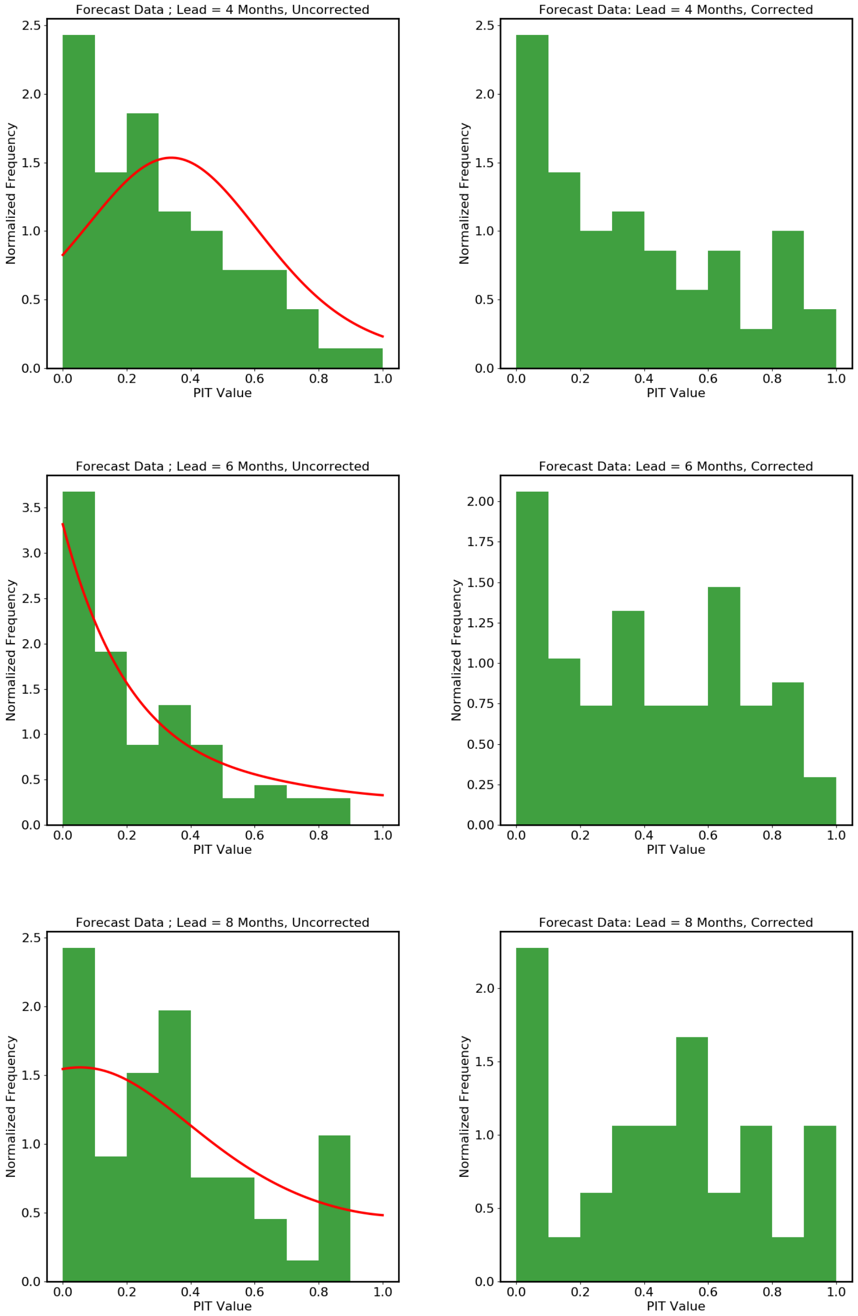

Figure 5: PIT histograms.

Left column: Uncorrected forecasts. The solid red line shows the fit to the training data histogram. Right column: Recalibrated forecasts. From top to bottom, forecast lead times of four, six, and eight months.

a training set, to say nothing of a test set.

We compromise, *faute de mieux*, by thinning by a factor of five. Below this factor, we find that correlations compromise the fit of $\pi(F|\mathcal{F}, C)$ unacceptably; above it, we have too few forecasts to work with. We choose a recalibration training set of 64 forecasts, leaving $394 - 5 \times 64 = 74$ forecasts in the test set. For each lead time, we fit $\pi(F|\mathcal{F}, C)$ to the corresponding PIT histogram and use it first to compute $EI[\pi(F|\mathcal{F}, C)]$, $\overline{\Delta S}$, $\mathrm{Var}(\Delta S)$ and FAM, then to run 74 rounds of the entropy game between the BMA forecasts and the recalibrated forecasts.

Figure 5 shows the result of the recalibration procedure on the PIT histograms for forecast lead times of four, six, and eight months. The 74 test forecasts are presented as histograms with 20 bins. The recalibrated forecasts (right column) improve the probabilistic calibration over the published forecasts (left column). The effect is not as dramatic as for the circuit data in Section 4.1, in part because of the paucity of data and in part no doubt because of model inadequacy due to residual correlations in the data. Nonetheless, the improvement in calibration is clear.

Figure 6 shows a few sample published and recalibrated forecasts for leads of four, six, and eight months. A comparison with Figure 5 shows that the recalibration attempts to correct the bias by shifting the mass of the probability distributions to lower values of the NINO 3.4 index.

The top-left panel of Figure 7 shows FAM as a function of the forecast lead. FAM increases dramatically from a lead of one month to a lead of two months, then settles down to a value of between one and two, suggesting moderate confidence in the performance of the recalibrated forecast for months two and later. The top-right panel shows $EI\left[\pi(F|\mathcal{F}, C)\right]$, which is fairly steady with a shallow peak near a lead of seven months, which indicates that the quality of the fit of $\pi(F|\mathcal{F}, C)$ to the PIT training data is fairly uniform across lead times.

The lower-left panel of Figure 7 shows average winnings in the entropy game (red dots) compared with the performance measures $\overline{\Delta S}$ (blue dots) and $\mathrm{Var}(\Delta S)$ (blue error bars). The performance measures do modestly well in predicting winnings, given the relatively modest size of the available training set and the correlations that still reside therein. The performance advantage of the recalibrated forecasts is striking, especially beginning at a lead of around four months.

The lower-right panel of Figure 7 shows a histogram of outcomes over 74 rounds of the entropy game in the eight-month forecast lead case. Again, the histogram structure can be interpreted in terms of the PIT histogram fit in the middle-

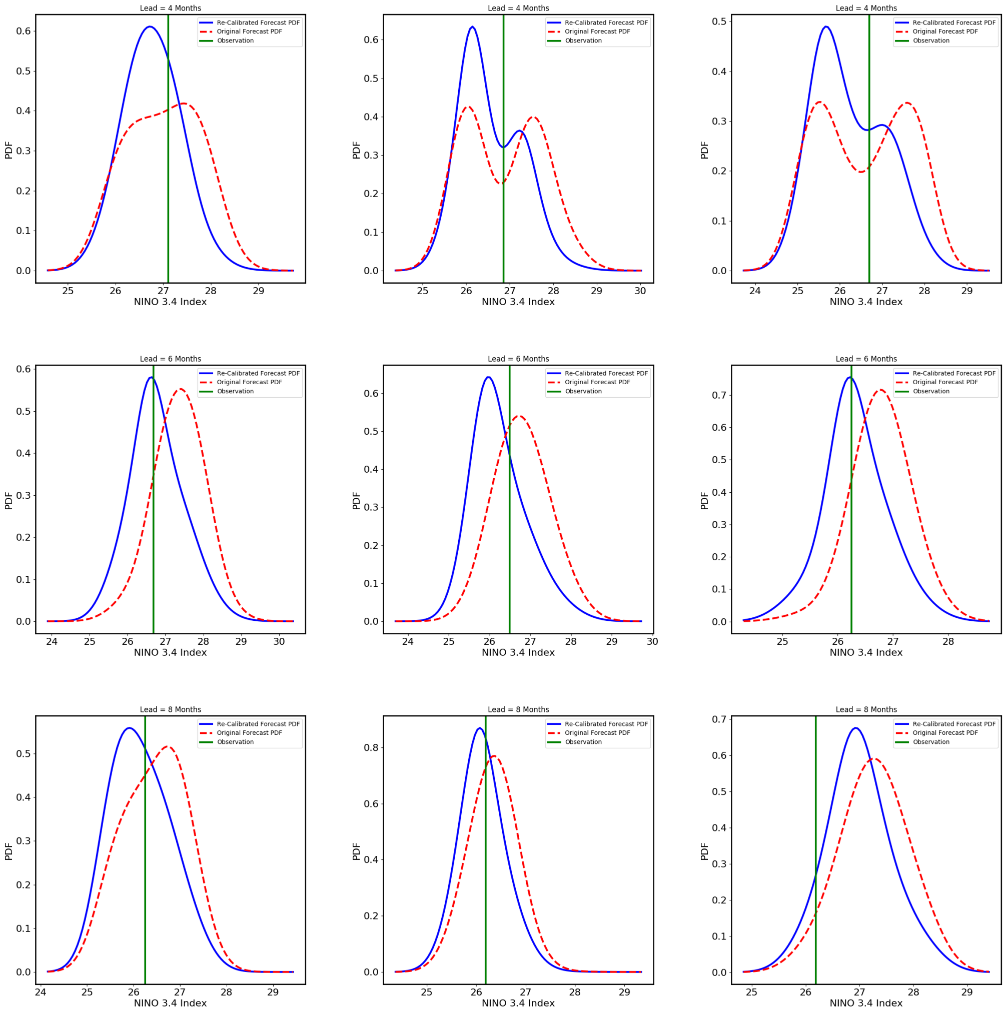

Figure 6: ENSO3.4 forecasts.
The panels show ensemble forecasts, recalibrated forecasts, and observations for the first three forecasts in the set of 74 that comprise the test set. The dashed red curve is the ensemble forecast, the solid blue curve is the recalibrated forecast, and the green vertical line shows the observation. Top row: four-month forecast lead. Middle row: six-month forecast lead. Bottom row: eight-month forecast lead.

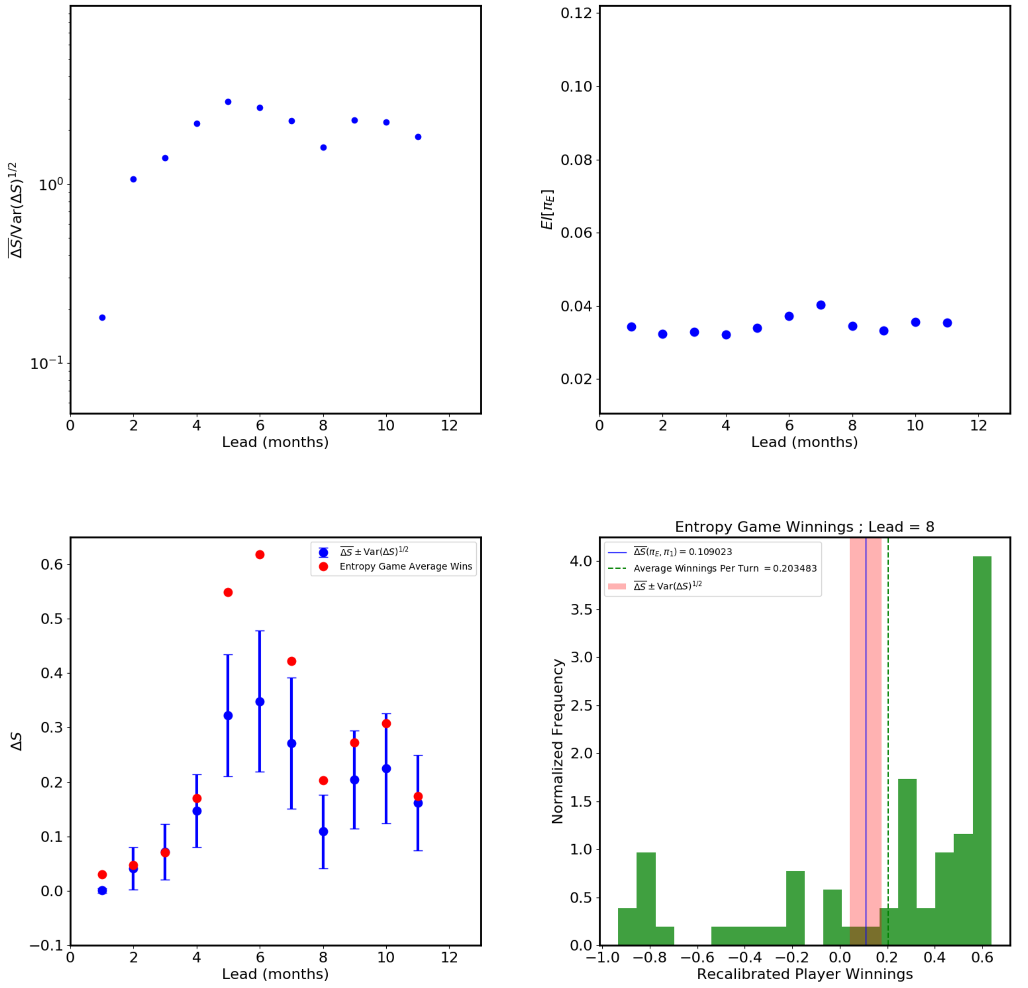

Figure 7: Results of the recalibration of BMA ensemble forecasts of the NINO 3.4 index. Top left panel: FAM plot, now as a function of the forecast lead time. Top right panel: EI plot, again as a function of the lead time. Lower left panel: Expected entropy game winnings (red dots) and predictions (blue dots and error bars) as a function of the lead time. Lower right panel: Histogram of outcomes over 74 rounds of the entropy game for the eight-month lead time case. The green dashed line is the empirical mean, the blue solid line is the predicted mean, and the pink band is the predicted $1 - \sigma$ interval.

left panel of Figure 5, with the mode at $\log_2 1.5 \approx 0.6$ and a tail extending to $\log_2 0.5 = -1$.

The empirical average entropy game winnings are in the range 0.2–0.6 bits, which corresponds to a range of per-turn wealth multipliers of $2^{0.2}$–$2^{0.6}$ = 1.15–1.52 in a Kelly-style odds-making/betting game. Even with the limitations set by the small amount of training data available, the recalibrated forecaster can expect to bankrupt the BMA forecaster after relatively few turns, especially if betting on leads of about six months or so.

5. Discussion

We have gone to great lengths in this work to emphasize the importance of weighing information, and of using quantitative measures of information, when reasoning about probabilistic forecasts of continuous variables. We have shown that the information interpretation of calibration is useful because we may use it to build, out of an arbitrary forecast system, a related, recalibrated system that is expected to be *much* better calibrated than the original system in those cases in which the probabilistic calibration of the original system was noticeably poor.

We validated the forecast recalibration theory on two very different examples: (1) a nonlinear circuit of which the output is forecast by iterating a radial basis function model constructed in delay space (see Machete, 2013, for details) and a smoothing that uses kernel dressing and climatology blending, and (2) 30 years of monthly NINO 3.4 index observations, using forecasts generated from NWP ensembles smoothed by BMA. The nature of the observational data, the input elements to the forecast system, and the method of generating probabilistic forecasts were different in each case, and we emphasize that the recalibration procedure was successful in both cases, producing objectively superior forecasts (as measured by entropy game outcomes) without needing to care much about the inner nature of the forecasts upon which it improves.

In the ENSO study, the recalibrated forecasts were able to outperform the BMA forecasts easily, despite the modest size of the training set. This result is particularly striking in view of the fact that the training was performed on simulations and data spanning about 27 years (after thinning), during which time some secular evolution of the dynamics of the NINO 3.4 index certainly occurred due to carbon forcing, so that the test set that comprised the remaining data necessarily represents a somewhat different climatology from the training set. The success of the method suggests that though the climatology may evolve, the *miscalibration* of the forecast system may be more stable over time, and hence may remain a reliable guide to recalibration.

We showed that the recalibrated forecasts have better ignorance scores than the original published forecasts, and can win bets consistently in the entropy game — a game that, while not fair (because the recalibrated forecast has more information than the original forecast, and hence the player that wields it has an edge), is not biased by its rules toward one player or another in any way. We also pointed out that, while the entropy game is an abstract game, its expected winnings are related directly to the wealth multiplication factor of the player with the recalibrated forecast in Kelly-style odds-setting-and-betting games on outcomes such as percentiles of the original forecast.

For recalibration to work well, much depends on the power and flexibility of the modeling system that is used to fit the training set of PIT values; and for the performance of the recalibrated forecast to be predictable, the modeling system must give access to entropy measures of the distributions being estimated. The Gaussian process measure estimation scheme described in Appendix A provides a highly satisfactory solution for this application, by providing a hyperparametric regression estimate of the PIT measure that yields estimates of Kullback-Leibler divergences from the true distribution.

This is far from saying that further development is unnecessary. In the first place, the studies presented in this work employed only the simplest and most basic GP kernel, the squared exponential, when modeling PIT distributions. In fact, Section 4.1 showed evidence, in the top-right panel of Figure 2, that the $EI[\pi(F|\mathcal{F},C)]$ plot deviates from the expected N_t^{-1} behavior (see Eq. A.51) at the largest training-set sizes, which could indicate a model defect that is masked by noise for smaller training sets. This sort of situation is probably not rare, so it would be worth investigating the effectiveness of more flexible covariances, and possibly mixtures of such covariances.

Furthermore, the GPME model's reliance on i.i.d. training data for creating a regression model $\pi\,(F|\mathcal{F},C)$ qualifies the success of the recalibration procedure in removing the i.i.d. restrictions of the methods described by DHT and KFE. In addition, the necessity of thinning forecasts to create an approximately uncorrelated PIT training sample can be a daunting prospect in cases where the data are not sufficiently abundant to support adequate thinning. Fortunately, an acceptable compromise was found in the ENSO case. Nonetheless, thinning seems an undesirable nuisance imposed by a somewhat simplistic model, namely the log-Gaussian Cox process, which leads to simple closed-form expressions at the cost of requiring the data to be statistically independent. It would be interesting and useful to develop the modeling in such a way as to account for correlations instead of ignoring them, possibly by modeling the PIT distribution in two dimensions, PIT value and time, through the use of a two-dimensional Gaussian process. Other alternatives are certainly worth considering.

This work has also argued from a perspective of forecast quality assessment that emphasizes the importance of decision support. It can often be difficult to interpret forecast skill scores in terms of decision support if one wishes to ascertain the superiority of one forecast set over another, and since different choices of skill score do not agree on a unique sort order of forecast excellence, the process of preferring certain forecast sets over others on the basis of skill has something of a beauty-contest air about it (Smith et al., 2015). We have seen that one can rate the performances of forecast sets concretely, in terms of their ability to consistently win bets against other forecast sets. It would perhaps be well to emphasize this concrete interpretation of skill, since it would seem to translate more directly into actionable decision-making information.

Acknowledgements

The authors wish to thank both referees and this journal's associate editor for critiques that materially strengthened this article, and L. A. Smith for valuable discussions. This material was based upon work supported by the US Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, under Contract DE-AC02-06CH11347. Jennifer Adams acknowledges support from NSF (1338427), NOAA (NA14OAR4310160) and NASA (NNX14AM19G).

The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory ("Argonne"). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan http://energy.gov/downloads/doe-public-access-plan.

Appendix A. Gaussian process probability measure estimation

There is a large body of literature on probability density estimation, along with a number of popular techniques, including kernel density estimation (KDE; see Scott, 2015, Chapter 6), nearest-neighbor estimation (Ivezić, Connolly, VanderPlas, and Gray, 2014, p. 257), and Gaussian mixture modeling (Ivezić et al., 2014, p. 259), as well as more sophisticated methods based on stochastic processes that are adapted well to density estimation, such as Dirichlet processes (Escobar and West, 1995).

For this work, we chose a nonparametric approach based on Gaussian process (GP) modeling. GP modeling is a popular approach for modeling spatial, time series, and spatiotemporal data (Stein, 2012; Cressie and Wikle, 2015), and recently received a lucid introductory treatment by Rasmussen and Williams (2006). Some work applying GP modeling to density estimates from Poisson-process data has appeared recently (Flaxman, Wilson, Neill, Nickisch, and Smola, 2015). We have selected and developed this technique because it is easy to implement and leads to readily-computed closed-form expressions for the Shannon entropies that we require.

Our scheme builds on the same foundation as was described by Flaxman et al. (2015): we model a log-Gaussian Cox process (LGCP), wherein a Gaussian process model is placed on the log-density of an inhomogeneous Poisson process, which is described further below. However, our scheme has a few new features relative to the work described by Flaxman et al. (2015): we show how to approximately normalize the density distributions so that they are in fact approximately probability distributions; we exhibit closed-form expressions for the Shannon entropies that are associated with fit uncertainties in the estimated densities; and we point out a curious—and, to the best of our knowledge, previously unrecognized—feature of this LGCP, namely that the Laplace method approximation to the Poisson likelihood yields a much better approximation to a Gaussian when the log-density is approximated than when the density is approximated directly.

Appendix A.1. Poisson number density estimation

Suppose that we have some i.i.d. sample points from some space. How do we estimate the distribution that gave rise to those points?

More precisely: suppose we that have an absolutely continuous finite measure μ over a set $\Gamma \subset \mathrm{R}^D$, which can be represented by a density $\rho(\boldsymbol{x})$, so that

$d\mu(\boldsymbol{x}) = \rho(\boldsymbol{x}) d^D\boldsymbol{x}$. For any subset $b \subset \Gamma$, we interpret $\mu(b)$ as the mean of a Poisson distribution that describes the number of events of some type that may occur in b. Evidently, $\mu(\Gamma)$ is the total expected number of events in Γ, and $\mu/\mu(\Gamma)$ is a probability measure over Γ, represented by a normalized probability density $\pi(\boldsymbol{x}) = \rho(\boldsymbol{x})/\mu(\Gamma)$. We are given N samples from this probability measure, denoted by $\boldsymbol{x}_k$, $k = 1, \ldots, N$. Suppose that $N \gg 1$. In this situation, which is common in many fields, often one would like a sensible way of estimating μ, or, equivalently, $\rho(\boldsymbol{x})$.

The density $\rho(\boldsymbol{x})$ is known imperfectly, since it must be estimated from the data $\boldsymbol{x}_k$. Just as the statistical estimation of a real-valued scalar quantity v leads naturally to the consideration of a real-valued scalar random variable V of which the possible realizations are values of v, the estimation of a density $\rho(\boldsymbol{x})$ from data leads naturally to the consideration of a density-function-valued random variable $R(\boldsymbol{x})$ of which the realizations are possible density functions $\rho(\boldsymbol{x})$. The simplest nontrivial theory of such stochastic functions is the theory of Gaussian processes (Stein, 2012; Rasmussen and Williams, 2006), which is what we exploit here.

We partition Γ into bins, which will constitute measurable training sets. In order to avoid a coarse binning of the space that smears out any spatial structure in $\rho(\boldsymbol{x})$, the bins should be geometrically small compared to the length scales in the measure. On the other hand, we would like to write down a likelihood for the data that somehow makes use of the Gaussian nature of the GP model. However, the Poisson likelihood, which is appropriate for this kind of data, is acceptably Gaussian in μ (the Poisson mean) only when μ is dispiritingly large, $\mu > 15$ or so. Hence, in principle, our requirement for small bins is in tension with our requirement for populous bins.

Furthermore, a GP model of $\rho(\boldsymbol{x})$ certainly will not respect the positivity constraint $\rho(\boldsymbol{x}) > 0$. It might do so approximately, for certain choices of the kernel function or in regions with abundant sample points, but we would like our model to apply correctly to sparsely sampled regions as well as to crowded ones.

While these obstacles seem considerable, they are not insurmountable. The key element of the estimation procedure described below is a log-Gaussian Cox process (LGCP), which models $\ln \rho$ rather than modeling ρ directly, so that ρ satisfies the positivity constraint automatically. By a stroke of good luck, it turns out that choosing to model $\ln \rho$ as a GP (instead of modeling ρ directly) solves two problems at once. In the first place, such a model automatically satisfies the positivity constraint $\rho(\boldsymbol{x}) > 0$. More subtly, the Laplace approximation to the Poisson likelihood *approaches the Gaussian regime much more rapidly* as a

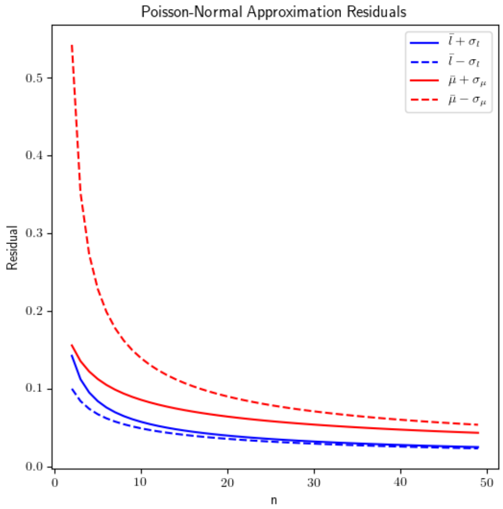

Figure A.8: Poisson-normal approximation residuals.
The blue curves show the residual magnitudes $|R_l(n)|$ at a deviation of $1-\sigma$ above (solid) and below (dashed) the mode $\bar{l}$. The red curves display the analogous behavior for $|R_\mu(n)|$.

function of $\ln\rho$ than it does as a function of ρ!

This is not difficult to demonstrate. Consider a single Poisson variate n with mean $\mu = \exp(l)$. The Poisson likelihood for an observation of n is $\pi = e^{-\mu}\mu^n$. The normal approximation is achieved by expanding $\ln\pi$ in a Taylor series about its maximum. As a function of μ, this is saying

$$\ln\pi = -n + n\ln n - \frac{1}{2}\frac{1}{\sigma_\mu^2}(\mu-\bar{\mu})^2 + R_1(n,\mu), \tag{A.1}$$

where $\bar{\mu} = n$ and $\sigma_\mu^2 = n$, and where we denote the approximation residual by $R_1(n,\mu)$. As a function of l, we may expand around the maximum similarly, obtaining

$$\ln\pi = -n + n\ln n - \frac{1}{2}\frac{1}{\sigma_l^2}(l-\bar{l})^2 + R_2(n,l), \tag{A.2}$$

with $\bar{l} = \ln n$ and $\sigma_l^2 = 1/n$. The functions R_1 and R_2 are defined implicitly by Eqs. (A.1)–(A.2), and may be computed from those formulae directly. The result is plotted in Figure A.8.

The figure shows the values of the residuals computed at their respective $\pm\sigma$ points about their respective means; that is, $R_1(n, \bar{\mu} \pm \sigma_\mu)$ and $R_2(n, \bar{l} \pm \sigma_l)$. It shows that the residual R_2 is considerably smaller than the corresponding R_1 at these characteristic values of the normal distribution to be approximated, especially at the $-\sigma$ points. In fact, the accuracy attained by the Gaussian approximation in μ at $n = 14$ is exceeded at $n = 3$ by the accuracy of the approximation in l. By $n = 10$ or so, the accuracy of the normal approximation to l exceeds that attained by the approximation to μ at $n > 40$. This is reassuring because it suggests that the Laplace approximation error can be controlled well even for small bin counts of the order 5–10.

Now suppose that the space Γ has volume $\int_\Gamma d^D\boldsymbol{x} = \Omega$. We break up the space into B disjoint bins $b_\nu, \nu = 1, \ldots, B$, which satisfy $\bigcup_\nu b_\nu \subset \Gamma$, $b_\nu \bigcap b_\beta = \emptyset$ if $\nu \neq \beta$, $\int_{b_\nu} d^D\boldsymbol{x} \equiv v_\nu$. We associate each bin b_ν with a coordinate label $\boldsymbol{x}_\nu$, which is usually the location of the center of the bin. In what follows, we will assume that the bins are small in the sense that the density $\rho(\boldsymbol{x})$ is approximately constant over each bin.

Since the density is approximately constant over each bin, we may set $\int_{b_\nu} d^D\boldsymbol{x}\, \rho(\boldsymbol{x}) = \rho(\boldsymbol{x}_\nu)v_\nu$. We want to model the log of ρ rather than ρ directly. To do so, we need a reference volume scale in order to nondimensionalize ρ, and the volume Ω will do nicely. We therefore set

$$l_\nu \equiv \ln\left(\rho(\boldsymbol{x}_\nu)\Omega\right), \tag{A.3}$$

from which it follows that

$$\begin{aligned} \rho(\boldsymbol{x}_\nu)v_\nu &= \frac{v_\nu}{\Omega} e^{l_\nu} \\ &\equiv \omega_\nu e^{l_\nu}, \end{aligned} \tag{A.4}$$

which defines the dimensionless volume element $\omega_\nu \equiv v_\nu/\Omega$.

Suppose that we observe n_ν samples from the density in bin b_ν. If we are given the density $\rho(\boldsymbol{x})$, we may then compute the Poisson process likelihood $\mathcal{L}(\boldsymbol{X}|\boldsymbol{l})$ of the data. Setting $\rho(\boldsymbol{x}_\nu) \equiv \rho_\nu$ and $\boldsymbol{X} = \{\boldsymbol{x}_k, k = 1, \ldots, N\}$ for brevity, we have

$$\begin{aligned} \mathcal{L}(\boldsymbol{X}|\boldsymbol{l}) &= \left[\prod_{\nu=1}^{B} \exp\left(-\omega_\nu e^{l_\nu} + n_\nu l_\nu\right) \times \omega_\nu^{n_\nu}\right] \\ &\approx \left[\prod_{\nu=1}^{B} \exp\left(-n_\nu + n_\nu \ln n_\nu - \frac{n_\nu}{2}\left(l_\nu - \ln\left(n_\nu/\omega_\nu\right)\right)^2\right)\right] \\ &= \text{const.} \times \exp\left[-\frac{1}{2}\left(\boldsymbol{l} - \boldsymbol{l}_1\right)^T \boldsymbol{D}^{-1}\left(\boldsymbol{l} - \boldsymbol{l}_1\right)\right], \end{aligned} \tag{A.5}$$

where we have appealed to the normal approximation discussed above, and we define

$$\begin{aligned} \boldsymbol{l} &\equiv [l_1, \ldots l_B]^T, & \text{(A.6)} \\ \boldsymbol{l}_1 &\equiv [\ln(n_1/\omega_1), \ldots, \ln(n_B/\omega_B)]^T, & \text{(A.7)} \\ \boldsymbol{D} &\equiv \mathrm{diag}\left[n_1^{-1}, \ldots, n_B^{-1}\right]. & \text{(A.8)} \end{aligned}$$

It is clear that at this point we have committed to having no empty bins, since an empty bin completely compromises the normal approximation that we have just introduced. In fact, as was discussed above, we will be happy with bin sample counts in the region 5–10.

Note that we have assumed implicitly here that the data are described well by a Poisson point process, and in particular, that the $\boldsymbol{x}_k$ are i.i.d. If this assumption is incorrect, then Eq. (A.5) is not the correct expression for the likelihood. In some cases where the $\boldsymbol{x}_k$ arise from a stationary time series with nonzero correlations $< \boldsymbol{x}_k \boldsymbol{x}_l >= f(|k - l|)$, we may be able to thin out the data by a factor that is suggested by the shape of the autocorrelation function, so as to obtain an approximately uncorrelated data sample that may be modeled correctly using Eq. (A.5).

We will model the function $l(\boldsymbol{x})$ using a constant-mean Gaussian process

$$l \sim GP\left(l_0(\boldsymbol{x}), K(\boldsymbol{x}, \boldsymbol{x}'; \theta)\right), \tag{A.9}$$

where the mean function $l_0(\boldsymbol{x})$ is in fact a constant l_0 and the covariance $K(\boldsymbol{x}, \boldsymbol{x}'; \theta)$ is a positive-definite (as an integral kernel) function, parametrized by some hyperparameters that are denoted by θ. For example, we might choose a stationary kernel with a scale hyperparameter σ and an amplitude hyperparameter A, that is

$$K(\boldsymbol{x}, \boldsymbol{x}') = Ak\left(\frac{\boldsymbol{x} - \boldsymbol{x}'}{\sigma}\right), \tag{A.10}$$

in which case $\theta = (A, \sigma)$. The specific form of the covariance function is not needed here, and it could be chosen in general from the many known valid covariance forms, as seems suited to the type of measure being modeled (Rasmussen and Williams, 2006, Chapter 4).

The choice of a parametrized mean level l_0 is made here because a zero-mean GP (the more usual choice) effectively makes a choice of amplitude scale for ρ that is not selected by the data. It is common to obviate this kind of issue by mean-subtracting the data $\boldsymbol{l}_1$. However, many weighted means of the data in $\boldsymbol{l}_1$

could be chosen for this purpose, and it is not clear a priori whether the usual unweighted mean $\bar{l} = B^{-1} \sum_{\nu=1}^{B} l_\nu$ will be the optimal one of these. Setting the mean level as an adjustable parameter selects a certain weighted mean as the maximum-likelihood estimate (MLE) of l_0 and turns out to be computationally inexpensive, as we will see below.

The covariance matrix $\boldsymbol{Q}$ that arises from the GP model is just the Gram matrix of the covariance function,

$$[\boldsymbol{Q}]_{\nu\nu'} = K(\boldsymbol{x}_\nu, \boldsymbol{x}_{\nu'}; \theta), \tag{A.11}$$

where the indices ν, ν' range over the B bins. The mean vector that arises from the process is $\bar{\boldsymbol{l}} = l_0 \boldsymbol{u}_B$, where $\boldsymbol{u}_B$ is the B-dimensional "one" vector, $[\boldsymbol{u}_B]_\nu = 1$, $\nu = 1, \ldots, B$, and l_0 is a parameter to be estimated, residing in the mean function rather than in the covariance kernel.

In terms of $\boldsymbol{Q}$ and $\bar{\boldsymbol{l}}$, the probability of a density ρ that is represented by B-dimensional vector $\boldsymbol{l}$ is

$$\pi\left(\boldsymbol{l}|I\right) d^B\boldsymbol{l} = (2\pi)^{-N/2} \left[\det \boldsymbol{Q}\right]^{-1/2} \exp\left[-\frac{1}{2}\left(\boldsymbol{l} - l_0\boldsymbol{u}_B\right)^T \boldsymbol{Q}^{-1}\left(\boldsymbol{l} - l_0\boldsymbol{u}_B\right)\right] d^B\boldsymbol{l}, \tag{A.12}$$

where we have symbolically collected in I all conditioning information such as the parameters θ, l_0 and the covariance kernel choices.

The Poisson process likelihood of Eq. (A.5) can be marginalized over the GP distribution for ρ of Eq. (A.12) in order to produce the *marginal likelihood*:

$$\begin{aligned}
\mathcal{L}\left(\boldsymbol{X}|I\right) &= \int d^B\boldsymbol{l}\, \mathcal{L}\left(\boldsymbol{X}|\boldsymbol{l}\right) \pi(\boldsymbol{l}|I) \\
&= (2\pi)^{-N/2} \left[\det \boldsymbol{Q}\right]^{-1/2} \times \int d^B\boldsymbol{l}\, \exp\left[-\frac{1}{2}\left(\boldsymbol{l} - l_0\boldsymbol{u}_B\right)^T \boldsymbol{Q}^{-1}\left(\boldsymbol{l} - l_0\boldsymbol{u}_B\right)\right. \\
&\qquad \left. -\frac{‘1}{2}\left(\boldsymbol{l} - \boldsymbol{l}_1\right)^T \boldsymbol{D}^{-1}\left(\boldsymbol{l} - \boldsymbol{l}_1\right)\right].
\end{aligned} \tag{A.13}$$

Unsurprisingly, this is the form for the marginal likelihood of a GP trained on noisy data $\boldsymbol{l}_1$ with noise covariance $\boldsymbol{D}$. We may therefore take over the standard result for the marginal likelihood (Rasmussen and Williams, 2006, Eq. 2.30) that is adapted for the case of a non-constant mean and heteroskedastic noise:

$$\mathcal{L}\left(\boldsymbol{X}|I\right) = \text{const.} \times \left[\det\left(\boldsymbol{Q} + \boldsymbol{D}\right)\right]^{-1/2} \exp\left[-\frac{1}{2}w\right], \tag{A.14}$$

where

$$w \equiv (\boldsymbol{l}_1 - l_0 \boldsymbol{u}_B)\left(\boldsymbol{Q} + \boldsymbol{D}\right)^{-1}(\boldsymbol{l}_1 - l_0 \boldsymbol{u}_B)\,. \tag{A.15}$$

We may use Eq. (A.14) to obtain an MLE of l_0 as a function of the kernel parameters θ. We define the function $\tilde{l}_0(\theta)$ by

$$\tilde{l}_0(\theta) = \frac{\boldsymbol{l}_1^T\left(\boldsymbol{Q} + \boldsymbol{D}\right)^{-1}\boldsymbol{u}_B}{\boldsymbol{u}_B^T\left(\boldsymbol{Q} + \boldsymbol{D}\right)^{-1}\boldsymbol{u}_B}. \tag{A.16}$$

Then, $l_0 = \tilde{l}_0(\theta)$ is the conditional MLE of l_0 given θ. Defining $\boldsymbol{g} \equiv \left(\boldsymbol{Q} + \boldsymbol{D}\right)^{-1}\boldsymbol{u}_B$, we may write

$$\tilde{l}_0(\theta) = \frac{\sum_{\nu=1}^{B} l_{1\nu} g_\nu}{\sum_{\nu-1}^{B} g_\nu}, \tag{A.17}$$

which shows that the MLE of l_0 is in fact a weighted average of $\boldsymbol{l}_1$, as was asserted earlier.

Combining Eqs. (A.16) and (A.15), we obtain

$$w\Bigg|_{l_0=\tilde{l}_0(\theta)} = \boldsymbol{l}_1^T\left(\boldsymbol{Q} + \boldsymbol{D}\right)^{-1}\boldsymbol{l}_1 - \frac{\left[\boldsymbol{l}_1^T\left(\boldsymbol{Q} + \boldsymbol{D}\right)^{-1}\boldsymbol{u}_B\right]^2}{\boldsymbol{u}_B^T\left(\boldsymbol{Q} + \boldsymbol{D}\right)^{-1}\boldsymbol{u}_B}, \tag{A.18}$$

which, by the Schwartz inequality, is non-negative definite and can attain a zero value only when $\boldsymbol{l}_1 = \alpha \boldsymbol{u}_B$ for some scalar α.

Combining Eq. (A.18) with the negative log of Eq. (A.14), we conclude that the MLE of θ, l_0 are obtained by minimizing the objective function $S(\theta)$:

$$S(\theta) \equiv \ln\det\left(\boldsymbol{Q} + \boldsymbol{D}\right) + \boldsymbol{l}_1^T\left(\boldsymbol{Q} + \boldsymbol{D}\right)^{-1}\boldsymbol{l}_1 - \frac{\left[\boldsymbol{l}_1^T\left(\boldsymbol{Q} + \boldsymbol{D}\right)^{-1}\boldsymbol{u}_B\right]^2}{\boldsymbol{u}_B^T\left(\boldsymbol{Q} + \boldsymbol{D}\right)^{-1}\boldsymbol{u}_B}, \tag{A.19}$$

$$\begin{aligned} \theta^{(MLE)} &= \arg\min_{\theta} S(\theta), && \text{(A.20)} \\ l_0^{(MLE)} &= \tilde{l}_0\left(\theta^{(MLE)}\right). && \text{(A.21)} \end{aligned}$$

Suppose that we would like to estimate the density $\rho(\boldsymbol{y}_a)$ — or, more to the point, the log-density $l_a^{(pred)} \equiv \ln\left(\rho(\boldsymbol{y}_a)\Omega\right)$ — at a set of points $\boldsymbol{y}_a$, $a = 1, \ldots, P$. We may do this by following the standard methodological path of GP regression. We first extend the covariance matrix to an $(B+P)\times(B+P)$ matrix $\hat{\boldsymbol{Q}}$, writing

$$\hat{\boldsymbol{Q}} \equiv \begin{bmatrix} \boldsymbol{Q}^{(pred)} & \boldsymbol{k}^T \\ \boldsymbol{k} & \boldsymbol{Q} \end{bmatrix}, \tag{A.22}$$

with

$$\left[\boldsymbol{Q}^{(pred)}\right]_{ab} \equiv K(\boldsymbol{y}_a, \boldsymbol{y}_b; \theta), \tag{A.23}$$

$$[\boldsymbol{k}]_{\nu a} \equiv K(\boldsymbol{x}_\nu, \boldsymbol{y}_a; \theta). \tag{A.24}$$

We further define $\boldsymbol{l}^{(pred)} = [l_1^{(pred)}, \ldots, l_P^{(pred)}]^T$, and the P-dimensional "one" vector $[\boldsymbol{u}_P]_a = 1$, $a = 1, \ldots, P$. Then, we may take over the standard formula for predictions using a GP trained with noisy data (Rasmussen and Williams, 2006, pp. 16–18), again adapted for a nonzero mean and heteroskedastic noise. That is, $\boldsymbol{l}^{(pred)} \sim \mathcal{N}(\boldsymbol{\lambda}^{(pred)}, \boldsymbol{C}^{(pred)})$, with

$$\boldsymbol{\lambda}^{(pred)} = l_0\boldsymbol{u}_P + \boldsymbol{k}_y^T\,(\boldsymbol{Q}+\boldsymbol{D})^{-1}\,(\boldsymbol{l}_1 - l_0\boldsymbol{u}_B) \tag{A.25}$$

and

$$\boldsymbol{C}^{(pred)} = \boldsymbol{Q}^{(pred)} - \boldsymbol{k}^T\,(\boldsymbol{Q}+\boldsymbol{D})^{-1}\,\boldsymbol{k}. \tag{A.26}$$

This is essentially a new, updated Gaussian process, with a "trained" mean function

$$\lambda(\boldsymbol{x}) = l_0 + \sum_{\nu=1}^{B} K(\boldsymbol{x}_\nu, \boldsymbol{x}; \theta)\left[(\boldsymbol{Q}+\boldsymbol{D})^{-1}\,(\boldsymbol{l}_1 - l_0\boldsymbol{u}_B)\right]_\nu, \tag{A.27}$$

and a "trained" covariance function

$$C(\boldsymbol{x},\boldsymbol{y}) = K(\boldsymbol{x},\boldsymbol{y};\theta) - \sum_{\nu,\mu=1}^{B} K(\boldsymbol{x},\boldsymbol{x}_\mu;\theta)K(\boldsymbol{y},\boldsymbol{x}_\nu;\theta)\left[(\boldsymbol{Q}+\boldsymbol{D})^{-1}\right]_{\nu\mu}. \tag{A.28}$$

Returning to the higher-level "random function" view, we may summarize the story so far as follows. There is an unknown unnormalized density $\rho(\boldsymbol{x})$, from which a set of points $\boldsymbol{X} = \{\boldsymbol{x}_k, k = 1, \ldots, N\}$ is i.i.d. sampled. Since $\rho(\boldsymbol{x})$ is known imperfectly, we represent it by a function-valued random variable $R(\boldsymbol{x})$ and its scaled logarithm $l(\boldsymbol{x}) = \ln(\Omega\rho(\boldsymbol{x}))$ by a function-valued random variable $L(\boldsymbol{x})$ that we estimate by GPME. The realizations of $L(\cdot)$ are possible log-density functions $l(\cdot)$. Similarly, the realizations of the function-valued random variable $R(\cdot) = \Omega^{-1}\exp(L(\cdot))$ are possible density functions $\rho(\cdot)$.

The prior distribution over $L(\cdot)$ is a hierarchical model that features a Gaussian process with a constant mean function l_0 and a covariance function $K(\boldsymbol{x}, \boldsymbol{y}; \theta)$, as well as some prior distribution over l_0, θ that we will not need to specify since we will proceed by maximizing the likelihood with respect to these parameters (when the parameter priors change slowly compared with the likelihood, this is approximately MAP estimation). We represent this prior distribution by the notation $L(\cdot)|I \sim GP[l_0, K(\cdot, \cdot)]$. Training with the data $\boldsymbol{X}$ yields an updated posterior distribution for $L(\cdot) \,|\, (\boldsymbol{X}, I)$, which is a Gaussian process with mean function $\lambda(\boldsymbol{x})$ and covariance function $C(\boldsymbol{x}, \boldsymbol{y})$. Notationally, $L(\cdot) \,|\, (\boldsymbol{X}, I) \sim GP\,[\lambda(\cdot), C(\cdot, \cdot)]$.

So far, we have modeled the imperfectly known log-density $L(\boldsymbol{x})$ rather than $R(\boldsymbol{x})$. This choice has consequences for the inferred Poisson process density, which is not, as one might assume naively, simply a constant times $\exp(\lambda(\boldsymbol{x}))$. In a small volume $v \equiv \Omega\omega$ about a location $\boldsymbol{x}$, the expected number of events n *given* the imperfectly known log-density $L(\boldsymbol{x})$ is

$$vR(\boldsymbol{x}) = \omega \exp\left[L(\boldsymbol{x})\right]. \tag{A.29}$$

Given the GP posterior predictive distribution $L(\cdot) \,|\, \boldsymbol{X} \sim GP[\lambda(\cdot), C(\cdot, \cdot)]$, the effective number density $\rho_E(\boldsymbol{x})$ at $\boldsymbol{x}$ is given by

$$\begin{aligned} \rho_E(\boldsymbol{x}) &= E_{L(\cdot)\,|\,(X,I)}\left\{R(\boldsymbol{x})\right\} \\ &= \Omega^{-1}\left(2\pi C(\boldsymbol{x}, \boldsymbol{x})\right)^{-1/2} \int dl\ \exp\left\{-\frac{1}{2}\frac{[l - \lambda(\boldsymbol{x})]^2}{C(\boldsymbol{x}, \boldsymbol{x})}\right\} \times \exp(l) \\ &= \Omega^{-1}\exp\left[\lambda(\boldsymbol{x})\right] \times \exp\left[\frac{1}{2}C(\boldsymbol{x}, \boldsymbol{x})\right]. \end{aligned} \tag{A.30}$$

We see that the log expected number of events is shifted with respect to the log-density $l(\boldsymbol{x})$ by the non-constant factor $C(\boldsymbol{x}, \boldsymbol{x})/2$.

Appendix A.2. Probability density estimation

The posterior predictive probability density $\pi(\boldsymbol{x}|\boldsymbol{X}, I)$ that a future event should occur within a differential volume $d^D\boldsymbol{x}$ of $\boldsymbol{x}$ is the normalized version of $\rho_E(\boldsymbol{x})$:

$$\pi(\boldsymbol{x}|\boldsymbol{X}, I) = \rho_E(\boldsymbol{x}) / \int d^D\boldsymbol{x}\, \rho_E(\boldsymbol{x}), \tag{A.31}$$

where the notation $\pi(\boldsymbol{x}|\boldsymbol{X}, I)$ will be justified below. In the asymptotic limit $N \longrightarrow \infty$, the normalization constant may be estimated from the data directly. We replace the integral by a sum over the training bins, in effect selecting the same prediction points as training bin centers:

$$A \equiv \int d^D\boldsymbol{x}\, \rho_E(\boldsymbol{x}) \approx \sum_{\nu=1}^{B} \omega_\nu \exp\{\lambda_\nu\} \times \exp \frac{1}{2} C_{\nu\nu}, \tag{A.32}$$

where

$$\begin{aligned} \boldsymbol{\lambda} &= l_0 \boldsymbol{u}_B + \boldsymbol{Q}\,(\boldsymbol{Q}+\boldsymbol{D})^{-1}\,(\boldsymbol{l}_1 - l_0 \boldsymbol{u}_B) \\ &= \boldsymbol{l}_1 - \boldsymbol{D}(\boldsymbol{Q}+\boldsymbol{D})^{-1}\,(\boldsymbol{l}_1 - l_0 \boldsymbol{u}_B) \end{aligned} \tag{A.33}$$

and

$$\begin{aligned} \boldsymbol{C} &= \boldsymbol{Q} - \boldsymbol{Q}\,(\boldsymbol{Q}+\boldsymbol{D})^{-1}\,\boldsymbol{Q} \\ &= \boldsymbol{D} - \boldsymbol{D}\,(\boldsymbol{Q}+\boldsymbol{D})^{-1}\,\boldsymbol{D}. \end{aligned} \tag{A.34}$$

In the asymptotic limit, $\boldsymbol{D} \longrightarrow 0$, and we have $\boldsymbol{\lambda} \approx \boldsymbol{l}_1$, $\boldsymbol{C} \approx 0$. Since $[\boldsymbol{l}_1]_\nu = \ln \frac{n_\nu}{\omega_\nu}$, we have

$$\begin{aligned} A &\approx \sum_{\nu=1}^{B} \omega_\nu \times \left(\frac{n_\nu}{\omega_\nu}\right) \\ &= N, \end{aligned} \tag{A.35}$$

which is an entirely unsurprising result.

Combining Eqs. (A.30), (A.31), and (A.32), we may write

$$\pi(\boldsymbol{x}|\boldsymbol{X}, I) = (A\Omega)^{-1} \exp\left(\lambda(\boldsymbol{x}) + \frac{1}{2} C(\boldsymbol{x}, \boldsymbol{x})\right). \tag{A.36}$$

We have seen that GPME furnishes a tractable posterior distribution over $R(\boldsymbol{x})$, the imperfectly known Poisson number density that estimates $\rho(\boldsymbol{x})$. Now consider the normalized probability density $\pi(\boldsymbol{x}) = \rho(\boldsymbol{x})/J[\rho]$, where $J[\rho] \equiv \int d^D\boldsymbol{x}\, \rho(\boldsymbol{x})$. We would like to use GPME to obtain a posterior distribution over $\Pi(\boldsymbol{x})$, the imperfectly known probability density (a probability-density-valued random variable) that estimates $\pi(\boldsymbol{x})$.

An unfortunate property of the GPME scheme is that, while it yields a tractable distribution over the log number density $L(\boldsymbol{x})$, it does not yield a tractable distribution over the probability density $\ln \Pi(\boldsymbol{x})$. The reason for this is that the transformation from $L(\boldsymbol{x})$ to $\ln \Pi(\boldsymbol{x})$ is nonlinear and does not transform the Gaussian distribution in L into another tractable distribution over either Π or $\ln \Pi$. The normalization factor A discussed above pertains to the effective number density $\rho_E(\boldsymbol{x})$, which, according to the first line of Eq. (A.30), is the expectation of the density $R(\boldsymbol{x})$ over the posterior distribution of the GP. This factor allows us to transition from the expected density to the posterior predictive probability density $\pi(\boldsymbol{x}|\boldsymbol{X}, I)$. In general, the factor A is not the normalization that is appropriate to the imperfectly known density $R(\boldsymbol{x})$.

However, if we are satisfied with approximate normalization, the factor A *is* an appropriate normalization. As we now show, the reason for this is that, in the asymptotic limit $E_{L(\cdot)\,|\,X}\left\{J\right\} = A \approx N$, $\mathrm{Var}_{L(\cdot)\,|\,X}\left\{J\right\} \approx N$, so that $E_{L(\cdot)\,|\,X}\left\{J\right\} / \sqrt{\mathrm{Var}_{L(\cdot)\,|\,X}\left\{J\right\}} \approx N^{-1/2}$. As a consequence, for large N, only a small error is committed by replacing $J[\rho]$ with A, and we may set

$$\Pi(\boldsymbol{x}) \approx A^{-1} R(\boldsymbol{x}) = (A\Omega)^{-1} e^{L(\boldsymbol{x})}. \tag{A.37}$$

This amounts to a constant offset of $\ln \Pi$ from L, so that the Gaussian distribution over $L(\boldsymbol{x})$ is simply mean-shifted by $-\ln(A\Omega)$ to produce the Gaussian distribution over $\ln \Pi$.

We show the required expectations by again approximating the integral J, a random variable, by the sum over observed bins,

$$\begin{aligned} J &= \int d^D\boldsymbol{x}\, R(\boldsymbol{x}) \\ &\approx \sum_{\nu=1}^{B} \omega_\nu e^{L(\boldsymbol{x}_\nu)}, \end{aligned} \tag{A.38}$$

so that

$$\begin{aligned} E_{L(\cdot)\,|\,X}\left\{J\right\} &\approx \sum_{\nu=1}^{B} \omega_\nu E_{L(\cdot)\,|\,X}\left\{e^{L(\boldsymbol{x}_\nu)}\right\} \\ &= \sum_{\nu=1}^{B} \omega_\nu \exp\left[\lambda(\boldsymbol{x})\right] \times \exp\left[\frac{1}{2} C(\boldsymbol{x}, \boldsymbol{x})\right] \\ &= A. \end{aligned} \tag{A.39}$$

Furthermore,

$$E_{L(\cdot)\,|\,X}\left\{J^2\right\} \approx \sum_{\nu=1}^{B}\sum_{\mu=1}^{B}\omega_\nu\omega_\mu E_{L(\cdot)\,|\,X}\left\{e^{L(\boldsymbol{x}_\mu)+L(\boldsymbol{x}_\nu)}\right\}. \tag{A.40}$$

Defining the B-dimensional vector $\boldsymbol{m}$ by

$$[\boldsymbol{m}(\mu,\nu)]_\sigma = \delta_{\nu\sigma} + \delta_{\mu\sigma}, \tag{A.41}$$

we may write this as

$$\begin{aligned}
E_{L(\cdot)\,|\,X}\left\{J^2\right\} &\approx \sum_{\nu=1}^{B}\sum_{\mu=1}^{B}\omega_\nu\omega_\mu(2\pi)^{-B/2}\left(\det\boldsymbol{C}\right)^{-1/2} \\
&\quad\times\int d^B\boldsymbol{l}\,\exp\left\{-\frac{1}{2}\left(\boldsymbol{l}-\boldsymbol{\lambda}\right)^T\boldsymbol{C}^{-1}(\boldsymbol{l}-\boldsymbol{\lambda})+\boldsymbol{m}(\mu,\nu)^T\boldsymbol{l}\right\} \\
&= \sum_{\nu=1}^{B}\sum_{\mu=1}^{B}\omega_\nu\omega_\mu\exp\left[\boldsymbol{m}(\mu,\nu)^T\boldsymbol{\lambda}\right]\times\exp\left[\frac{1}{2}\boldsymbol{m}(\mu,\nu)^T\boldsymbol{C}\boldsymbol{m}(\mu,\nu)\right] \\
&= \sum_{\nu=1}^{B}\sum_{\mu=1}^{B}\omega_\nu\omega_\mu\exp\left[\lambda_\nu+\lambda_\mu\right]\times\exp\left[\frac{1}{2}C(\boldsymbol{x}_\nu,\boldsymbol{x}_\nu)+\frac{1}{2}C(\boldsymbol{x}_\mu,\boldsymbol{x}_\mu)+C(\boldsymbol{x}_\nu,\boldsymbol{x}_\mu)\right].
\end{aligned} \tag{A.42}$$

We then have

$$\begin{aligned}
\mathrm{Var}_{L(\cdot)\,|\,X}\left\{J\right\} &= E_{L(\cdot)\,|\,X}\left\{J^2\right\}-\left[E_{L(\cdot)\,|\,X}\left\{J\right\}\right]^2 \\
&\approx \sum_{\nu=1}^{B}\sum_{\mu=1}^{B}\omega_\nu\omega_\mu\exp\left[\lambda_\nu+\lambda_\mu\right] \\
&\quad\times\exp\left[\frac{1}{2}C(\boldsymbol{x}_\nu,\boldsymbol{x}_\nu)+\frac{1}{2}C(\boldsymbol{x}_\mu,\boldsymbol{x}_\mu)\right]\left\{\exp\left[C(\boldsymbol{x}_\nu,\boldsymbol{x}_\mu)\right]-1\right\}.
\end{aligned} \tag{A.43}$$

In the asymptotic limit, by Eq. (A.34), $\boldsymbol{C}\longrightarrow\boldsymbol{D}$, which is diagonal and has small matrix elements $1/n_\nu$, so that

$$\begin{aligned}
\mathrm{Var}_{L(\cdot)\,|\,X}\left\{J\right\} &\approx \sum_{\nu=1}^{B}\sum_{\mu=1}^{B}\omega_\nu\omega_\mu \exp\left[\lambda_\nu+\lambda_\mu\right]\times\exp\left[\frac{1}{2}C(\boldsymbol{x}_\nu,\boldsymbol{x}_\nu)+\frac{1}{2}C(\boldsymbol{x}_\mu,\boldsymbol{x}_\mu)\right][\boldsymbol{D}]_{\nu\mu} \\
&\approx \sum_{\nu=1}^{B}\omega_\nu^2\times\left(\frac{n_\nu}{\omega_\nu}\right)^2\times n_\nu^{-1} \\
&= N. \qquad \text{(A.44)}
\end{aligned}$$

Eqs. (A.39) and (A.44) are the required relations that allow us to approximate $\Pi(\boldsymbol{x})\approx A^{-1}L(\boldsymbol{x})$ in the asymptotic regime, and hence to approximate the posterior distribution over $\ln\Pi$ by a Gaussian process. In this light, we may cast the posterior predictive distribution $\pi(\boldsymbol{x}|\boldsymbol{X},I)$, defined in Eq. (A.31), as

$$\begin{aligned}
\pi(\boldsymbol{x}|\boldsymbol{X},I) &= \rho_E(\boldsymbol{x})/A \\
&= E_{\Pi\,|\,(X,I)}\left\{\Pi(\boldsymbol{x})\right\}. \qquad \text{(A.45)}
\end{aligned}$$

This equation provides the justification for attaching the notation $\pi(\boldsymbol{x}|\boldsymbol{X},I)$ to the posterior predictive distribution.

Appendix A.3. Entropy estimation

The practical output of the probability density estimation procedure is the posterior predictive probability density $\pi(\boldsymbol{x}|\boldsymbol{X},I)$. One might ask how far this is from the imperfectly known true probability density $\Pi(\boldsymbol{x})$. The Kullback-Leibler divergence between the two distributions is

$$KL\left[\Pi\,||\,\pi(\boldsymbol{x}|\boldsymbol{X},I)\right]=\int d^D\boldsymbol{x}\,\Pi(\boldsymbol{x})\ln\frac{\Pi(\boldsymbol{x})}{\pi(\boldsymbol{x}|\boldsymbol{X},I)}, \qquad \text{(A.46)}$$

which is a random variable that measures the departure of $\Pi(\boldsymbol{x})$ from $\pi(\boldsymbol{x}|\boldsymbol{X},I)$. We may calculate the expected divergence

$$\begin{aligned}
ES\left[\pi\left(\boldsymbol{x}|\boldsymbol{X}\right)\right] &\equiv E_{\Pi\,|\,X}\left\{KL\left[\Pi\,||\,\pi(\boldsymbol{x}|\boldsymbol{X},I)\right]\right\} \\
&= \Omega^{-1}\int d^D\boldsymbol{x}\,E_{\Pi\,|\,X}\left\{A^{-1}e^{L(\boldsymbol{x})}\left(L(\boldsymbol{x})-\ln A\right)\right\} \\
&\quad -\int d^D\boldsymbol{x}\,\pi(\boldsymbol{x}|\boldsymbol{X},I)\ln\left(\Omega\pi(\boldsymbol{x}|\boldsymbol{X},I)\right). \qquad \text{(A.47)}
\end{aligned}$$

We have

$$\begin{aligned} E_{\Pi\,|\,X}\left\{A^{-1}e^{L(\boldsymbol{x})}\right\} &= A^{-1}e^{\lambda(\boldsymbol{x})+\frac{1}{2}C(\boldsymbol{x},\boldsymbol{x})} \\ &= \Omega\pi(\boldsymbol{x}|\boldsymbol{X},I) \end{aligned} \tag{A.48}$$

and

$$\begin{aligned} E_{\Pi\,|\,X}\left\{A^{-1}L(\boldsymbol{x})e^{L(\boldsymbol{x})}\right\} &= A^{-1}\left(2\pi C(\boldsymbol{x},\boldsymbol{x})\right)^{-1/2}\int dl\,\exp\left[-\frac{1}{2}\frac{(l-\lambda(\boldsymbol{x}))^2}{C(\boldsymbol{x},\boldsymbol{x})}\right]\times l\,e^{l} \\ &= \Omega\pi(\boldsymbol{x}|\boldsymbol{X},I)\left[\ln\left(\Omega\pi(\boldsymbol{x}|\boldsymbol{X},I)\right)+\ln A+\frac{1}{2}C(\boldsymbol{x},\boldsymbol{x})\right]. \end{aligned} \tag{A.49}$$

Putting all of this together, we find

$$ES\left[\pi(\boldsymbol{x}|\boldsymbol{X},I)\right]=\int d^D\boldsymbol{x}\,\pi(\boldsymbol{x}|\boldsymbol{X},I)\times\frac{1}{2}C(\boldsymbol{x},\boldsymbol{x}). \tag{A.50}$$

This is an intuitively reasonable result: the expected divergence between the estimated probability density and the true density is proportional to the average of the posterior variance weighted by the posterior predictive density. As the quality of the fit improves, the variance decreases, taking $ES\left[\pi(\boldsymbol{x}|\boldsymbol{X},I)\right]$ down with it.

We may obtain an asymptotic estimate of $ES\left[\pi(\boldsymbol{x}|\boldsymbol{X},I)\right]$ by using the training bins as prediction points and approximating the integral by its finite Riemann sum, as we did above. Then

$$\begin{aligned} ES\left[\pi(\boldsymbol{x}|\boldsymbol{X},I)\right] &\approx \frac{1}{A}\sum_{\nu=1}^{B}\omega_\nu e^{\lambda_\mu+C_{\nu\nu}}\times\frac{1}{2}C_{\nu\nu} \\ &\approx \frac{1}{2N}\sum_{\nu=1}^{B}\omega_\nu\left(\frac{n_\nu}{\omega_\nu}\right)\times\frac{1}{n_\nu} \\ &= B/2N. \end{aligned} \tag{A.51}$$

This asymptotic behavior is reassuring, since its simple dependence on the average number of events per bin is in accordance with intuition. Of course, the rapidity with which the asymptotic result becomes a reasonable approximation depends on how quickly the first term in Eq. (A.34) eclipses the second term as $N\longrightarrow\infty$, which is to say, on the choice of the covariance kernel function $K(\boldsymbol{x}_1,\boldsymbol{x}_2)$, on the best-fit hyperparameters, and therefore, ultimately, on the data and the distribution that gave rise to it.

Suppose that someone has proposed a different probability density $p(\boldsymbol{x})$ as the source of the data. Is it possible to tell whether $\pi(\boldsymbol{x}|\boldsymbol{X}, I)$ is an improvement on $p(\boldsymbol{x})$?

Define

$$\begin{aligned}\Delta S\left[\pi(\boldsymbol{x}|\boldsymbol{X}, I), p(\boldsymbol{x})\right] &\equiv KL\left[\Pi(\boldsymbol{x})||p(\boldsymbol{x})\right] - KL\left[\Pi(\boldsymbol{x})||\pi(\boldsymbol{x}|\boldsymbol{X}, I)\right].\\ &= \int d^D\boldsymbol{x}\,\Pi(\boldsymbol{x})\,\ln\frac{\pi(\boldsymbol{x}|\boldsymbol{X}, I)}{p(\boldsymbol{x})}. \qquad \text{(A.52)}\end{aligned}$$

The quantity ΔS is a random variable, in consequence of the uncertainty in the imperfectly known distribution $\Pi(\boldsymbol{x})$. Taking the expectation value of Eq. (A.52), we find

$$\overline{\Delta S} = E_{\Pi\,|\,(X,I)}\left\{\Delta S\right\} = KL\left[\pi(\boldsymbol{x}|\boldsymbol{X}, I)||p(\boldsymbol{x})\right]. \qquad \text{(A.53)}$$

By the properties of the entropy, we have $\overline{\Delta S} \geq 0$, with equality holding only if $p(\boldsymbol{x}) = \pi(\boldsymbol{x}|\boldsymbol{X}, I)$ almost everywhere. However, this is not to say that $\pi(\boldsymbol{x}|\boldsymbol{X}, I)$ is always superior to any other distribution $p(\boldsymbol{x})$. (What if p were, in fact, the ideal distribution $\pi(\boldsymbol{x}|I)$?) The quantity ΔS is uncertain and may in fact be negative; $\overline{\Delta S}$ is merely its expected value. The distribution for ΔS is too difficult to compute, but its variance is straightforward to compute:

$$E_{\Pi\,|\,(X,I)}\left\{(\Delta S)^2\right\} = \int d^D\boldsymbol{x}_1 d^D\boldsymbol{x}_2\,\ln\frac{\pi(\boldsymbol{x}_1|\boldsymbol{X})}{p(\boldsymbol{x}_1)}\,\ln\frac{\pi(\boldsymbol{x}_2|\boldsymbol{X})}{p(\boldsymbol{x}_2)}\,E_{\Pi\,|\,X}\left\{\Pi(\boldsymbol{x}_1)\Pi(\boldsymbol{x}_2)\right\}. \qquad \text{(A.54)}$$

However,

$$\begin{aligned}E_{\Pi\,|\,(X,I)}\left\{\Pi(\boldsymbol{x}_1)\Pi(\boldsymbol{x}_2)\right\} &= (A\Omega)^{-2}(2\pi)^{-1}\left(\det C_2\right)^{-1/2}\\ &\quad\times\int d^2\boldsymbol{l}\,\exp\left\{-\frac{1}{2}\left(\boldsymbol{l}-\boldsymbol{\lambda}_2\right)^T C_2^{-1}\left(\boldsymbol{l}-\boldsymbol{\lambda}_2\right)+\boldsymbol{u}_2^T\boldsymbol{l}\right\}, \qquad \text{(A.55)}\end{aligned}$$

where $\boldsymbol{\lambda}_2^T = [\lambda(\boldsymbol{x}_1), \lambda(\boldsymbol{x}_2)]$, $[C_2]_{ij} = C(\boldsymbol{x}_i, \boldsymbol{x}_j)$ with $i, j = 1, 2$, and $\boldsymbol{u}_2$ is the two-dimensional "one" vector. In other words,

$$\begin{aligned}E_{\Pi\,|\,(X,I)}\left\{\Pi(\boldsymbol{x}_1)\Pi(\boldsymbol{x}_2)\right\} &= (A\Omega)^{-2}\exp\left\{\lambda(\boldsymbol{x}_1)+\lambda(\boldsymbol{x}_2)+\frac{1}{2}C(\boldsymbol{x}_1, \boldsymbol{x}_1)+\frac{1}{2}C(\boldsymbol{x}_2, \boldsymbol{x}_2)+C(\boldsymbol{x}_1, \boldsymbol{x}_2)\right\}\\ &= \pi(\boldsymbol{x}_1|\boldsymbol{X}, I)\pi(\boldsymbol{x}_2|\boldsymbol{X}, I)\exp\left\{C(\boldsymbol{x}_1, \boldsymbol{x}_2)\right\}. \qquad \text{(A.56)}\end{aligned}$$

Substituting in Eq. (A.54), we have

$$\begin{aligned} E_{\Pi\,|\,(X,I)}\left\{(\Delta S)^2\right\} &= \int d^D\boldsymbol{x}_1 d^D\boldsymbol{x}_2 \left(\pi(\boldsymbol{x}_1|\boldsymbol{X},I)\ln\frac{\pi(\boldsymbol{x}_1|\boldsymbol{X},I)}{p(\boldsymbol{x}_1)}\right) \\ &\quad \times\left(\pi(\boldsymbol{x}_2|\boldsymbol{X},I)\ln\frac{\pi(\boldsymbol{x}_2|\boldsymbol{X},I)}{p(\boldsymbol{x}_2)}\right)\times\exp\left\{C(\boldsymbol{x}_1,\boldsymbol{x}_2)\right\}. \end{aligned} \tag{A.57}$$

It follows immediately that

$$\begin{aligned} \mathrm{Var}_{\Pi\,|\,(X,I)}\left\{\Delta S\right\} &= \int d^D\boldsymbol{x}_1 d^D\boldsymbol{x}_2 \left(\pi(\boldsymbol{x}_1|\boldsymbol{X},I)\ln\frac{\pi(\boldsymbol{x}_1|\boldsymbol{X},I)}{p(\boldsymbol{x}_1)}\right) \\ &\quad \times\left(\pi(\boldsymbol{x}_2|\boldsymbol{X},I)\ln\frac{\pi(\boldsymbol{x}_2|\boldsymbol{X},I)}{p(\boldsymbol{x}_2)}\right) \\ &\quad \times\left(\exp\left\{C(\boldsymbol{x}_1,\boldsymbol{x}_2)\right\}-1\right). \end{aligned} \tag{A.58}$$

The asymptotic approximations for $\overline{\Delta S}$ and Var(ΔS) are

$$\lim_{N\to\infty}\overline{\Delta S} = \sum_{\nu=1}^{B}\left(\frac{n_\nu}{N}\right)\left[\ln\left(\frac{n_\nu}{N v_\nu}\right)-\ln p(\boldsymbol{x}_\nu)\right], \tag{A.59}$$

$$\lim_{N\to\infty}\mathrm{Var}(\Delta S) = \frac{1}{N}\sum_{\nu=1}^{B}\left(\frac{n_\nu}{N}\right)\left[\ln\left(\frac{n_\nu}{N v_\nu}\right)-\ln\pi_1(\boldsymbol{x}_\nu)\right]^2. \tag{A.60}$$

Since in general n_ν/N tends to a finite value in the limit, we see that Var(ΔS) $\sim \mathcal{O}(N^{-1})$ and tends to zero in the limit. On the other hand, if $p(\boldsymbol{x})$ is misspecified, one expects $\lim_{N\to\infty}\overline{\Delta S}$ to be a finite positive value. It follows that the GP measure estimate $\pi(\boldsymbol{x}|X,I)$ can achieve significant performance improvements over a misspecified distribution $p(\boldsymbol{x})$ in the limit of large data, and that we would expect to be able to exploit this superior performance (in betting against the owner of $p(\boldsymbol{x})$, say) once N is large enough that $\overline{\Delta S}/\sqrt{\mathrm{Var}(S)} \gg 1$.

If $p(\boldsymbol{x})$ is *not* misspecified — that is, if it happens to be the true distribution $\pi(\boldsymbol{x})$ — then we can set $n_\nu/N = \pi(\boldsymbol{x}_\nu)(1+\epsilon)$, where $\epsilon \sim n_\nu^{-1/2}$, in the limit of large N. We therefore have $\overline{\Delta S} \sim \mathcal{O}(N^{1/2})$ in this limit, so that asymptotically $\overline{\Delta S}/\sqrt{\mathrm{Var}(S)}$ tends to a constant.

References

Araújo, M. B., New, M., 2007. Ensemble forecasting of species distributions. Trends in Ecology & Evolution 22 (1), 42–47.

Bengtsson, L., Ghil, M., Källén, E., 1981. Dynamic Meteorology: Data Assimilation Methods. Vol. 36. Springer.

Brier, G. W., 1950. Verification of forecasts expressed in terms of probability. Monthly Weather Review 78 (1), 1–3.

Bröcker, J., Smith, L. A., 2008. From ensemble forecasts to predictive distribution functions. Tellus A 60 (4), 663–678.

Bross, D., Bross, J., 1953. Design for Decision. The Free Press; New York; Collier Macmillan Limited; London.

Casillas-Olvera, G., Bessler, D. A., 2006. Probability forecasting and central bank accountability. Journal of Policy Modeling 28 (2), 223–234.

Coelho, C., Pezzulli, S., Balmaseda, M., Doblas-Reyes, F., Stephenson, D., 2004. Forecast calibration and combination: A simple bayesian approach for enso. Journal of Climate 17 (7), 1504–1516.

Cressie, N., Wikle, C., 2015. Statistics for Spatio-Temporal Data. Wiley.
URL https://books.google.com/books?id=KsDdCgAAQBAJ

Dawid, A. P., 1984. Present position and potential developments: Some personal views: Statistical theory: The prequential approach. Journal of the Royal Statistical Society. Series A (General) 147 (2), 278–292.
URL http://www.jstor.org/stable/2981683

Dawid, A. P., Sebastiani, P., 1999. Coherent dispersion criteria for optimal experimental design. Annals of Statistics, 65–81.

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society. Series B (methodological), 1–38.

Diebold, F. X., Gunther, T. A., Tay, A. S., 1998. Evaluating density forecasts with applications to financial risk management. International Economic Review 39 (4), 863–883.

Diebold, F. X., Hahn, J., Tay, A. S., 1999. Multivariate density forecast evaluation and calibration in financial risk management: high-frequency returns on foreign exchange. Review of Economics and Statistics 81 (4), 661–673.

Dutton, J. A., James, R. P., Ross, J. D., 2013. Calibration and combination of dynamical seasonal forecasts to enhance the value of predicted probabilities for managing risk. Climate Dynamics 40 (11-12), 3089–3105.

Escobar, M. D., West, M., 1995. Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association 90 (430), 577–588.

Flaxman, S., Wilson, A., Neill, D., Nickisch, H., Smola, A., 2015. Fast Kronecker inference in Gaussian processes with non-Gaussian likelihoods. In: Proceedings of the 32nd International Conference on Machine Learning. pp. 607–616.

Fraley, C., Raftery, A. E., Gneiting, T., 2010. Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. Monthly Weather Review 138, 190.

Fritsch, J. M., Carbone, R., 2004. Improving quantitative precipitation forecasts in the warm season: A uswrp research and development strategy. Bulletin of the American Meteorological Society 85 (7), 955–965.

Gamerman, D., Lopes, H., 2006. Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Second Edition. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.
URL https://books.google.com/books?id=yPvECi_L3bwC

Glahn, B., Peroutka, M., Wiedenfeld, J., Wagner, J., Zylstra, G., Schuknecht, B., Jackson, B., 2009. MOS uncertainty estimates in an ensemble framework. Monthly Weather Review 137 (1), 246–268.

Glahn, H. R., Lowry, D. A., 1972. The use of model output statistics (MOS) in objective weather forecasting. Journal of Applied Meteorology 11 (8), 1203–1211.

Gneiting, T., Balabdaoui, F., Raftery, A. E., 2007. Probabilistic forecasts, calibration and sharpness. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69 (2), 243–268.

Gneiting, T., Katzfuss, M., 2014. Probabilistic forecasting. Annual Review of Statistics and Its Application 1 (1), 125–151.
URL https://doi.org/10.1146/annurev-statistics-062713-085831

Gneiting, T., Raftery, A. E., 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102 (477), 359–378.

Gneiting, T., Raftery, A. E., Westveld, A. H., Goldman, T., 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Monthly Weather Review 133, 1098.

Gneiting, T., Ranjan, R., et al., 2013. Combining predictive distributions. Electronic Journal of Statistics 7, 1747–1782.

Good, I. J., 1952. Rational decisions. Journal of the Royal Statistical Society. Series B (Methodological), 107–114.

Greenwood-Nimmo, M., Nguyen, V. H., Shin, Y., 2012. Probabilistic forecasting of output growth, inflation and the balance of trade in a gvar framework. Journal of Applied Econometrics 27 (4), 554–573.

Hagedorn, R., Smith, L. A., 2009. Communicating the value of probabilistic forecasts with weather roulette. Meteorological Applications 16 (2), 143–155.

Hamill, T. M., 2001. Interpretation of rank histograms for verifying ensemble forecasts. Monthly Weather Review 129, 550.

Held, L., Meyer, S., Bracher, J., 2017. Probabilistic forecasting in infectious disease epidemiology: The thirteenth Armitage lecture. bioRxiv, 104000.

Ivezić, Ž., Connolly, A., VanderPlas, J., Gray, A., 2014. Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data. Princeton Series in Modern Observational Astronomy. Princeton University Press.
URL https://books.google.com/books?id=2fM8AQAAQBAJ

Kagan, Y. Y., Jackson, D. D., 2000. Probabilistic forecasting of earthquakes. Geophysical Journal International 143 (2), 438–453.

Kamstra, M., Kennedy, P., Suan, T.-K., 2001. Combining bond rating forecasts using logit. Financial Review 36 (2), 75–96.

Kelly, J. L., July 1956. A new interpretation of information rate. The Bell System Technical Journal 35 (4), 917–926.

Kirtman, B. P., Min, D., Infanti, J. M., James L. Kinter, I., Paolino, D. A., Zhang, Q., van den Dool, H., Saha, S., Mendez, M. P., Becker, E., Peng, P., Tripp, P., Huang, J., DeWitt, D. G., Tippett, M. K., Barnston, A. G., Li, S., Rosati, A., Schubert, S. D., Rienecker, M., Suarez, M., Li, Z. E., Marshak, J., Lim, Y.-K., Tribbia, J., Pegion, K., Merryfield, W. J., Denis, B., Wood, E. F., 2014. The North American Multimodel Ensemble: Phase-1 Seasonal-to-Interannual Prediction; Phase-2 toward Developing Intraseasonal Prediction. Bulletin of the American Meteorological Society 95 (4), 585–601.
URL https://doi.org/10.1175/BAMS-D-12-00050.1

Knutti, R., Sedláček, J., 2013. Robustness and uncertainties in the new cmip5 climate model projections. Nature Climate Change 3 (4), 369–373.

Krzysztofowicz, R., Aug. 2001. The case for probabilistic forecasting in hydrology. Journal of Hydrology 249, 2–9.

Krzysztofowicz, R., 2014. Probabilistic flood forecast: Exact and approximate predictive distributions. Journal of Hydrology 517, 643–651.
URL http://www.sciencedirect.com/science/article/pii/S0022169414003242

Kuleshov, V., Fenner, N., Ermon, S., 10–15 Jul 2018. Accurate uncertainties for deep learning using calibrated regression. In: Dy, J., Krause, A. (Eds.), Proceedings of the 35th International Conference on Machine Learning. Vol. 80 of Proceedings of Machine Learning Research. PMLR, Stockholmsmï¿œssan, Stockholm Sweden, pp. 2796–2804.
URL http://proceedings.mlr.press/v80/kuleshov18a.html

Kullback, S., Leibler, R. A., 03 1951. On information and sufficiency. Ann. Math. Statist. 22 (1), 79–86.
URL https://doi.org/10.1214/aoms/1177729694

Lutz, W., Sanderson, W., Scherbov, S., 2001. The end of world population growth. Nature 412 (6846), 543–545.

Machete, R. L., 2007. Modelling a Moore-Spiegel electronic circuit: The imperfect model scenario. Ph.D. thesis, University of Oxford.
URL https://ora.ox.ac.uk/objects/uuid:0186999b-3e62-4e18-9ca9-9603be0acae2

Machete, R. L., 2013. Model Imperfection and Predicting Predictability. International Journal of Bifurcation and Chaos 23 (8).

Machete, R. L., Smith, L. A., 2016. Demonstrating the value of larger ensembles in forecasting physical systems. Tellus A: Dynamic Meteorology and Oceanography 68 (1), 28393.
URL https://doi.org/10.3402/tellusa.v68.28393

Maciejowska, K., Nowotarski, J., Weron, R., 2016. Probabilistic forecasting of electricity spot prices using factor quantile regression averaging. International Journal of Forecasting 32 (3), 957–965.
URL http://www.sciencedirect.com/science/article/pii/S0169207014001848

Marzocchi, W., Woo, G., 2007. Probabilistic eruption forecasting and the call for an evacuation. Geophysical Research Letters 34, L22310.

Moore, D. W., Spiegel, E. A., 1966. A thermally excited non-linear oscillator. The Astrophysical Journal 143, 871.

Moran, K. R., Fairchild, G., Generous, N., Hickmann, K., Osthus, D., Priedhorsky, R., Hyman, J., Del Valle, S. Y., 2016. Epidemic forecasting is messier than weather forecasting: The role of human behavior and internet data streams in epidemic forecast. The Journal of Infectious Diseases 214 (suppl_4), S404–S408.

Pinson, P., 2012. Very-short-term probabilistic forecasting of wind power with generalized logit–normal distributions. Journal of the Royal Statistical Society: Series C (Applied Statistics) 61 (4), 555–576.

Raftery, A. E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using Bayesian model averaging to calibrate forecast ensembles. Monthly Weather Review 133 (5), 1155–1174.

Rasmussen, C., Williams, C., 2006. Gaussian Processes for Machine Learning. Springer.

Reynolds, R. W., Rayner, N. A., Smith, T. M., Stokes, D. C., Wang, W., 2002. An improved in situ and satellite sst analysis for climate. Journal of Climate 15 (13), 1609–1625.

Roulston, M. S., Smith, L. A., 2002. Evaluating probabilistic forecasts using information theory.

Monthly Weather Review 130 (6), 1653–1660.
URL https://doi.org/10.1175/1520-0493(2002)130<1653:EPFUIT>2.0.CO;2

Scott, D., 2015. Multivariate Density Estimation: Theory, Practice, and Visualization. Wiley Series in Probability and Statistics. Wiley.
URL https://books.google.com/books?id=XZ03BwAAQBAJ

Sharma, A., 2000. Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 1 - a strategy for system predictor identification. Journal of Hydrology 239 (1), 232 – 239.
URL http://www.sciencedirect.com/science/article/pii/S0022169400003462

Smith, L. A., 2016. Integrating information, misinformation and desire: Improved weather-risk management for the energy sector. In: UK Success Stories in Industrial Mathematics. Springer, pp. 289–296.

Smith, L. A., Suckling, E. B., Thompson, E. L., Maynard, T., Du, H., 2015. Towards improving the framework for probabilistic forecast evaluation. Climatic Change 132 (1), 31–45.

Stein, M., 2012. Interpolation of Spatial Data: Some Theory for Kriging. Springer Series in Statistics. Springer New York.
URL https://books.google.com/books?id=aZXwBwAAQBAJ

Stern, H., Davidson, N. E., 2015. Trends in the skill of weather prediction at lead times of 1-14 days. Quarterly Journal of the Royal Meteorological Society 141 (692), 2726–2736.
URL http://dx.doi.org/10.1002/qj.2559

Suckling, E. B., Smith, L. A., 2013. An evaluation of decadal probability forecasts from state-of-the-art climate models. Journal of Climate 26 (23), 9334–9347.

Tanner, M., 1993. Tools for Statistical Inference: Observed Data and Data Augmentation Methods. Lecture Notes in Statistics. Springer New York.
URL https://books.google.com/books?id=SoVtQgAACAAJ

Tippett, M. K., Ranganathan, M., L'Heureux, M., Barnston, A. G., DelSole, T., 2017. Assessing probabilistic predictions of ENSO phase and intensity from the North American Multimodel Ensemble. Climate Dynamics, 1–22.

Zhang, Y., Wang, J., Wang, X., 2014. Review on probabilistic forecasting of wind power generation. Renewable and Sustainable Energy Reviews 32, 255–270.