Designing Multi-model Applications with Surrogate Forecast Systems

Leonard A. Smith[1,2], Hailiang Du[1,3] and Sarah Higgins[1]

[1]Centre for the Analysis of Time Series,

London School of Economics, London WC2A 2AE. UK

[2]Pembroke College, Oxford, UK

[3]Department of Mathematical Sciences,

Durham University, Durham, UK

Email: hailiang.du@durham.ac.uk

*

ABSTRACT

Probabilistic forecasting is common in a wide variety of fields including geoscience, social science and finance. It is sometimes the case that one has multiple probability-forecasts for the same target. How is the information in these multiple nonlinear forecast systems best "combined"? Assuming stationarity, in the limit of a very large forecast-outcome archive, each model-based probability density function can be weighted to form a "multi-model forecast" which will, in expectation, provide at least as much information as the most informative single model forecast system. If one of the forecast systems yields a probability distribution which reflects the distribution from which the outcome will be drawn, Bayesian Model Averaging will identify this forecast system as the preferred system in the limit as the number of forecast-outcome pairs goes to infinity. In many applications, like those of seasonal weather forecasting, data are precious; the archive is often limited to fewer than 2^6 entries. In addition, no perfect model is in hand. It is shown that in this case forming a single "multi-model probabilistic forecast" can be expected to prove misleading. These issues are investigated in the surrogate model (here a forecast system) regime; where using probabilistic forecasts of a simple mathematical system allows many limiting behaviours of forecast systems to be quantified and compared with those under more realistic conditions.

1. Introduction

Forecasters are often faced with an ensemble of model simulations which are to be incorporated into quantitative forecast system and presented as a probabilistic forecast. Indeed, ensembles of initial conditions have been operational in weather centers in both the USA (Kirtman et al. 2014) and Europe (Palmer et al. 2004; Weisheimer et al. 2009) since the early 1990s and there is a significant literature on their interpretation (Raftery et al. 2005; Hoeting et al. 1999; Roulston and Smith 2003; Wang and Bishop 2004; Wilks 2006; Wilks and Hamill 2007). There is significantly less work on the design and interpretation of ensembles over model structures, although such ensembles are formed on weather (TIGGE (Bougeault et al. 2010)), seasonal (ENSEMBLES (Weisheimer et al. 2009)) and climate (CMIP5 (Taylor et al. 2012)) forecast lead times. This paper focuses on the interpretation of multi-model ensembles in situations where data are precious, that is where the forecast-outcome archive is relatively small. Archives for seasonal forecasts fall into this category, typically limited to between 32 and 64 forecast-outcome pairs.[1] At times, the forecaster has only an "ensemble of convenience" composed by collecting forecasts made by various groups for various purposes. Alternatively, multi-model ensembles could be formed in collaboration using an agreed experimental design. This paper was inspired by the ENSEMBLES project (Weisheimer et al. 2009), in which seven seasonal models were run in concert, with nine initial condition simulations under each model (Hewitt and Griggs 2004). Small-archive parameters[2] (SAP) forecast systems are contrasted with large-archive parameters (LAP) forecast systems using the lessons learned in experimental design based on the results originally reported by Higgins (2015).

[1]The observational data available for initialization and evaluation of the forecasts is very different before the satellite era.

[2]Here the parameters refer to the parameters involved in transforming multi-model ensemble into predictive distribution, for example the model weights, dressing and blending parameters (see Appendix A1) and they are estimated from an archive which is sometimes large and sometimes small.

We adopt the surrogate model context, taking relatively simple models of a chaotic dynamical system, then contrasting combinations of model to gain insight in how to build and test multi-model ensembles in a context where the data are not precious and a “perfect model” (the system) is known. In this context a robust experimental design can be worked out. There is, of course, an informal subjective judgement regarding how closely the consideration in the surrogate experiments map back into the real-world experiment. This is illustrated using a relatively simple chaotic dynamical system. Specifically, the challenges posed when evaluation data are precious are illustrated by forecasting a simple one-dimensional system using four imperfect models. A variety of ensemble forecast system designs are considered: the selection of parameters and the relative value of “more” ensemble members from the “best” model are discussed. This consideration is addressed in a new generalization of the surrogate modeling framework (Smith (1992) and references therein); it is effectively a “surrogate forecasting system” approach, of value when practical constructions rule out the use of the actual forecast system of interest, as is often the case. In the large forecast-archive limit, the selection of model weights is shown to be straightforward and the results are robust as expected; when a unique set of weights are not well defined, the results remain robust in terms of predictive performance. It is shown that when the forecast-outcome archive is nontrivial but small, as it is in seasonal forecasting, uncertainty in model weights is large. The parameters of the individual model probabilistic forecasts vary widely between realizations in the SAP case; they do not do so in the LAP case. This does not guarantee that the forecast skill of SAP is significantly inferior to that of LAP, but it is shown that in this case the SAP forecast systems are significantly (several bits) less skillful. The goal of this paper is to refocus attention on this issue, not to claim to have resolved it. When evaluating models which push the limits of computational abilities of the day, one is forced to use systems simpler than those targeted by operational models to investigate ensemble forecasting. And whenever simplified models are employed, there is a question as to whether the results hold in larger (imperfect) models. This question of "Even In Or Only In" (EIOOI) was discussed in Smith and Gilmour (1998).

Turning to the question of forming a multi-model forecast system, it is shown that (a) the model weights assigned given SAP are significantly inferior to those under LAP (and, of course, to the using ideal weights). (b) Estimating the best model in SAP is problematic when the models have similar skill. (c) Multi-model "out-of-sample" performance is often degraded due to the assignment of low (zero) weights to useful models. Potential approaches to this challenge (other than waiting for decades) are discussed. It is not possible, given the current archive, to establish the extent to which these results are relevant. The aim of the paper can only be to suggest a more general experimental design in operational studies which would identify or rule out the concerns quantified above. The paper merely raises a concern to which no exceptions are known, it does not attempt (nor could any paper today succeed) in showing this clear and present challenge to multi-model forecasting that dominates seasonal (or other) operational forecasts. It does, by considering well designed surrogate forecasting systems, provide insight into challenges likely to be faced by any multimodel forecast system of a design similar to the real forecast system of interest.[3]

[3] After reading this section, a reviewer asks if these results are relevant to readers of MWR? Consider the related question: what evidence is in hand that any approach is robust in operational models? Detailed questions of how large an ensemble should be or how a multi-model should be weighted (or even constructed (Du and Smith 2017)) can not be explored with operational models due to the extreme computational cost of such an evaluation. One could not evaluate, say, Figure 13 using operational models. The aim of surrogate modeling is to address such questions and demonstrate the robustness of the results for simpler target systems. The weakness of surrogate forecast systems is interpreting their relation of these results to those of operational systems of interest. The alternative is to have no well quantified and evaluated insight into the robustness at all. Were the results of Hide (1958) and Read (1992) useful to numerical weather forecasting? Were the many systems of mathematical equations constructed by Lorenz (1963, 1995)? Were the circuit studies on ensemble size by Machete and Smith (2016)? Surrogate forecast systems can aid in the design of operational test-beds and support their finding. The answer in our particular case appears to be that they are relevant.

2. From Ensemble(s) to Predictive Distribution

The ENSEMBLES project considered seasonal forecasts from seven different models; an initial condition ensemble of nine members was made for each model and launched four times a year (in the months of February, May, August and November). The maximum lead time was seven months, except for the November launch which extended to 14 months. Details of the project can be found in Alessandri et al. (2011); Doblas-Reyes et al. (2010); Weigel et al. (2008); Hewitt and Griggs (2004); Weisheimer et al. (2009); Smith et al. (2015).

The models are not exchangeable in terms of the performance of their probabilistic forecasts. Construction of predictive functions via kernel dressing and blending with climatology (see Brocker and Smith (2008) and Appendix A1. for mathematical details) for each initial condition ensemble of simulations is discussed in Smith et al. (2015) (under various levels of cross-validation). Note that kernel dressing is not kernel density estimation (Silverman 1986); asked to suggest a reference that clarifies this common confusion of the two procedures, Silverman replied "As for anything in print, this is like asking for something in print that says the earth is round rather than flat? (B. Silverman private communication, 12 Apr 2018 12:54:11). Kernel Dressing does aim to reproduce the imperfect-model distribution from which it was drawn; Kernel Density Estimation always and only attempts to reproduce the distribution from which the ensemble members were drawn. Throughout the current paper, skill is quantified with I.J. Good's logarithmic score (Good 1952; Roulston and Smith 2002); this score is sometimes (and in this paper) referred to as Ignorance (IGN) (Roulston and Smith 2002). As noted in Smith et al. (2015); Du and Smith (2017), IGN is the only proper and local score for continuous variables (Bernardo 1979; Raftery et al. 2005; Brocker and Smith 2006), and is defined by:

$$S(p(y),Y) = -\log_2(p(Y)), \tag{1}$$

where Y is the outcome and $p(Y)$ is the probability of the outcome Y. In practice, given K forecast-outcome pairs $\{(p_i, Y_i) \mid i = 1, \ldots, K\}$, the empirical average Ignorance score of a forecast system is then

$$S_E(p(y), Y) = \frac{1}{K} \sum_{i=1}^{K} -\log_2(p_i(Y_i)), \tag{2}$$

In practice, the skill of a forecast system can be reflected by the Ignorance of the forecast system relative a reference forecast p_{ref}:

$$S_{rel}(p(y), Y) = \frac{1}{K} \sum_{i=1}^{K} -\log_2[(p_i(Y_i))/p_{ref}(Y_i)]. \tag{3}$$

Climatological forecast (climatology) is a commonly used reference forecast in meteorology.

3. Simple Chaotic System Models Pair

Without any suggestion that probabilistic forecasting of a one-dimensional chaotic map reflects the complexity or the dynamics of seasonal forecasting of the Earth System, this paper draws parallels. Parallels between challenges to probabilistic forecasting of scalar outcomes using multiple models with different structural model errors and a small forecast-outcome archive in low-dimensional systems and those in high-dimensional systems. These challenges occur both in low-dimensional systems and in high-dimensional systems. Whether or not suggestions inspired by the low-dimensional case below generalize to high-dimensional cases (or other low-dimensional cases, for that matter), would have to be evaluated on a case-by-case basis. The argument below is that the challenges themselves can be expected in high-dimensional cases, leading to the suggestion that they should be considered in the design of all multi-model forecast experiments.

The system to be forecast throughout this paper is the Moran-Ricker Map (Moran 1950; Ricker 1954) given in Equation 4. Selection of a simple, mathematically defined system allows the option of examining the challenges of a small forecast-outcome archive in the context of results based on very large archives. This is rarely possible for a physical system (see however Machete (2007); Smith et al. (2015)). In this section the mathematical structure of the system and four imperfect models of it are specified. The specific structure of these models reflects a refined experimental design in light of the results of Higgins (2015).

Let $\tilde{x}_t$ be the state of the Moran-Ricker Map at time $t \in \mathbb{Z}$. The evolution of the system state $\tilde{x}$ is given by

$$\tilde{x}_{t+1} = \tilde{x}_t e^{\lambda(1-\tilde{x}_t)}. \tag{4}$$

In the experiments presented in this paper, $\lambda = 3$, where the system is somewhat "less chaotic" than using the value adopted in Sprott (2003) (Figure 1 shows the Lyapunov exponent as a function of system parameter λ (Glendinning and Smith 2013)), in order to ease the construction of models with comparable forecast skill. Define the observation at time t to be $s_t = \tilde{x}_t + \eta_t$, where the observational noise, η_t, is independent and normally distributed ($\eta_t \sim N(0, \sigma^2_{noise})$)[4].

Four one-dimensional deterministic models are constructed, each one being an imperfect model of the Moran-Ricker system. In the experiments presented here, the focus is on designing a multi-model ensemble scheme and effective parameter selection for producing predictive distribution from multiple models. Therefore the imperfect models as well as their parameter values are fixed. These four models share the same state space as the system, and the observations are complete. Note in practice, it is almost always the case that the model state x lies in a different space from the system state $\tilde{x}$. The models are:

- **Model I**, $G_1(x)$, is built by first expanding the exponential term in Equation 4 to the 12^{th} order:

$$x_{t+1} = x(1 + \lambda(1-x) + \frac{1}{2!}(\lambda(1-x))^2 + \cdots + \frac{1}{12!}(\lambda(1-x))^{12}). \tag{5}$$

[4] Observations are restricted to positive values.

The coefficient of each polynomial term is then truncated at the 3^{rd} decimal place:

$$x_{t+1} = x(1+3(1-x)+4.5(1-x)^2+\cdots+0.004(1-x)^{11}+0.001(1-x)^{12}). \tag{6}$$

- **Model II**, $G_2(x)$, is derived by first taking the logarithm of Equation 4 and expanding to the 8^{th} order:

$$logx_{t+1} = logx+\lambda-\lambda x = logx+\lambda-\lambda e^{logx} \tag{7}$$

$$logx_{t+1} = -2logx-\frac{3}{2!}(logx)^2-\frac{3}{3!}(logx)^3-\cdots-\frac{3}{8!}(logx)^8 \tag{8}$$

The coefficient of each polynomial term is then truncated at the 4^{th} decimal place:

$$logx_{t+1} = -2logx-1.5(logx)^2-0.5(logx)^3-\cdots-0.0006(logx)^7-0.0001(logx)^8 \tag{9}$$

- **Model III**, $G_3(x)$, is obtained by expanding the right-hand side of Equation 4 in a Fourier series over the range $0 \le \tilde{x} \le \pi$. This series is then truncated at the 10^{th} order to yield

$$x_{t+1} = \frac{a_0}{2}+\sum_{i=1}^{10}[a_i cos(2ix_t)+b_i sin(2ix_t)],$$

where the coefficients a_i and b_i are obtained by

$$a_i = \frac{2}{\pi}\int_0^{\pi} xe^{\lambda(1-x)}cos(2ix)dx \tag{10}$$

$$b_i = \frac{2}{\pi}\int_0^{\pi} xe^{\lambda(1-x)}sin(2ix)dx \tag{11}$$

- **Model IV**, $G_4(x)$, is obtained by expanding the right-hand side of Equation 4 by Laguerre Polynomials truncated at the 20^{th} term.

$$x_{t+1} = \sum_{i=0}^{20} c_i L_i(x),$$

where $L_i(x) = \sum_{k=0}^{i} \frac{(-1)^k}{k!} \binom{N}{k} x^k$ are the Laguerre Polynomials and the coefficients c_i are obtained by

$$c_i = \int_0^\infty w(x) L_i(x) x e^{\lambda(1-x)} dx \tag{12}$$

with the weighting function $w(x) = e^{-x}$. Laguerre Polynomials are orthogonal and orthonormal.

Notice that the order of the truncation for Model I, II, III and IV differ. These are chosen so that each model represents the system dynamics well and the scales of their forecast skill are comparable. Figure 2 plots the dynamical function of each model together with the system dynamics. Figure 3 presents the histogram of the 1-step model error over 2048 different initial conditions which are uniformly sampled between the minimum and maximum of the Moran-Ricker system. It appears that Model I simulates the system dynamics well except when the initial condition is near the maximum value of the system. For Model II, a large difference between the model dynamics and the system dynamics appears only when the initial condition is near the minimum value obtained by the system. Model III does not match the system dynamics well where $x \gtrsim 1.5$ and where the forward model reaches the maximum value of the map. Model IV matches the system less well for initial conditions near the maximum value of the map.

Figure 4 plots the two-step model error for each model, while Figure 5 presents the histogram of the 2-step model error. Generally, the structure of the model error is different. Different models have different scales of model error in different local state space.

Again, there is, of course, no suggestion that the Moran-Ricker system resembles the dynamics of the Earth. Rather, the framework presented here (and in Higgins (2015)) provides probabilistic forecasts from structurally flawed models; the model-based forecasts (and ideal probabilistic forecasts formed using the perfect model) differ nontrivially from each other, and as the models are nonlinear the forecast distributions are non-Gaussian. It is these challenges to multi-model forecast system development that are illustrated in this paper, which should (of course) not be taken to present an actual geophysical forecast system; indeed the verifications in the observational record rules out examination of LAP in geophysical systems, while computational requirements rule out extensive examination of SAP in "state-of-the-art" geophysical models.

4. Ensemble Formation

a. Initial Condition Ensembles for Each Model

In the experiments presented in this paper, each model produces ensemble forecasts by iterating an ensemble of initial conditions (IC). The initial condition ensemble (ICE) is formed by perturbing the observation with random draws from a Normal distribution, $N(0, \kappa_\tau^2)$. If the model were perfect and the observation were exact, κ_τ would be zero; as neither of these conditions is met one does not expect κ_τ to be zero. Such a perturbation parameter κ_τ is chosen to minimize the Ignorance score at lead time τ. When making medium-range forecasts, the European Centre for Medium-Range Weather Forecasts (ECMWF) selects a perturbation size such that the RMS error between the ensemble members and the ensemble mean at a lead time of two days is roughly equal to the RMS of the ensemble mean and the outcome at two days.

In the experiments presented below, each initial condition ensemble will contain $N_e = 9$ members, following the ENSEMBLES protocol. Consider first the case of a large archive, with $N_a = 2048$. For a given κ and lead time τ, the kernel dressing and climatology-blend parameter values are fitted using a training forecast-outcome archive which contains $N_l = 2048$ forecast-outcome pairs. The Ignorance score is then calculated using an independent testing forecast-outcome set which contains $N_t = 2048$ forecast-outcome pairs. Figure 6a shows the optimal perturbation parameter κ for each model varies with lead time.[5] The Ignorance score for each model at different lead time, using the values of κ in Figure 6a, is shown in Figure 6b. The sampling uncertainty across forecast launches is represented by a bootstrap resampling procedure, which resamples the set of forecast Ignorance scores for each model, with replacement. The bootstrap resampling intervals are shown as vertical bars in Figure 6 as a $5-95\%$ interval. As seen in Figure 6a, for each model, the preferred value of κ varies significantly (by about a factor of 2) between different forecast lead times. Defining a N_e-member forecast system requires selecting a specific value of κ for each model. In this paper, the value of κ for each model is chosen by optimizing the forecast Ignorance score at lead time 1. Sensitivity tests have been conducted and the Ignorance score at other lead times is much less sensitive to κ than that at lead time 1. Bias correction in the dressing blending approach is another concern. Hodyss et al. (2016) discussed bias in a real-world context. The dressing blending approach can be generalized by including a shifting parameter (see Brocker and Smith (2008)) to account for model bias. Including the shifting parameter does, in fact, improve the Ignorance score out-of-sample (in each model at almost all lead times) in this case. As the improvement is typically less than one 20th of a bit (sometimes zero), such shifting parameter is not included in the dressing blending throughout the experiments presented in the current paper.

b. On the Number of IC Simulations in Each Ensemble

Forecast system design relies on the knowledge of the relationship between the size of the forecast ensemble and the information content of the forecast (Smith et al. 2015). Usually, the cost of developing a brand new model is tremendously larger than the cost of increasing the number of

[5] As noted by a reviewer, there is uncertainty in the κ values reported in Figure 6a. To quantify this uncertainty, the estimate of κ was bootstrap resampled. The results (not shown) show variation in κ, at lead time 1 always less than 50%, but very little variation in the corresponding Ignorance value for each model.

ensemble members[6]. Furthermore the cost of increasing the ensemble size increases only (nearly) linearly and decreases as technology improves.

As the number of ensemble members increases, the true limitation due to structural model error becomes more apparent. Figure 7 shows that forecast Ignorance varies as ensemble size increases. Improvement from the additional ensemble members can be noted, especially at shorter lead times.

5. Forecast System Design and Model Weighting When Data Are Precious

a. Forecasts With a Large Forecast-outcome Archive

As N_a, the size of the forecast-outcome archive, increases, one expects robust results since large training sets and large testing sets are considered. To examine this, 512 different training sets are produced, each contains 2048 forecast-outcome pairs. For each archive, the kernel width σ and climatology-blend weight α are fitted for each model's forecasts at lead time. Figures 8a and 8b show the fitted values of the dressing parameters and climatology-blend weights. The error bars reflect the central 90^{th} percentile over 512 samples. The variation of the weight assigned to the model appears small. The variation of the fitted kernel width is small at short lead times and large at long lead times. Especially at lead time 5, the fitted value for Model IV has relatively large variation. This, however, does not indicate that the estimate is not robust but suggests the Ignorance score function in the parameter space is relatively flat near the minimum. To demonstrate this, the empirical Ignorance is calculated for each archive of kernel width and climatology-blend weight based on the same testing set (which contains another 2048 forecast-outcome pairs). Figure 8c plots the Ignorance score and its 90^{th} percentile as a function of lead time. Notice the 90^{th} percentile ranges are always very narrow.

---

[6] And financially, the cost falls on the current account not the capital account.

The next two paragraphs echo Smith et al. (2015). There are many ways to combine multiple single model forecast distributions into a single probabilistic (multi-model) forecast distribution (Hagedorn et al. 2005; Brocker and Smith 2008). A simple approach is to treat each model equally and therefore apply equal weight to each individual model (see, for example, Weisheimer et al. (2009)). In general, different models perform differently in terms of forecasts, for example, the ECMWF model significantly outperforms other models in seasonal forecasts (Smith et al. 2015). Therefore, applying non-equal weights to all contributing models might provide more skillful multi-model forecast distribution (see, for example, Rajagopalan et al. (2002)). Following Doblas-Reyes et al. (2005) and Smith et al. (2015), define the combined multi-model forecast distribution to be the weighted linear sum of the constituent distributions:

$$p_{mm} = \sum_i \omega_i p_i, \tag{13}$$

where p_i is the individual forecast distribution from the i^{th} model and ω_i ($\sum_i \omega_i$=1) the corresponding weight. The weighting parameters ω_i may be determined according to their performance in a past forecast-outcome archive. The weights of individual models are expected to vary as a function of lead time.

It is computationally costly and potentially results in ill-fitted model weights, if all the weights are fitted simultaneously. To avoid both issues, a simple iterative approach (Du and Smith 2017) is adopted. For each lead time, the best (in terms of Ignorance) model is first combined with the second-best model to form a combined forecast distribution (by assigning weights to both models that optimize the Ignorance of the combined forecast). The combined forecast distribution is then combined with the third-best model to update the combined forecast distribution. This process is repeated until inclusion of the "worst" model is considered. Note each time a new model is included in the combined model, only two weights need to be assigned. Figure 8d shows the weights assigned to each model as a function of lead time. The cyan line in Figure 8c shows the variation of the Ignorance score for the multi-model forecast given those estimated model weights is very small.

b. Forecast With a Small Forecast-outcome Archive

When given a small forecast-outcome archive (e.g. from a $\sim$ 40-year seasonal forecast-outcome archive), one does not have the luxury of exploring a large collection of independent training and testing sets. Cross-validation is often approached by adopting a leave-one-out approach. The robustness of parameter fitting in such cases is of concern. To examine such robustness, a large number of forecast-outcome archives are considered. Each archive contains the same numbers of forecast-outcome pairs. For each archive, the parameter values are fitted via leave-one-out cross-validation. The distribution of fitted values over these small forecast-outcome archives is then compared with the fitted value from the $N_a = 2048$ forecast-outcome archives above. Figure 9 plots the histograms of the fitted climatology-blend weights given 512 forecast-outcome archives each containing $N_a = 40$ forecast-outcome pairs. Notice that, in most cases, the distributions are very wide although they cover the value fitted given the large training set. There are some cases in which about 90 percent of the estimates are larger or smaller than the values fitted by the large archive, e.g. lead time 1 of Model I and Model II and lead time 4 of Model III and Model IV. It therefore appears that the robustness of fitting varies with lead time and the model. For shorter lead times, however, the weights are more likely to be over-fitted and, for longer lead times, the weights are more likely to be under-fitted. This is because at short lead times the model forecasts are relatively good; only a few forecast systems yield predictions that are worse than the climatological forecast. Small forecast-outcome archives, on the other hand, may not contain any model busts and so often overestimate the weights. The longer lead time case can be explained similarly. Figure 10 plots the histogram of fitted kernel widths. Again, observe that there is much larger variation of the estimates here than when fitting with large forecast-outcome archives.

Poor estimation of the kernel width and climatology-blend weight will cause the forecast to lose skill and appear to underperform out-of-sample (due to inappropriately high expectations). This could, of course, be misinterpreted as climate change. For each of the 512 fitted kernel widths and climatology-blend weights, the Ignorance scores are calculated over the same testing set of 2048 forecast-outcome pairs. Figure 11 plots the histogram of the Ignorance score for each model. Using parameters fitted with small archives often results in significant degrading (~ 1 bit) of the Ignorance score of the forecasts. Correctly blending with the climatological distribution would yield a forecast score which, in expectation, is never worse than the climatology. When the blending parameter is determined using the small archive, however, the average relative Ignorance can be worse than climatology out-of-sample at long lead times (see for example in Figure 11). Figure 12 plots the histogram of multi-model weights. Clearly the variation of the model weights based on a small archive are much larger. Weights of zero are often assigned to model forecasts which contain useful information, for example.

6. Multi-model vs Single Best Model

As noted in Smith et al. (2015)[7], it is sometimes said that a multi-model ensemble forecast is more skillful than any of its constituent single-model ensemble forecasts (see, for example, Palmer et al. (2004); Hagedorn et al. (2005); Bowler et al. (2008); Weigel et al. (2008); Weisheimer et al. (2009); Alessandri et al. (2011)). One common "explanation" (Weigel et al. 2008; Weisheimer et al. 2009; Alessandri et al. 2011) for this is that individual model tends to be overconfident with its forecast and a multi-model forecast reduces such overconfidence, which leads to a more

[7] These first two sentences are taken from Smith et al. (2015).

skillful forecast performance. As shown in Section 6, single model SAP forecast systems are typically between half a bit and two bits less skillful than a LAP system based on the same model. Can a SAP multi-model forecast system regain some of this potential skill? Figure 12 shows that this is unlikely, as the determination of model-weights given SAP varies tremendously relative to their LAP values. Again, it is the performance of the combination of weights that determine the skill of the forecasts, so this variation need not always be deadly.

Figure 13 shows the skill of the multi-model forecast system relative to the forecast system based on the single best model. Both the SAP and the LAP forecast systems show that the multi-model system usually outperforms the single model. Comparing SAP multi-model systems with the single best model SAP system (Figure 13b), the advantage of the multi-model system(s) is stronger when the best model (as well as all the parameters: model weights and dressing and climatology-blended parameters) are ill-identified. Comparing SAP multi-model systems with the single best model LAP system (Figure 13c), however, the advantage of the multi-model system(s) is weaker. Multi-model systems do **not** always outperform the single best model, especially at longer lead times.

At this point, one faces questions of resource distribution. A fair comparison of an N-model forecast system would be against a single model with n-times larger ensemble. (This, of course, ignores the operational fact that it is much more demanding to maintain an ensemble of models than to maintain a large ensemble under one model.) Secondly, note that for each model, κ was a function of lead time. At the cost of making ensemble members non-exchangeable, one could draw ensembles from distinct groups, and weight these members differently for each lead time. Finally, one could develop methods which treat the raw ensemble members from each of the models as non-exchangeable and use a more complex interpretation to form the forecast. While the simple forecast framework of this paper is an ideal place to explore such questions, they lie beyond the scope of this paper. Instead, the extent to which the multi-model forecast system is more misleading than the single model systems concludes the discussion in the next section.

7. Discussion and Conclusions

A significant challenge to the design of seasonal probabilistic forecasting has been discussed and illustrated in a simple system where multiple models can easily be explored in long time limits. The challenge has been addressed within the surrogate modeling paradigm. In the actual system of interest, empirical data is precious: we have very few relevant out-of-sample forecasts, and doubling the current sample size will take decades. For these reasons we consider surrogate systems with sufficient similarity given the questions we wish to ask. We are forced to assume that the results obtained are general enough to make them informative for design in the real-world system; in this particular case we believe that they are: the challenges of interpreting small ensembles in any multi-model context are arguably quite similar. Similarly, the convergence to a clear conclusion in the limit of large ensembles is also arguably quite similar. The details of the rate at which information increases as the ensemble size increases will depend on the details of the dynamics of the system, the quality of the models, and so on. That said, there is sufficient evidence from the study above to show that some current multi-model ensemble studies do not employ initial condition ensembles of sufficient size to achieve robust results.

There is no statistical fix to the challenges of "lucky strikes" when a generally poor model places an ensemble member near an outcome "by chance", and that particular outcome was not well predicted by the other forecast systems. Similarly "hard busts" in a small archive can distort the parameters of the forecast systems, when an outcome occurs relatively far from each ensemble member. In this case, wider kernels and/or heavier weighting on the climatology results. This may be due to structural model failure, or merely to a "rare" event, where rare is related to the ensemble size. Given a sufficiently large ensemble, the forecast system could have assigned an (appropriately low) probability to the observed "bust" event.

In short, the brief duration of the forecast-outcome archive, typically less than 40 years in seasonal forecasting, limits the clarity both with which probability distributions can be derived from individual models and with which model weights can be determined. No clear solution to this challenge has been proposed, and while improvements on current practice can be made, it is not clear that this challenge can be met. Over long periods, like 512 years, the climate may not be well-approximated as stationary. In any event, both observational systems and the models themselves can evolve significantly on much shorter timescales, perhaps beyond recognition.

One avenue open to progress is in determining the relative skill of "the best model" (or a small subset) and the full diversity of models. Following Brocker and Smith (2008) it is argued that a forecast system under the best model with a large ensemble may well outperform the multi-model ensemble forecast system when both systems are given the same computer power. To test this in practice requires access to larger ensembles under the best model. This paper argues future studies, such as ENSEMBLES, could profitably adjust their experimental design to take this into account (see also Machete and Smith (2016)).

A second avenue is to reduce the statistical uncertainty of model fidelity within the available archive. This can be done by running large ensembles (much greater than "9", indeed greater than might be operationally feasible) under each model. This would allow identification of which models have significantly different probability distributions, and the extent to which they are (sometimes) complementary. Tests with large ensembles also reveal the "bad busts" due to small ensemble size to be what they are. It can also suggest that those which remain are indeed due to structural model error.

In closing, it is suggested that perhaps the most promising way forward is to step away from the statistics of the ensembles, and consider the physical realism of the individual trajectories. One can look for shadowing trajectories in each model, and attempt to see what phenomena limit the model's ability to shadow. Identifying these phenomena, and the phenomena that cause them, would allow model improvement independent of the probabilistic skill of ensemble systems. This approach is not new, of course, but the traditional physical approach to model improvement which dates back to Charney. Modern forecasting methods do offer some new tools (Judd et al. 2008), and the focus on probabilistic forecasting is well placed in terms of prediction. The point here is merely that probabilistic forecast skill, while a sharp tool for decision support, may prove a blunt tool for model improvement when the data are precious.

APPENDIX

A1. From Simulation to a Predictive Distribution

This appendix is taken from Smith et al. (2015) Appendix A.

An ensemble of simulations is transformed into a probabilistic distribution function by a combination of kernel dressing and blending with climatology (Brocker and Smith 2008). An N-member ensemble at time t is given as $X_t = [x_t^1, ..., x_t^N]$, where x_t^i is the value of a observable quantity for the i^{th} ensemble member. For simplicity, ensemble members given a model are considered exchangeable. Kernel dressing defines the model-based component of the density as:

$$p(y : X, \sigma) = \frac{1}{N\sigma} \sum_{i}^{N} K\left(\frac{y - (x^i)}{\sigma}\right), \tag{A1}$$

where y is a random variable (the correspondent of the density function p) and K is the kernel, taken here to be

$$K(\zeta) = \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}\zeta^2). \tag{A2}$$

Thus each ensemble member contributes a Gaussian kernel centred at x^i. For a Gaussian kernel, the kernel width σ is simply the standard deviation determined empirically as discussed below.

Even for an ensemble drawn from the the same distribution as the outcome, there remains the chance of $\sim \frac{2}{N}$ that the outcome lies outside the range of the ensemble. Given the nonlinearity of the model, such outcomes can be very far outside the range of the ensemble members. In addition to N being finite, the simulations are **not** drawn from the same distribution as the outcome, as the forecast system is never perfect in practice. To improve the skill of the probabilistic forecasts, the kernel dressed ensemble may be blended with an estimate of the climatological distribution of the system obtained by dressing the historical observations (see Brocker and Smith (2008) for more details, Roulston and Smith (2003) for alternative kernels and Raftery et al. (2005) for a Bayesian approach). The blended forecast distribution is then written as

$$p(\cdot) = \alpha p_m(\cdot) + (1-\alpha) p_c(\cdot), \tag{A3}$$

where p_m is the density function generated by dressing the model ensemble and p_c is the estimate of the climatological density. The blending parameter α determines how much weight is placed on the model. Specifying both values (kernel width σ, and climatology blended parameter α) at each lead time defines the forecast distribution. Both parameters are fitted simultaneously by optimizing the empirical Ignorance score over the training set.

Acknowledgments. This research was supported by the LSE's Grantham Research Institute on Climate Change and the Environment and the ESRC Centre for Climate Change Economics and Policy, funded by the Economic and Social Research Council and Munich Re; it was also supported as part of the EPSRC-funded Blue Green Cities (EP/K013661/1). Additional support for H.D. was also provided by the EPSRC-funded Uncertainty analysis of hierarchical energy systems models: Models versus real energy systems (EP/K03832X/1) and Centre for Energy Systems Integration (EP/P001173/1). L.A.S. gratefully acknowledges the continuing support of Pembroke College, Oxford.

References

Alessandri, A., A. Borrelli, A. Navarra, A. Arribas, P. R. M. Déqué, and A. Weisheimer, 2011: Evaluation of probabilistic quality and value of the ensembles multi-model seasonal forecasts: Comparison with demeter. *Mon. Wea. Rev.*, **2**, 139.

Bernardo, J. M., 1979: Expected information as expected utility. *Ann. Stat.*, **7**, 686–690.

Bougeault, P., and Coauthors, 2010: The thorpex interactive grand global ensemble (tigge). *Bull. Amer. Met. Soc.*, **91**, 1059–1072.

Bowler, N. E., A. Arribas, and K. R. Mylne, 2008: The benefits of multi-analysis and poor-mans ensembles. *Mon. Wea. Rev.*, **136**, 4113–4129.

Brocker, J., and L. Smith, 2006: Scoring probabilistic forecasts: On the importance of being proper. *Wea. Forecasting*, **22**, 382–388.

Brocker, J., and L. Smith, 2008: From ensemble forecasts to predictive distribution functions. *Tellus A*, **60**, 663–678.

Doblas-Reyes, F. J., R. Hagedorn, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting. part ii: Calibration and combination. *Tellus A*, **57**, 234–252.

Doblas-Reyes, F. J., A. Weisheimer, T. N. Palmer, J. M. Murphy, and D. Smith, 2010: Forecast quality assessment of the ensembles seasonal-to-decadal stream 2 hindcasts. *Technical Memorandum (ECMWF)*, 621.

Du, H., and L. Smith, 2017: Multi-model cross-pollination in time. *Physica D: Nonlinear Phenomena*, **353-354**, 31–38.

Glendinning, P., and L. Smith, 2013: Lacunarity and period-doubling. *Dynamical Systems*, **28**, 111–121.

Good, I., 1952: Rational decisions. *J. R. Stat. Soc.*, 107–114.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting. part i: Basic concept. *Tellus A*, **57**, 219–233.

Hewitt, C. D., and D. J. Griggs, 2004: Ensembles-based predictions of climate changes and their impacts. *Eos, Transactions American Geophysical Union*, **85**, 566.

Hide, R., 1958: An experimental study of thermal convection in a rotating fluid. *Phil. Trans. R. Soc. Lond.*, 441–478.

Higgins, S., 2015: Limitations to seasonal weather prediction and crop forecasting due to non-linearity and model inadequacy. Ph.D. thesis, The London School of Economics and Political Science, London, UK.

Hodyss, D., E. Satterfield, J. McLay, T. M. Hamill, and M. Scheuerer, 2016: Inaccuracies with multimodel postprocessing methods involving weighted, regression-corrected forecasts. *Monthly Weather Review*, **144 (4)**, 1649–1668.

Hoeting, J., D. Madigan, A. Raftery, and C. Volinsky, 1999: Bayesian model averaging: a tutorial. *Stat. Sci.*, **14**, 382–417.

Judd, K., C. A. Reynolds, T. E. Rosmond, and L. A. Smith, 2008: The geometry of model error. *J. Atmos. Sci.*, **65**, 1749–1772.

Kirtman, B., and Coauthors, 2014: The north american multimodel ensemble: Phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. *Earth Interactions*, **95 (4)**, 585–601, doi:10.1175/BAMS-D-12-00050.1.

Lorenz, E., 1995: Predictability: a problem partly solved. *Seminar on Predictability, 4-8 September 1995*, ECMWF, Shinfield Park, Reading, ECMWF, Vol. 1, 1-18.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, **20 (2)**, 130–141.

Machete, R., 2007: Modelling a moore-spiegel electronic circuit: the imperfect model scenario. Ph.D. thesis, University of Oxford, UK.

Machete, R., and L. Smith, 2016: Demonstrating the value of larger ensembles in forecasting physical systems. *Tellus A*, **68**, 28 393.

Moran, P., 1950: Some remarks on animal population dynamics. *Biometrics*, **6**, 250–258.

Palmer, T. N., and Coauthors, 2004: Development of a european multi-model ensemble system for seasonal to inter-annual prediction (demeter). *Bull. Amer. Meteorol. Soc.*, **85**, 853–872, doi: http://dx.doi.org/10.1175/BAMS-85-6-853.

Raftery, A., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174.

Rajagopalan, B., U. Lall, and S. E. Zebiak, 2002: Categorical climate forecasts through regularization and optimal combination of multiple gcm ensembles. *Mon. Wea. Rev.*, **130**, 1792–1811.

Read, P. L., 1992: *Rotating Annulus Flows and Baroclinic Waves*, 185–214. Springer Vienna, Vienna.

Ricker, W., 1954: Stock and recruitment. *J. Fisheries Res. Board Can.*, **11**, 559–623.

Roulston, M., and L. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Mon. Wea. Rev.*, **130**, 1653–1660.

Roulston, M., and L. Smith, 2003: Combining dynamical and statistical ensembles. *Tellus A*, **55**, 16–30.

Silverman, B., 1986: *Density Estimation for Statistics and Data Analysis*. 1st ed., Chapman and Hall, London.

Smith, L., 1992: Identification and prediction of low-dimensional dynamics. *Physica D*, **58**, 56–76.

Smith, L., and I. Gilmour, 1998: Enlightenment in shadows. *Nonlinear Dynamics and Stochastic Systems Near the Millennium*, J. B. Kadtke, and A. Bulsara, Eds., Vol. 411, AIP Conference Proceedings, American Institute of Physics, New York, United States, 201–213.

Smith, L. A., H. Du, E. B. Suckling, and F. Niehörster, 2015: Probabilistic skill in ensemble seasonal forecasts. *Quart. J. Roy. Meteor. Soc.*, **141**, 1085–1100.

Sprott, J., 2003: *Chaos and time-series analysis*. Oxford University Press, Oxford.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of cmip5 and the experimental design. *Bull. Amer. Meteor. Soc.*, **93**, 485–498.

Wang, X., and C. Bishop, 2004: Improvement of ensemble reliability with a new dressing kernel. *Quart. J. Roy. Meteor. Soc*, **131**, 965–986.

Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2008: Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quart. J. Roy. Meteor. Soc*, **134**, 241–260.

Weisheimer, A., and Coauthors, 2009: Ensembles: A new multi-model ensemble for seasonal-to-annual predictions–skill and progress beyond demeter in forecasting tropical pacific ssts. *Geophysical Research Letters*, **36 (21)**.

Wilks, D., 2006: Comparison of ensemble-mos methods in the lorenz 96 setting. *Meteorol. Appl.*, **13**, 243–256.

Wilks, D., and T. Hamill, 2007: Comparison of ensemble mos methods using gfs reforecasts. *Mon. Wea. Rev.*, **6**, 2379–2390.

LIST OF FIGURES

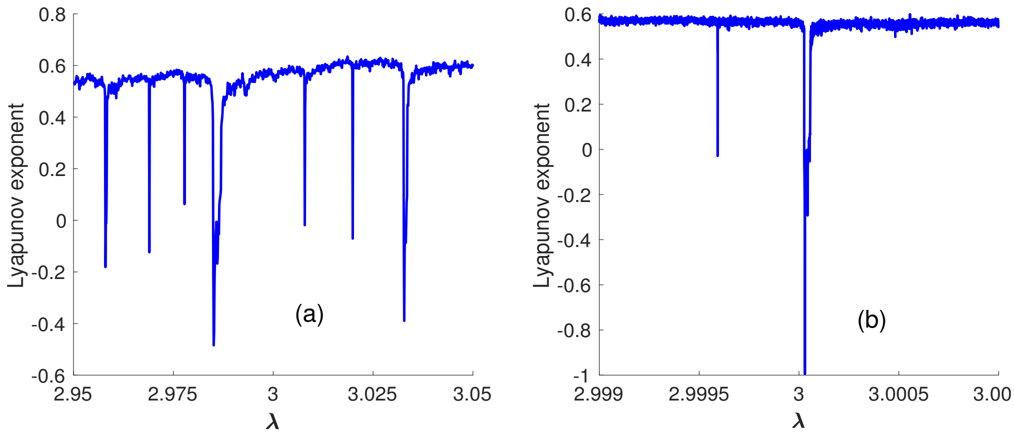
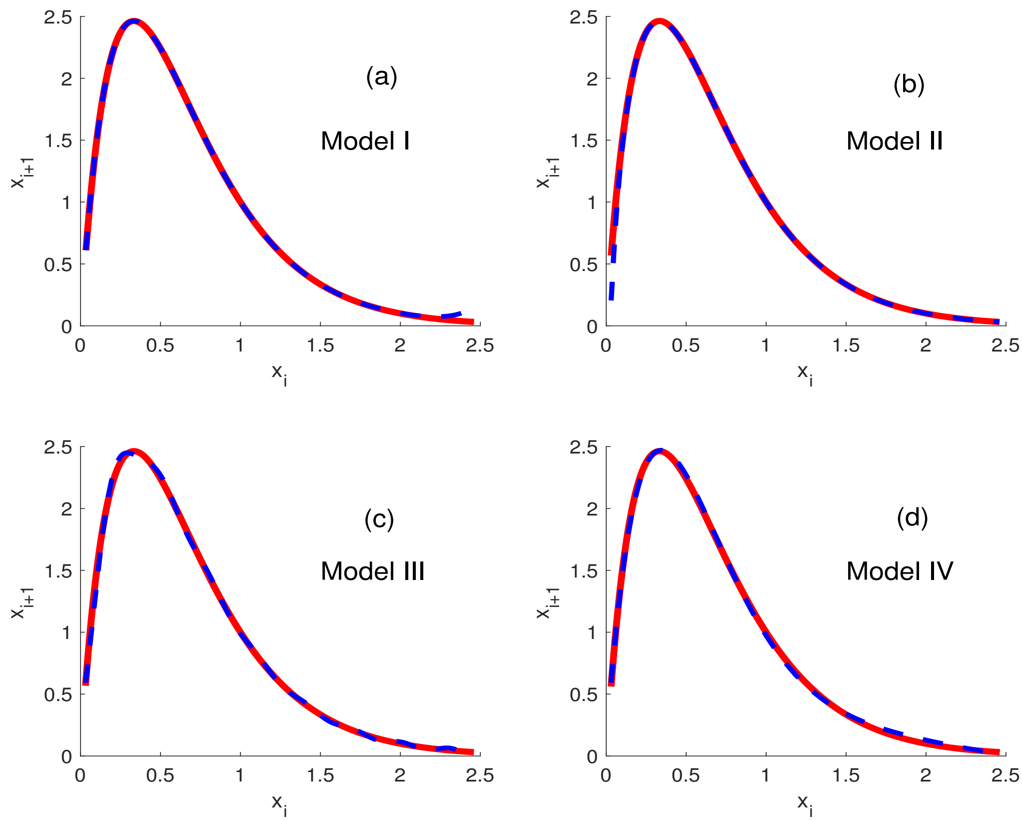
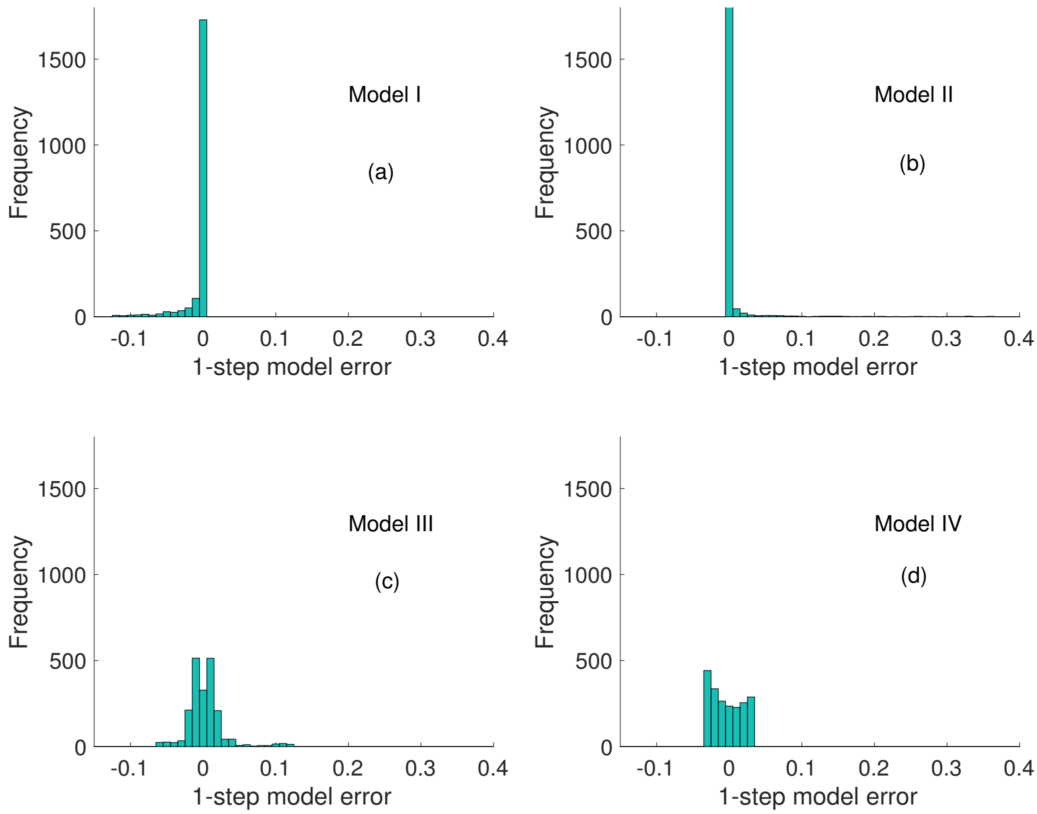


FIG. 1. Estimates of the Global Lyapunov exponent are plotted as a function of λ. a) 4096 values of λ uniformly random sampled between 2.95 and 3.05; b) 4096 values of λ uniformly random sampled between 2.999 and 3.001.



573 FIG. 2. Graphical presentation of the dynamics of four different models, the blue line represents model dynamics as a function of initial conditions and the red line represents the system dynamics.



575 FIG. 3. Histogram of the 1-step model errors, given 2048 different initial conditions with respect to natural
576 measure.

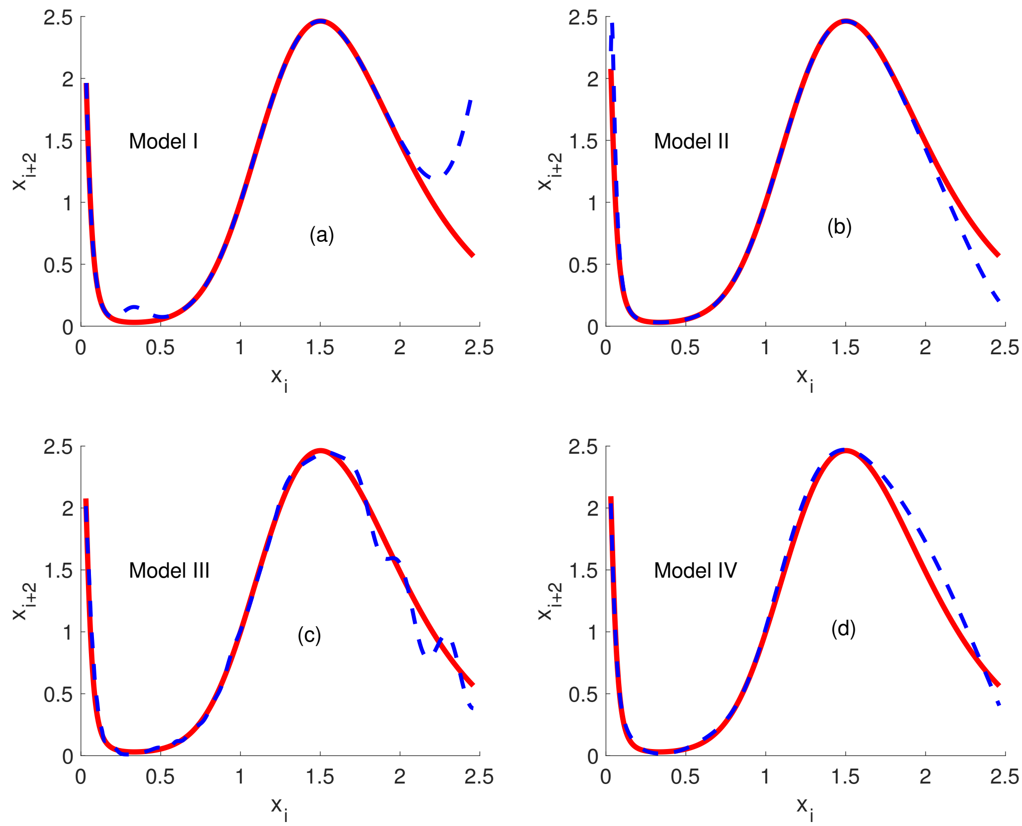


FIG. 4. Graphical presentation of the 2-step evolution of four different models, the blue line represents the 2-step model evolution as a function of initial conditions and the red line represents the 2-step evolution under the system.

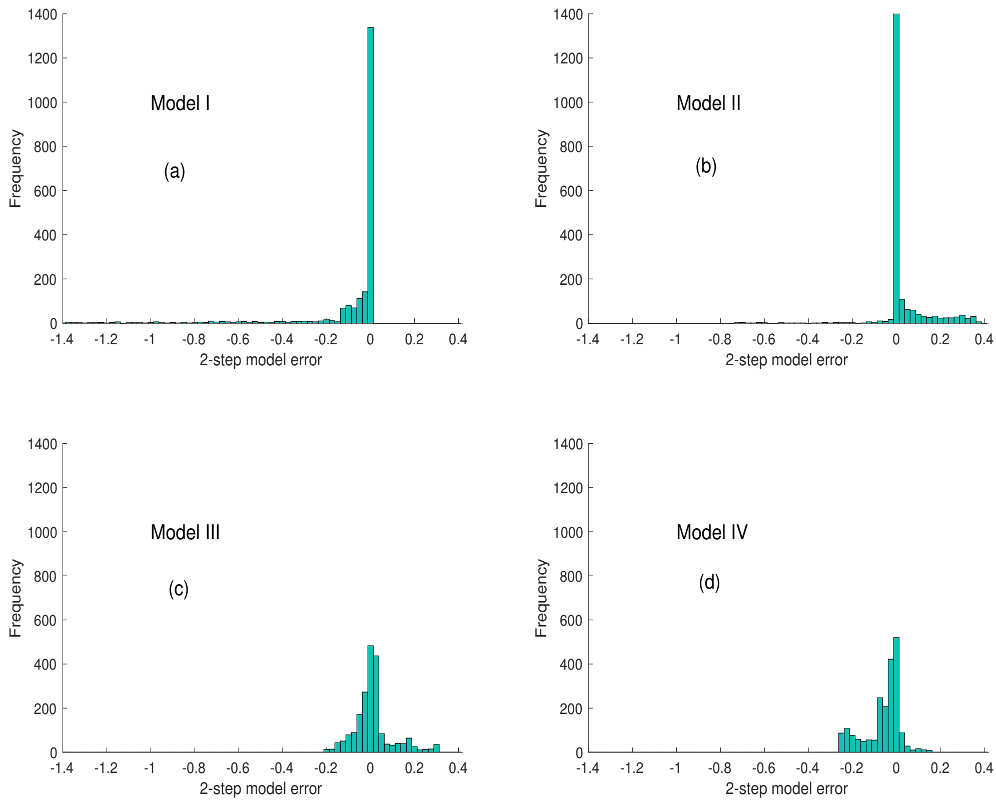
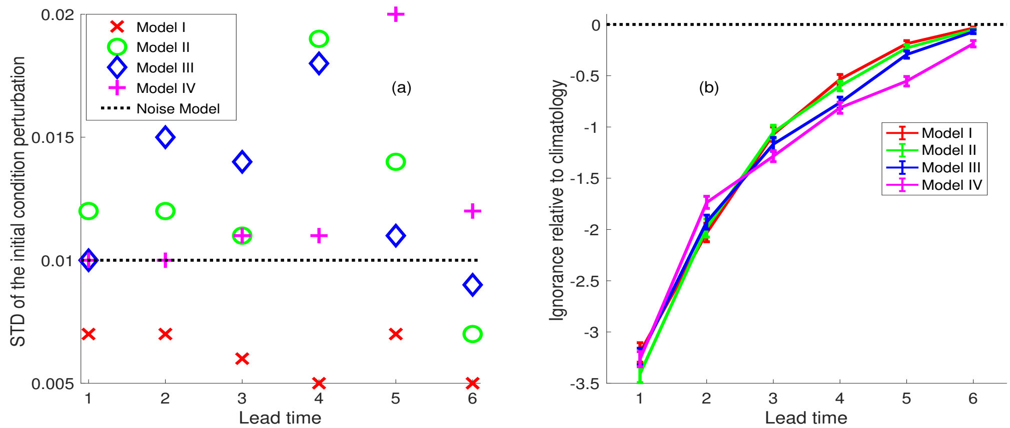


FIG. 5. Histogram of the 2-step model errors, given 2048 different initial conditions with respect to natural measure.



582 FIG. 6. a) The best found perturbation parameter values κ as a function of lead time for each model, the dashed black line reflects the standard deviation of the noise model. b) Ignorance score of each model as a function of lead time, the dashed black line reflects skill of climatology which defines zero skill.

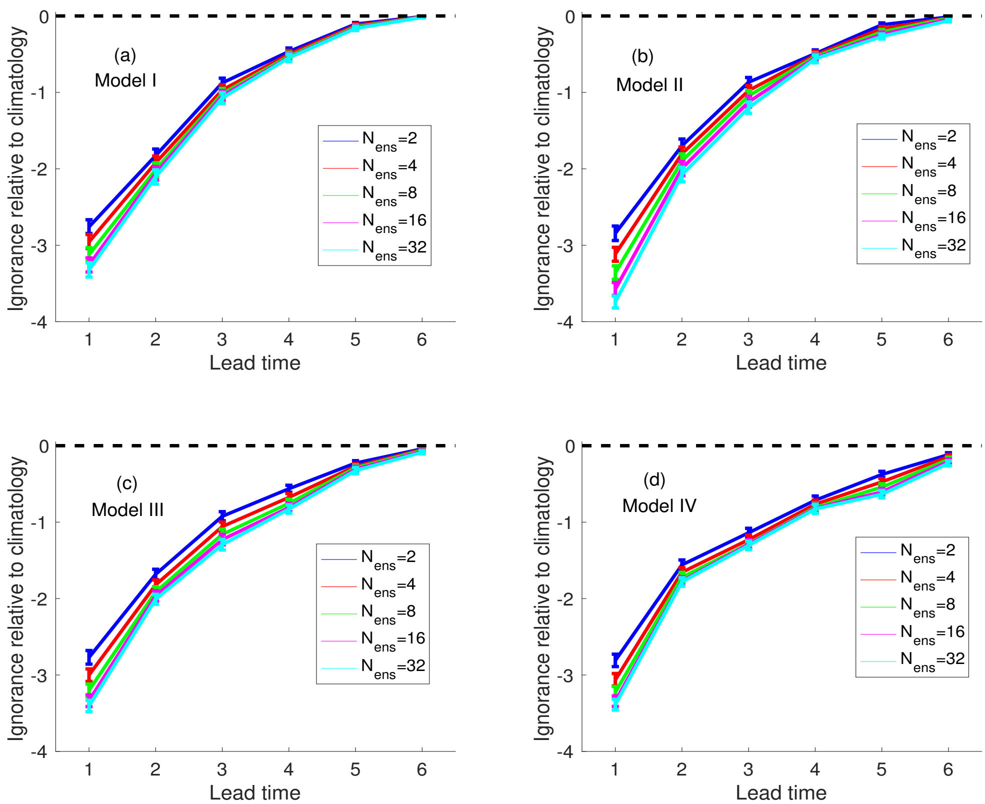


FIG. 7. The Ignorance score varies as the ensemble size increases for each model.

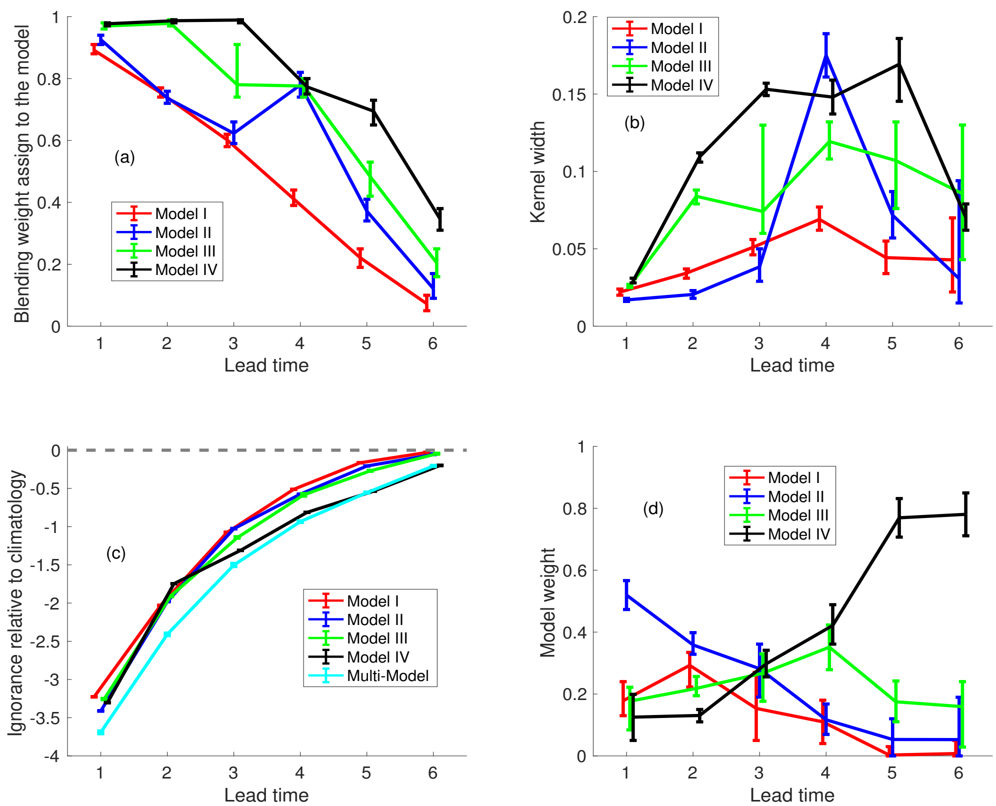


FIG. 8. Forecast Ignorance, climatology-blend weight assigned to the model, kernel width and weights assigned to each individual model are plotted as a function of lead time.

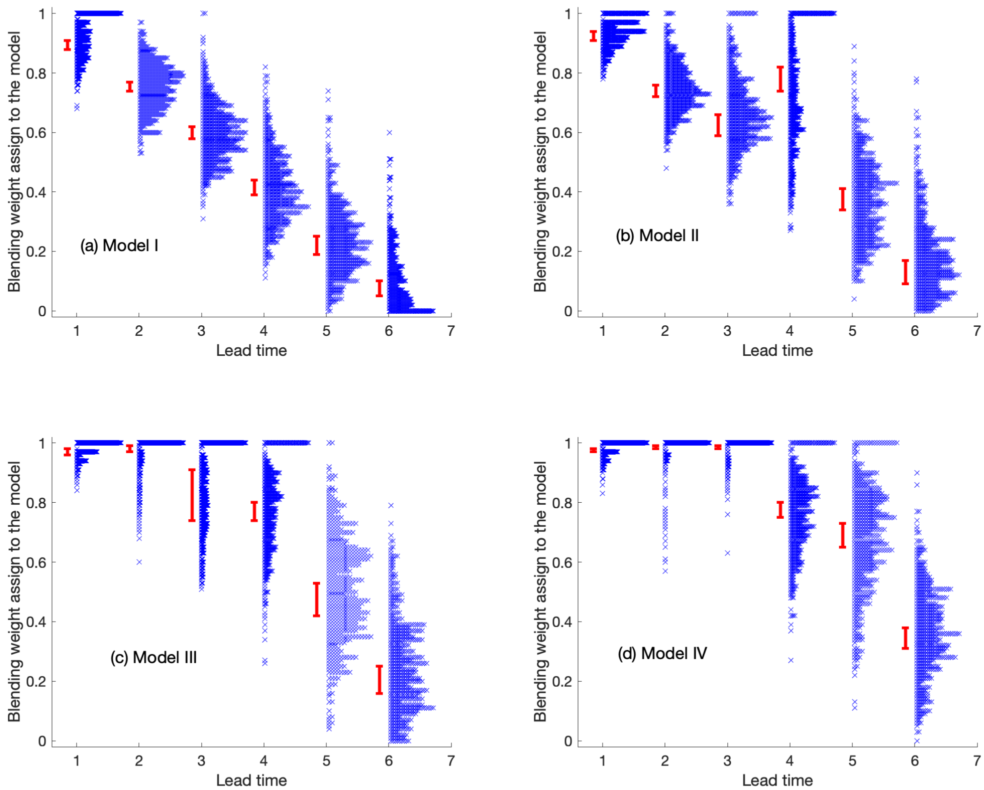


FIG. 9. Climatology-blend weights assigned to each model. The red bars are the 95^{th} percentile range of the fitted weights based on 512 forecast-outcome archives. Each contains 2048 forecast-outcome pairs. The blue crosses represent the histogram of the fitted weights based on 512 forecast-outcome archives. Each of these contains only 40 forecast-outcome pairs.

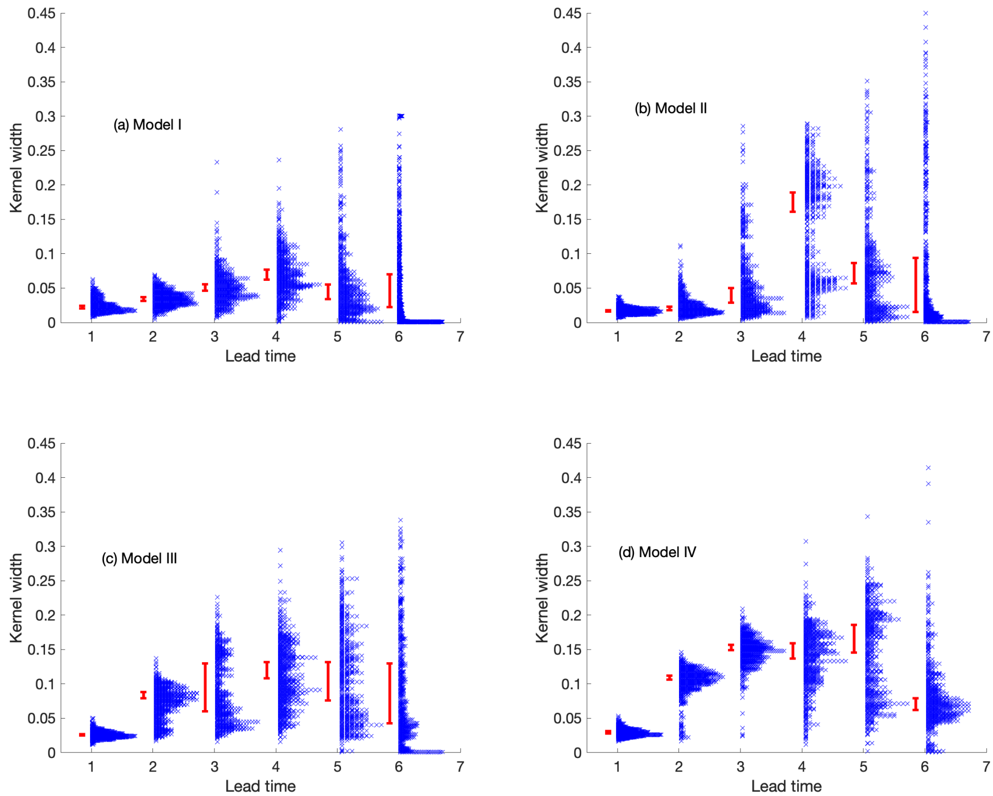


FIG. 10. Kernel width of each model's forecasts. The red bars are the 95^{th} percentile range of the fitted kernel width based on 512 forecast-outcome archives, each contains 2048 forecast-outcome pairs. The blue crosses represent the histogram of the fitted kernel width based on 512 forecast-outcome archives, each contains only 40 forecast-outcome pairs.

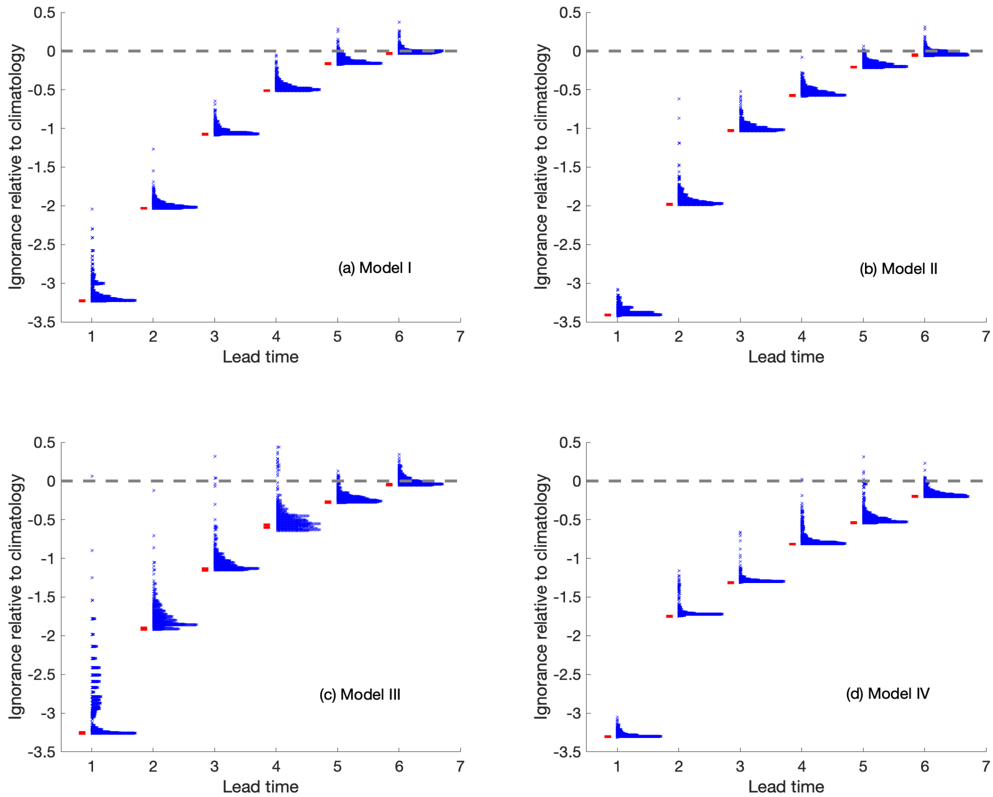


FIG. 11. Ignorance score of each model's forecasts. The red bars are the 95^{th} percentile range of Ignorance score calculated based on a testing set containing 2048 forecast-outcome pairs, using the climatology-blend weights and kernel widths fitted based on 512 forecast-outcome archives, each contains 2048 forecast-outcome pairs. The blue crosses represent the histogram of Ignorance score calculated based on the same testing set but using the climatology-blend weights and kernel widths based on 512 forecast-outcome archives, each contains only 40 forecast-outcome pairs.

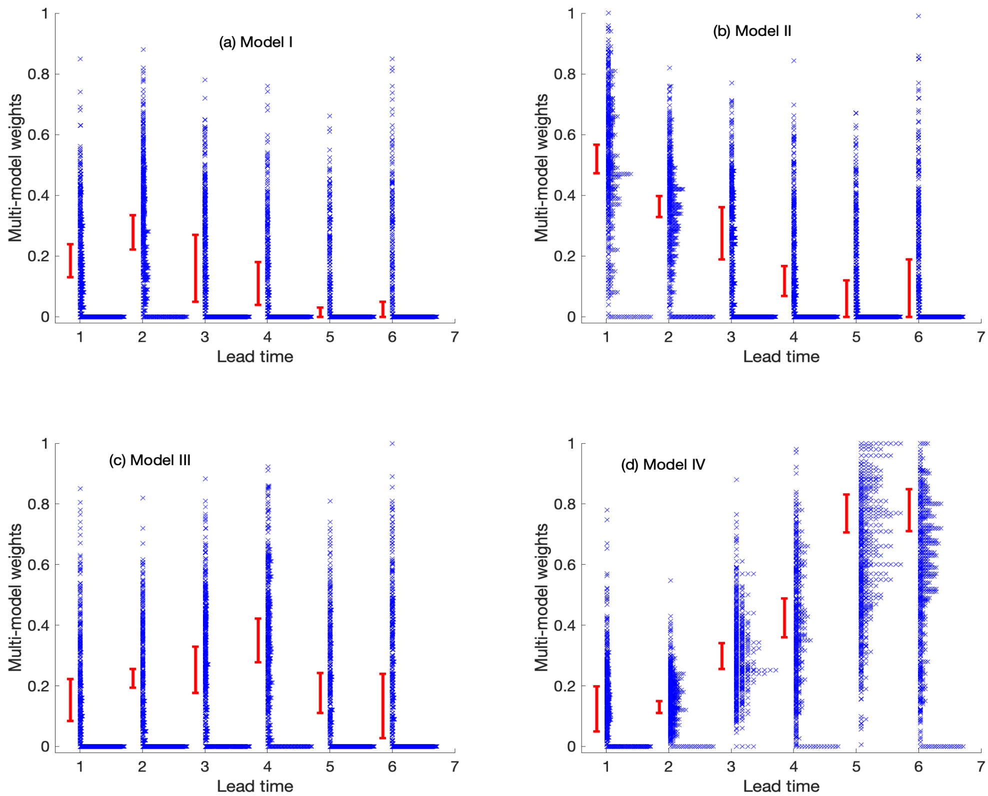
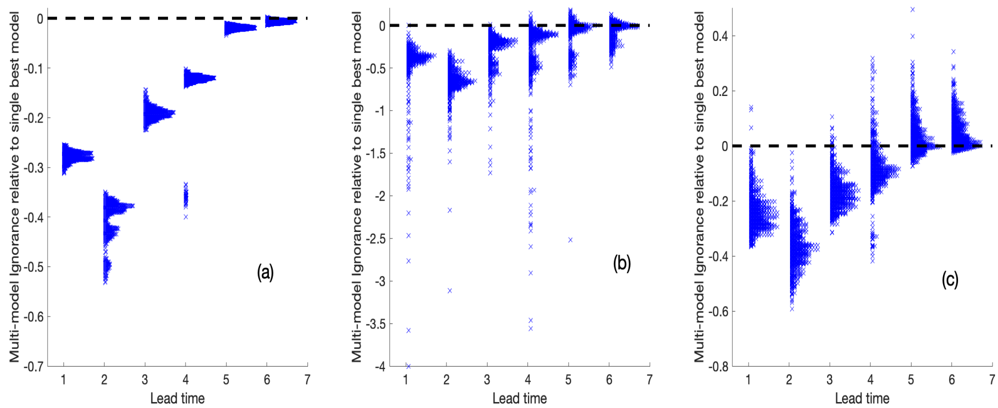


FIG. 12. Multi-model weights for each set of model forecasts. The red bars are the 95^{th} percentile range of model weights calculated based on a testing set containing 2048 forecast-outcome pairs, using the climatology-blend weights and kernel widths fitted based on 512 forecast-outcome archives, each contains 2048 forecast-outcome pairs. The blue crosses represent the histogram of model weights calculated based on the same testing set but using the climatology-blend weights and kernel widths based on 512 forecast-outcome archives, each contains only 40 forecast-outcome pairs.



607 FIG. 13. Ignorance of multi-model ensemble relative to the single best model. The blue crosses represent the histogram of the Ignorance of the multi-model ensemble relative to the single best model (black dashed line). (a) Model weights and dressing and climatology-blend parameters are fitted based on 512 large archives, each contains 2048 forecast-outcome pairs. (b) Model weights and dressing and climatology-blend parameters are fitted based on 512 small archives, each contains 40 forecast-outcome pairs. (c) The Ignorance of the multi-model ensemble is calculated using model weights and dressing and climatology-blend parameter which are fitted based on 512 small archives, while the Ignorance of the single best model is calculated based on 512 large archives.