Alexander S. Jarman and Leonard A. Smith

Quantifying the predictability of a predictand: demonstrating the diverse roles of serial dependence in the estimation of forecast skill

Article (Accepted version) (Refereed)

Original citation:
Jarman, Alexander and Smith, Leonard A. (2018) *Quantifying the predictability of a predictand: demonstrating the diverse roles of serial dependence in the estimation of forecast skill.* Quarterly Journal of the Royal Meteorological Society. ISSN 0035-9009

DOI: 10.1002/qj.3384

Quantifying the Predictability of a Predictand: Demonstrating the Diverse Roles of Serial Dependence in the Estimation of Forecast Skill

Alexander S. Jarman[a*] and Leonard A. Smith[ab]

[a] *Centre for the Analysis of Time Series, London School of Economics, London WC2A 2AE, UK*

[b] *Pembroke College, OX1 1DW, Oxford, UK*

[*] Correspondence to: Centre for the Analysis of Time Series, London School of Economics, London WC2A 2AE, UK. E-mail: a.s.jarman@lse.ac.uk

Predictability varies. In geophysical systems, and related mathematical dynamical systems, variations are often expressed as serial dependence in the skill with which the system is, or can be, predicted. It is well known, of course, that estimation is more complicated in cases where the time series sample in-hand does not reflect an independent from the target population; failure to account for this results in erroneous estimates both of the skill of the forecast system and of the statistical uncertainty in the estimated skill. This effect need not be indicated in the time series of the predictand; specifically: it is proven by example that linear correlation in the predictand is neither necessary nor sufficient to identify misestimation. Wilks [Quarterly Journal of the Royal Meteorological Society 136, 2109 (2010)] has shown that temporal correlations in forecast skill give rise to biased estimates of skill of a forecast system, and made progress on accounting for this effect in probability-of-precipitation forecasts. Related effects are explored in probability density forecasts of a continuous target in three different dynamical systems (demonstrating that linear correlation in the predictand is neither necessary nor sufficient), and a simple procedure is presented as a straightforward, good practice test for the effect when estimating the skill of forecast system.

Key Words: probabilistic forecasting, forecast skill, serial correlation

Received . . .

1. Introduction

The standard procedure for demonstrating the skill of a forecast system is to evaluate an out of sample sequence of forecasts with target data, and determine whether the skill of the forecast system is greater than that attributable to chance. Establishing statistical confidence in forecast skill, however, is complicated where sequential target data are not independent. In this case, the variances of sampling distributions of the corresponding skill estimates are altered relative to those computed with data drawn at random, resulting in inaccurate statistical inferences of skill.

The effects of serial dependence on the sampling properties of scoring rules have important implications for proving forecast skill, requiring sample size corrections to obtain reliable skill estimates, and the inter-comparison of forecast systems. In the case of real-time forecasting applications, larger samples of forecast evaluations, and hence longer durations of time, are required to establish statistically significant skill where confidence in skill has initially been overestimated. Wilks (2010) demonstrates the effects of serial dependence on estimates of the Brier score (BS) and Brier skill score (BSS) in a binary predictand scenario, and how statistical inference yields overconfident estimates of forecast skill (i.e. a higher probability of type I errors). To compensate for this effect on skill estimation, *effective sample size* (ESS) (Thiébaux and Zwiers, 1984) corrections are proposed to achieve more accurate estimation of skill (i.e. that made with serially independent data). Methods for accounting for the effect of serial dependence on forecast evaluation have also been considered in other studies (see, for example, Hamill, 1999; Ferro, 2007; Pinson et al., 2010, and references therein).

The investigation into the effects of serial dependence on forecast skill estimation is applied to probabilistic forecasts of continuous predictands. Accordingly, forecast evaluation is performed with the information-theoretical logarithmic score referred to as *ignorance* (Good, 1952; Roulston and Smith, 2002). Moreover, ignorance belongs to the class of *proper* scoring rules (Bröcker and Smith, 2007; Gneiting and Raftery, 2007), ensuring that, in the long run, no forecast system is expected to obtain a score superior to that of the probability(s) that generated the outcome. Wilks (2010) considered the estimates of skill for binary probability-of-precipitation forecast systems, contrasting skill estimated from independent events with the estimates of

Table 1. Example systems used for demonstrating the presence or absence of linear correlation (LC) in scores

	No LC in State	**LC in State**
No LC in Score	IID Gaussian process	AR(1) (Section 3.2)
LC in Score	Logistic map (Section 3.3)	· Linear-calibration/ beta-refinement model (Wilks, 2010) · Lorenz63 (Section 3.1)

the same forecast system from a time series of forecasts with serial dependence; the empirical sampling distributions differ from those expected in the case of **no** serial dependence. The current paper considers continuous target variables, and employs a resampling technique (Efron, 1981; Wilks, 2011) to emulate the conditions for serial independence. The inflationary effect on the variances of the score sampling distributions and estimation of forecast skill is illustrated and discussed. Serial dependence in target time series need **not** imply linear temporal correlation in the scores, however, and in such cases, estimates of forecast skill may not be biased. In fact, four distinct cases of (non-)effects of serial dependence on estimation of forecast skill are possible. Firstly, where linear correlation in target data is present in the scores; secondly, where linear correlation in target data is not present in the scores; thirdly, where nonlinear correlation is exhibited in target data, resulting in linear correlation present in the scores; and fourthly, where there is no serial dependence present in either the target data or scores[1]. It is argued here that those evaluating forecasts should aim to distinguish between these different cases of effects and non-effects so as to operate within a robust statistical framework, and therefore maximise the benefit of predictive information.

Section 2 explains the implications of serial dependence for the sampling properties of scoring rules, and derives an analytical expression for the variance of the sampling distribution (hereafter sampling variance) of the ignorance score (IGN) under conditions of serial independence, equivalent to that derived by Bradley et al. (2008) for the Brier score, and used by Wilks (2010) to formulate expressions for ESS corrections.

The first three of the above four distinct cases of effects/non-effects of serial dependence on the estimation of forecast skill are illustrated in Section 3 in the context of three different dynamical systems: the three-dimensional Lorenz63 flow, a first-order autoregressive process, and the logistic map. The fourth case of serial dependence being present in neither the target data nor the corresponding scores is demonstrable with, for example, an IID Gaussian process, and is not discussed further in this paper. Table 1 lists the four systems as examples of whether linear correlation, which is either present or absent in target data, is also present or absent in the corresponding time series of a given scoring rule.

Finally, a procedure for making effective sample size (ESS) corrections by comparing empirical estimates of the sampling variance of the score under serial dependence and under serial independence (using resampling) is proposed in Section 4. This approximate method contrasts with the approach of Wilks (2010) which requires an empirical fitting of the ratios of simulated to analytically derived sampling variances as functions of the system-model parameters. ESS corrections aim to provide an estimate of the minimum duration of observations required to determine the skill of a forecast system with the desired accuracy. Very small sample sizes, of course, always present a challenge to resampling methods which will become apparent in practice. As with any extrapolation into the future, large unforeseen alterations of the behaviour of the system can yield unforeseen changes in forecast skill (if, say, a large interstellar object impacts the Earth).

[1] While "serial dependence" is the general term used here; the cases above are distinguished with the terms "linear correlation" and "nonlinear correlation" .

2. Effects of serial dependence on sampling distributions of scoring rules

The effect of serial dependence on the sampling distributions of a statistic is commonly encountered in the statistical analysis of geophysical variables, and has been examined in depth in the literature (Leith, 1973; Jones, 1975; Albers, 1978; Trenberth, 1984; Thiébaux and Zwiers, 1984). Consider a random variable which has a population distribution with mean μ and standard deviation σ. An intuitive result of the Central Limit Theorem is that the finite-time average of a sample of N independent and identically distributed (IID) target data of the random variable is a normal random variable with mean μ and standard error $\sigma/\sqrt{N}$. Since geophysical phenomena are typically *red* processes, samples of data may be collected at time intervals which are too short for the assumption of independence to hold (Leith, 1973). The sampling variance of a finite-time average computed from serially dependent geophysical data need not scale as $1/\sqrt{N}$ (as is the case for independent data i.e. a *white-noise* process). As sample size increases, the rate of convergence of the sample averages on the true mean μ can be significantly slower (or faster) than those which are IID; referred to as inflation (or deflation) of the sampling variance by Wilks (2011). Consequently, the duration of time required to obtain realistic estimates of μ is increased (or decreased). Without accounting for the effect of serial dependence, textbook statistical inferences of an underlying statistic which are based upon the assumption of independence will yield biased results (Wilks, 2011). Indeed, linear correlation need not be detectable in observations of nonlinear systems with the autocorrelation function (Fraser and Swinney, 1986).

Figure 1 illustrates how the sampling variance of a scoring rule is inflated under serial dependence by providing a comparison of a time series of serially dependent scores and a series consisting of a random resampling of the time series. The latter series represents - by construction - an estimate of the *natural measure* of the system (Ott, 2002), and the sequential scores contained within are serially independent. The unbiased estimator for the sampling variance of 8 subsample estimates (sample size of $N = 16$) of ignorance ($\widehat{IGN_e}$) computed from the serially dependent scores ($s^2_{\widehat{IGN_e}} = 0.15$) is larger than the serially independent scores ($s^2_{\widehat{IGN^*_e}} = 0.05$). The degree of variance inflation is reduced with increase in sample size as $s^2_{\widehat{IGN_e}}$ and $s^2_{\widehat{IGN^*_e}}$ decrease and converge; for

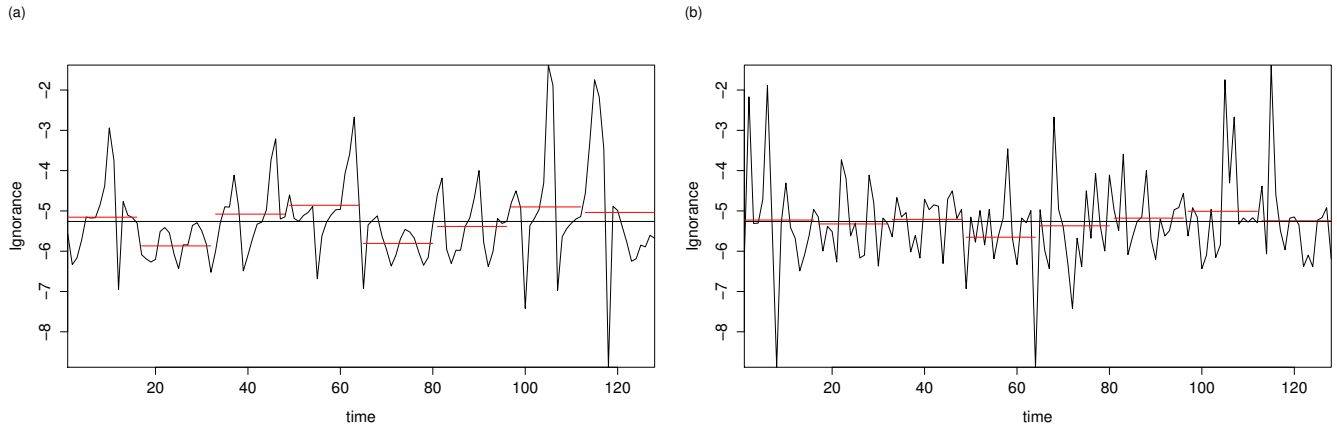

Figure 1. Time series of 2^7 forecast skill evaluations computed from (a) a time series of forecasts of observed states of the Lorenz63 system and (b) a random resample from that time series. The time series scores are serially dependent ($r_1(\widehat{IGN_e}) = 0.45$) while the randomly resampled scores are serially independent by construction ($r_1(\widehat{IGN_e^*}) \approx 0$). Averages over sequential samples of size $N = 16$ (red lines) tend to vary more about the IGN estimate over the entire time series ($\widehat{IGN_e} = -5.26$) in (a) compared to (b), resulting in a sampling distribution of the averages which is larger in variance. The sampling variances of the 8 subsamples are $s^2_{\widehat{IGN_e}} = 0.15$ and $s^2_{\widehat{IGN_e^*}} = 0.05$.

example, the sampling variances for 4 subsample estimates ($N = 32$) are $s^2_{\widehat{IGN_e}} = 0.12$ and $s^2_{\widehat{IGN_e^*}} = 0.01$. Serial dependence is measured here (and throughout the rest of the paper) using the lag 1 sample autocorrelation function (ACF) r_1.

Forecast evaluation is routinely carried out to monitor and improve the quality of forecast systems, yet its usefulness is limited by the frequent omission of sampling uncertainty inherent in the estimation of forecast skill (Joliffe, 2007). The sampling uncertainty of a scoring rule is dependent on both sample size and the statistical characteristics of the forecasts and observations (Bradley et al., 2008). In that sense, a scoring rule can be considered in the same way as standard statistical inference, where some underlying parameter or value θ is estimated, for example, by constructing a confidence interval for an empirical estimate $\hat{\theta}$ using a resampling method (Joliffe, 2007). This is a simple approach for determining sampling uncertainty, although it is also computationally inefficient.

An alternative approach requiring minimal computational effort would be to find analytical solutions for the sampling uncertainty of a scoring rule using sampling theory. Bradley et al. (2008) derive analytical expressions for the sampling variances of the Brier score and Brier skill score with respect to binary forecasts; the former being an exact solution due to the unbiasedness of the Brier score. The equivalent derivation of the ignorance score is presented below, and could, in theory, be used, along with empirical sampling variances under serial dependence, to measure the degree of inflation, and to determine expressions for ESS corrections (see Section 4).

2.1. *Ignorance*

The sample estimator of the ignorance score is expressed as

$$\widehat{IGN}(p(x), X) = -\frac{1}{N}\sum_{i=1}^{N} log_2(p(X_i)), \tag{1}$$

where $p(X_i)$ is the forecast probability that observation X_i will occur. The sampling variance of the ignorance score is then given by[2]

$$\begin{aligned} Var[\widehat{IGN}(p(x), X)] &= Var\left[-\frac{1}{N}\sum_{i=1}^{N} log_2 p(X_i)\right] \\ &= \frac{1}{N^2}\sum_{i=1}^{N} Var[log_2 p(X_i)] \\ &= \frac{1}{N} Var[log_2 p(x)]. \end{aligned} \tag{2}$$

The variance term on the RHS can be expanded as follows

$$\begin{aligned} Var[log_2 p(x)] &= E[log_2^2 p(x)] - E[log_2 p(x)]^2 \\ &= E[log_2^2 p(x)] - IGN^2. \end{aligned} \tag{3}$$

Therefore,

$$Var[\widehat{IGN}(p(x), X)] = \frac{1}{N}[E[log_2^2 p(x)] - IGN^2], \tag{4}$$

where

$$E[log_2^2 p(x)] = \frac{1}{N}\sum_{i=1}^{N} log_2^2 p(X_i). \tag{5}$$

The derivation above is based on the assumption that the forecast-observation pairs (p_i, X_i) are independent random samples from their joint distribution (Murphy and Winkler, 1987). This assumption is commonly (and erroneously) made in real world geophysical forecasting (Seaman, 1992; Seaman, Mason, and Woodcock, 1996; Wilks, 2010), and can lead to biased estimates of forecast skill.

In general, derivation of an exact solution of the sampling variance (as in Eqn. (4)) is not straightforward for scoring rules; estimates need to be evaluated with sufficient sample sizes to produce stable results indicating they are usefully accurate (Bradley et al., 2008; Wilks, 2010). Bradley et al. (2008) observe that, because of the inclusion of the higher moments of the joint distribution of the forecasts and observations in the Brier score, relatively large sample sizes are required; at least on the order of $\sim 10^2$. The required size increases for lower climatological event probability, and with a higher degree of forecast skill. Wilks

[2] $Var[\cdot]$ and $E[\cdot]$ denote the sample variance and the sample mean here. The sample variance is distinct from the sampling variance in that the sample variance is the variance of a given sample while the sampling variance is the variance of a sample estimator, given a series of samples.

(2010) concludes that a sample size of $N = 3000$ is sufficient when forecast skill is estimated with the Brier score. In light of the challenge for deriving analytical score sampling variances, and because illustration of the effects of serial dependence on forecast evaluation does not require inordinate sample sizes, the empirical approach for estimation of IGN and its sampling variance is used in this study.

Wilks (2010) exploits the fact that the analytical sampling variance of the Brier score is based on the serial independence of the forecasts p and target data x, and that it depends only on the moments of their joint distribution, to derive ESS corrections in terms of the parameters of the model. Derivation of such ESS corrections is more difficult for IGN because Eqn. (4) depends on $E[log_2^2 p(x)]$ rather than the moments of the joint distribution; that derivation is beyond the scope of this paper. An alternative method for ESS corrections is proposed here. This *approximate* method consists of finding the difference between sample sizes corresponding to a given empirical sampling variance computed respectively from serially dependent synthetic time series and serially independent random resamples; it is defined in Section 4. There exist other resampling methods for this purpose, but the method employed here is sufficient to simulate serial independence.

3. Estimation of forecast skill under serial dependence

The results of forecast skill estimation under serial dependence are presented and critically examined for three different dynamical systems. A Monte-Carlo approach to generate synthetic datasets of sequential forecast-observation pairs is employed. The statistics of the resulting time series of scores are then compared with those of an IID series (constructed by sampling with respect to the natural measure of the system i.e. random resampling) to detect and assess the magnitude of the effects of serial dependence on forecast skill estimation. The key idea here is to simply compare the sampling variances of the score estimates, and their rates of convergence, over increasing sample sizes. In each case, 2^8 simulations have been performed for each set of parameters to yield convergent score sampling distributions.

3.1. Case study 1: Lorenz63

The 3-dimensional Lorenz63 flow (Lorenz, 1963) is a suitable dynamical system to illustrate the case where serial dependence is present in both the forecast target and the corresponding forecast skill scores (see Table 1; bottom right). The evolution of the system state is governed by a discrete time, deterministic nonlinear dynamical system, defined by the following set of three ordinary differential equations (with respect to time):

$$\begin{aligned} \dot{x} &= -\sigma x + \sigma y \\ \dot{y} &= -xz + rx - y \\ \dot{z} &= xy - bz, \end{aligned} \tag{6}$$

where $x, y, z \in \mathbb{R}$ (see Appendix A.1 for full details of the system configuration used in this study). The trajectory of the Lorenz63 system is recognisable by its "butterfly wings" attractor which occupy two distinct regions of state space. Consequently, the x state variable exhibits bimodal behaviour (see Fig. 2) which can result in highly correlated forecast target data for sufficiently short time steps. Hence, for the purposes of this study, it is convenient to evaluate the forecasts solely on a scalar observation of x.

Consider the forecasts to be constructed from a perfect model[3] so that the system state and model state share the same state space,

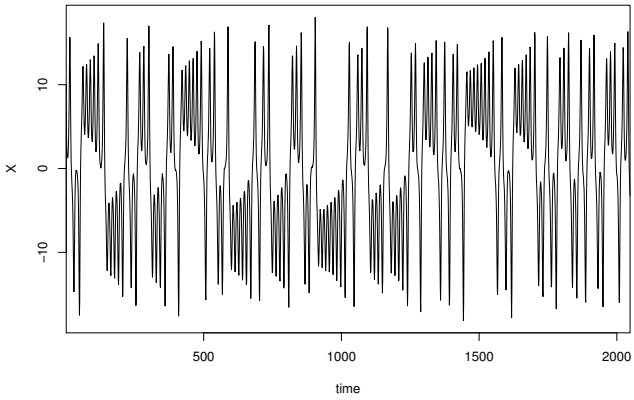

Figure 2. Time series of target variable x illustrating the bimodal behaviour of the Lorenz63 attractor. The target data have a strong degree of linear correlation ($r_1(x) \approx 0.96$ for a sample size $N = 2^{11}$ timesteps).

and state estimation is only subject to observational uncertainty (Smith, 2001, 2006). Given a time series of observations of x, denoted $s_1, \ldots, s_t, \ldots, s_N$, determined by the true value $\tilde{x}$ plus some additive observational noise ϵ, an observation s_t at time t is defined as

$$s_t = \tilde{x}_t + \epsilon_t, \tag{7}$$

where $\epsilon_t \overset{iid}{\sim} F(\cdot)$ reflects an observational noise term where a stochastic model $F(\cdot)$ is used to simulate observational noise in the target data, taken here to be $\mathcal{N}(0, \sigma^2)$. An ensemble forecast approach has been adopted here to sample the initial conditions $\mathbf{s}_1 = s_{1,1}, \ldots, s_{1,M}$ for an M member ensemble at each forecast initialisation at time $t = 1$ using two different data assimilation (DA) schemes. The more simple of the two DA schemes, referred to as *inverse noise* (IN), is based on adding draws from the inverse of the stochastic observational noise model (in this case, again a Gaussian distribution i.e. $\epsilon \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$) to the observation s to assign values to the ensemble members. The other scheme, called *pseudo-orbit data assimilation* (PDA) (Judd and Smith, 2004; Du and Smith, 2012), provides maximum likelihood estimates of the true system state using a large window of observations. The PDA scheme provides a better approximation of the initial conditions which are more consistent with the long term model dynamics (Du, 2009), and hence often produces more skilful forecasts than the IN scheme. Following the process of state estimation using these two DA schemes, probabilistic forecasts are constructed by kernel dressing and blending (Bröcker and Smith, 2008) the initial conditions (see also Appendix A.1).

Figure 3 shows the sampling variances of IGN estimates as a function of sample size for the two different forecast systems with forecast lead time $\iota = 0.1LTU$[4], illustrating the effect of serial dependence on the skill statistics. The PDA forecast system produces more skilful forecasts ($\widehat{IGN_e} = -5.34$) than the IN forecast system ($\widehat{IGN_e} = -3.57$) averaged over the entire time series ($N = 2^{14}$). The forecasts have been evaluated against a reference forecast constructed from the unconditional climatological distribution, a measure referred to as empirical ignorance, which is defined as (see Du and Smith, 2012)

$$\widehat{IGN_e}(p(x), X) = -\frac{1}{N}\sum_{i=1}^{N} log_2\left[\frac{p(X_i)}{p_{clim}(X_i)}\right]. \tag{8}$$

[3] In mathematical systems, this implies that one has access to the True probability distribution which determines that outcome. It is sometimes useful to speak of the True distribution even if it may not exist (Good, 1983).

[4] One Lorenz Time Unit (LTU) is analogous to 1.2 hours of standard time, and corresponds to 100 integration time steps - similar in approach to Stephenson (2004). See Appendix A.1.

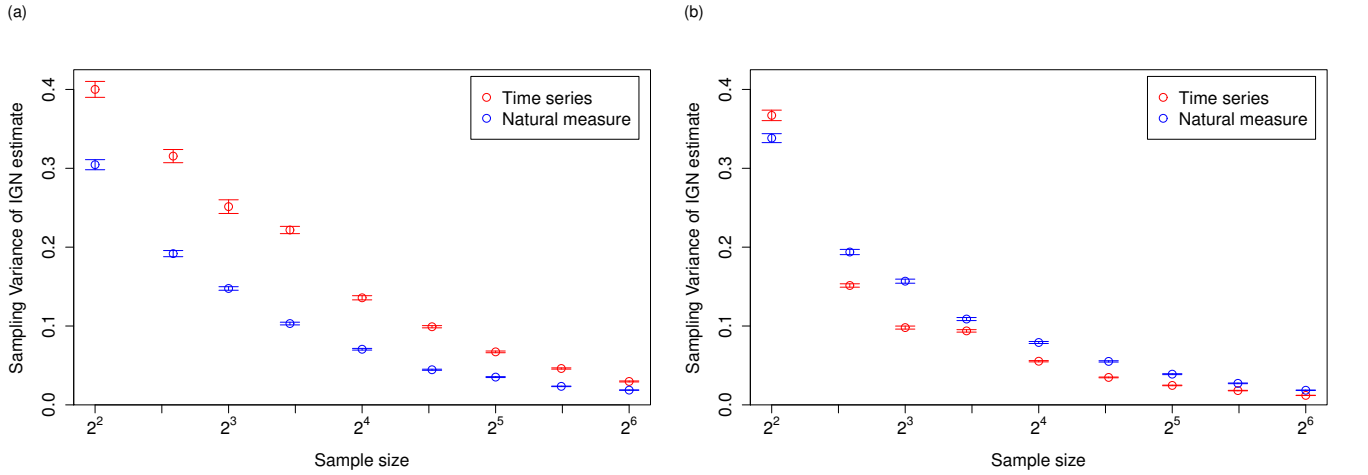

Figure 3. Results for Lorenz63: sampling variances of 2^8 IGN estimates computed from forecasts with lead time $\iota = 0.1LTU$ constructed from the (a) PDA and (b) IN models of a serially dependent time series of Lorenz63 target data ($r_1(s) \approx 0.94$; red circles) and an IID randomly resampled series of scores (equivalent to sampling the data with respect to the natural measure of the system; blue circles), both with $5\% - 95\%$ bootstrap uncertainty intervals. There is a clear inflation of the sampling variances up to at least sample size $N = 2^6$ for the PDA forecasts showing that linear correlation is exhibited by both the forecast target and corresponding ignorance scores ($r_1(IGN_e) \approx 0.5$ for the PDA forecasts, and $r_1(IGN_e) \approx 0.31$ for the IN forecasts). There is a slight deflation of the IN forecast skill sampling variances overall, reflecting their poorer skill, and that the effect of serial correlation on forecast skill estimation is minimal. The wider uncertainty intervals for the serially dependent series reflect the poorer estimates of skill, and hence, greater variability of the sampling variances of those estimates.

A negative score indicates that the forecast system has superior skill to the climatological forecast. A 3-D scatter plot of points on the Lorenz attractor coloured by the IGN score of the associated forecast (not shown, see Jarman, 2014) reveals large scale regions of similar predictability reminiscent of similar plots of the doubling time of infinitesimal uncertainties (Smith, 1999).

Linear correlation is present in both the target time series ($r_1(s) \approx 0.94$) and corresponding ignorance scores ($r_1(IGN_e) \approx 0.5$ for the PDA forecasts, and $r_1(IGN_e) \approx 0.31$ for the IN forecasts)[5]. There is a resulting inflation of the variance $s^2_{\widehat{IGN_e}}$ of the score sampling distribution with respect to that of the natural measure $s^2_{\widehat{IGN^*_e}}$ for the PDA forecasts, but a slight deflation for the IN forecasts. This contrasting outcome occurs because the time series of superior PDA forecasts is more likely to exhibit the underlying correlation structure, and hence, greater variability in skill estimates, than the poorer and more noisy (i.e. less correlated) skill of the IN forecasts. The sampling variances of the natural measure score series (blue points) in Fig. 3 differ because the forecasts are constructed using two different forecast systems, even though evaluation is performed with the same observational dataset.

The effect of serial dependence on estimation of the skill of the PDA forecast system is also demonstrated in Fig. 4. The upper plot compares the probability coverage of 95% confidence intervals for both the serially dependent and independent PDA forecast skill series shown in Fig. 3 above. The probability coverage is determined in a similar manner to Wilks (2010) by the relative frequency over 2^8 replications of the confidence interval containing an asymptotic estimate of the true score $\widehat{IGN_e}$[6]. The lower plot illustrates the relationship between forecast skill and confidence interval width. The insufficient probability coverage of confidence intervals, particularly at smaller sample sizes, and overconfidence in skill estimation are clearly evident in the plots.

Figure 5 shows the effects of sampling variance inflation using ellipses with semi-major and semi-minor axes determined by $s^2_{\widehat{IGN_e}}$ and $s^2_{\widehat{IGN^*_e}}$, respectively, for each sample size. The eccentricity of these ellipses is a function of the ratio of the two sampling variances, that is, it is a measure of the convergence of the sampling variances with increase in sample size. The expectation is therefore that, as sample size increases, the eccentricity of the ellipses approaches zero, that of a perfect circle.

Linear correlation in a score time series as a result of evaluating forecasts with serially dependent target data implies that the autocovariance of the score is expected to be non-zero. The autocovariance $R(\tau)$ of a score S is

$$\begin{aligned} R(\tau) &= E[(S_t - E[S_t])(S_{t+\tau} - E[S_{t+\tau}])] \\ &= E[S_t S_{t+\tau}] - E[S_t]E[S_{t+\tau}] \\ &\neq 0, \end{aligned} \tag{9}$$

where $E[S_t]$ and $E[S_{t+\tau}]$ are the means of the score distributions at time t and time $t+\tau$ (lag τ) respectively. The non-zero result arises under serial dependence since, only where S_t and $S_{t+\tau}$ are independent, is it true that

$$E[S_t S_{t+\tau}] = E[S_t]E[S_{t+\tau}]. \tag{10}$$

Although the inflation of the scoring rule sampling variance induced by serial dependence is demonstrated for probabilistic forecasts in this section, it can easily be shown for point forecasts. The influence of observational noise on the serial correlation of the scores is not quantified here. The effect of serial correlation is expected to be stronger in the absence of uncorrelated observational noise. Of course, even the probability distributions from a forecast system with a perfect dynamical model can be corrupted by failure to properly account for observational noise; this can, in turn, introduce serial dependence in cases where none would be found given the True distribution.

3.2. *Case study 2: AR(1) process*

While it is straightforward to demonstrate the misleading effect of serial dependence on forecast skill estimation, as in Section 3.1, serial dependence in forecast target data is not a sufficient condition for the presence of serial dependence in forecast skill. The case is now illustrated where the sampling distribution of a scoring rule is time-independent, and hence, the scores are serially

[5] With sample size $N = 2^{14}$. The difference in the degrees of lag 1 autocorrelation between the target data and scores is attributable to the fact that ignorance is a function of the joint distribution of forecasts and target data, and this is applicable to any scoring rule (e.g. the Brier score, see Wilks, 2010). Note that the autocorrelation function tends to be negatively biased at small sample sizes where comparing degrees of serial dependence for different statistics (see DeCarlo and Tyron, 1993)

[6] This differs to Wilks (2010) where the true score is known analytically in terms of the model's parameters

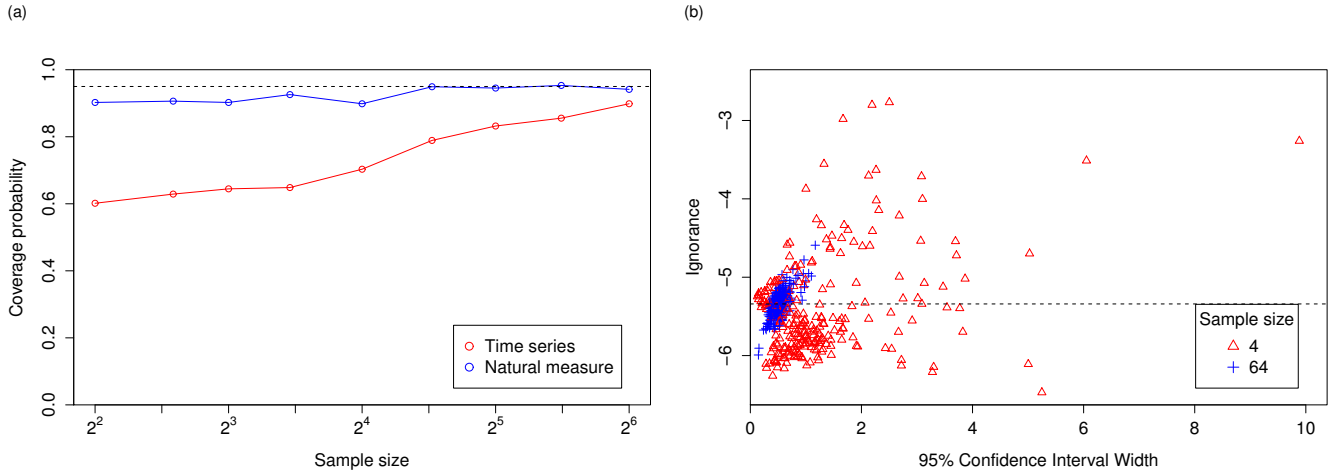

Figure 4. Results for Lorenz63: plot (a) shows the poorer probability coverage for the time series (red circles) compared to the IID series (blue circles) for smaller sample sizes, but approaches that of the IID series and the correct 95% coverage (dashed line) with increasing sample size. The slightly worse probability coverage of the confidence intervals for the IID series at smaller sample sizes reflects poorer estimates of skill, even without the presence of linear correlation. Plot (b) shows the distribution of time series skill estimates against their corresponding confidence interval widths, and how skill estimation is less precise for smaller sample sizes and better forecast skill. Note that the estimate for the true ignorance is $\widehat{IGN_e} = -5.34$ with sample size $N = 2^{14}$.

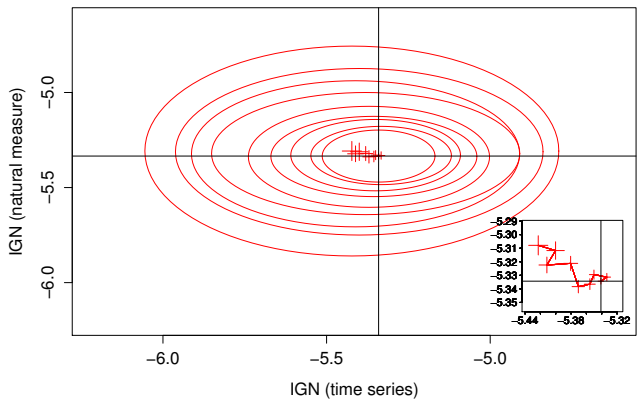

Figure 5. Results for Lorenz63: ellipses with semi-major axis determined by $s^2_{\widehat{IGN_e}}$ and semi-minor axis determined by $s^2_{\widehat{IGN^*_e}}$, and corresponding mean estimates plotted at their centres (coordinates are $\{\widehat{IGN_e}, \widehat{IGN^*_e}\}$; '+' symbols which shrink with increase in sample size). Each ellipse represents a sample size corresponding to Fig. 3. The black vertical and horizontal lines denote the mean estimates $\widehat{IGN_e} = \widehat{IGN^*_e} = -5.34$ with sample size $N = 2^{14}$.

independent, even where the forecast target data are not (see Table 1; top right). Without the inflationary effect induced by serial dependence on the sampling variance of the scoring rule, ESS corrections are not required and statistical inference of forecast skill can be made under the assumption of serial independence.

Consider a time series of target data s_t generated from a *first-order autoregressive* (AR(1)) process[7], first introduced by Yule (1927) to model sunspots. An observation s_t at time t is given by

$$s_t = \varphi s_{t-1} + \epsilon_t, \tag{11}$$

where $\epsilon_t \sim \mathcal{N}(0, \sigma^2_\epsilon)$ is the normally distributed random noise component of the AR(1) process. Since the observational noise ϵ_t is a Gaussian process, the target data s_t are also Gaussian distributed. The model parameter φ controls the degree of autocorrelation in the time series, and the process is *weak-sense stationary* for values $|\varphi| < 1$, implying that the mean $E[s_t]$ and covariance $Cov[s_t, s_{t+\tau}]$ are constant with respect to time. In that case, as φ approaches a value of 1, the dependence of s_t on the previous observation s_{t-1} increases.

Let a 1-step ahead singleton probabilistic dynamical forecast $p_t(x)$ of the system state at time t be constructed from an imperfect model based on the observation s_{t-1} so that

$$p_t(x) = \frac{1}{\sqrt{2\sigma^2_\epsilon \pi}} e^{-\frac{(x-\varphi s_{t-1})^2}{2\sigma^2_\epsilon}}. \tag{12}$$

Random draws from the forecast PDF are, like the target data, Gaussian distributed, and exhibit a similar degree of linear correlation determined by the parameter φ.

Figure 6 shows the IGN sampling variances for the serially dependent time series and random resampled series over increasing sample sizes for $\varphi = 0.9$ and $\sigma_\epsilon = 1.0$. Also shown are 95% uncertainty intervals constructed from IGN sampling variances for a time series of forecast-observation pairs where both variables are standard normal distributed and IID, implying that they are also jointly normally distributed. Hence, the resulting time series of scores is serially independent. The containment of the time series and random resampled IGN estimates within the uncertainty intervals indicates that their respective sampling variances are statistically indistinguishable both from the sampling variances of the standard normal forecast IGN estimates, and from each other. Hence, the scores of both the serially dependent forecasts and serially independent random resampled forecasts are Gaussian distributed and independent (IID) (i.e. the score distributions are non-time dependent). The serial independence of the scores is reflected by the lack of inflation of the sampling variances, and satisfies Eqn. (10). The **absence** of linear correlation in the score time series is also evident in Fig. 7 where a delay plot indicates a negligible linear relationship between ignorance at time $t-1$ and time t. To see this precisely, note that

$$p_t(x) \propto e^{-\frac{(x-\varphi s_{t-1})^2}{(2\sigma^2_\epsilon)}}, \tag{13}$$

so that

$$\begin{aligned} IGN(p_t(x), X = s_t) &\propto (s_t - \varphi s_{t-1})^2 \\ &= (\varphi s_{t-1} + \epsilon_{t-1} - \varphi s_{t-1})^2 \\ &= \epsilon^2_{t-1}. \end{aligned} \tag{14}$$

[7] the results hold for AR processes of any order p. $p = 1$ is taken for simplicity.

In this case, the time series of ignorance scores corresponds to squared, independent Gaussian noise (an IID χ^2 distribution), and is thus serially independent. When the skill of the forecast distribution is independent of the state of the system, as in the AR(1) case, then there is no serial dependence in the sample skill time series even if there is serial correlation in the predictand.

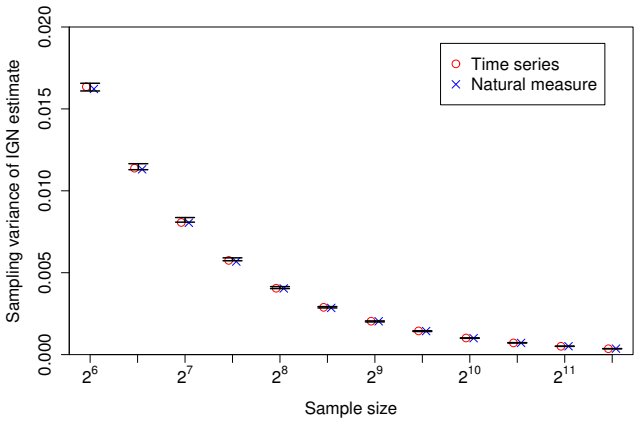

Figure 6. Results for AR(1) process: sampling variances of 2^8 ignorance estimates computed from forecasts of a serially dependent time series of AR(1) target data ($r_1(s) \approx 0.9$; red circles) and an IID randomly resampled series of scores (blue circles). All points lie within 95% uncertainty intervals constructed from 2^7 estimates of the sampling variance of standard normal distributed distributed forecasts showing that there is no statistically significant difference between either of the sampling variances and uncorrelated standard normal distributed forecasts. Note that the red and blue circle points have been shifted left and right respectively for clarity.

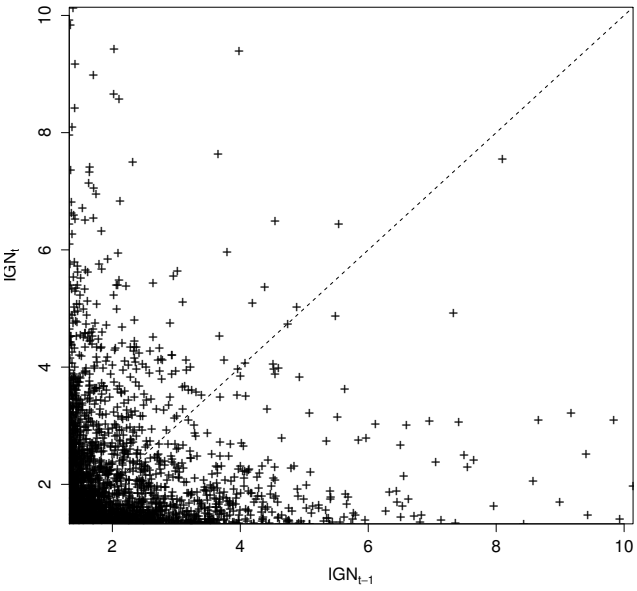

Figure 7. Results for AR(1) process: delay plot of scores at $t-1$ and t in a single ignorance time series of sample size $N = 2896$ computed from serially dependent target data ($\varphi = 0.9$). The lack of linear trend reflects the absence of linear correlation in the time series ($r_1(IGN) \approx 0$).

This example of non-effect of serial dependence on forecast skill estimation is now followed by a counterexample where the serial dependence in the time series of target data and forecasts generated under the AR(1) process **is** present in the corresponding time series of scores. In the case that the forecast distribution is independent of the state of the system (as it is when the forecast is always climatology) while there is serial correlation in the predictand, then there is serial dependence in the skill time series.

Consider a "perfect" climatological Gaussian forecast of the AR(1) system state with population mean $E(s_t) = 0$, which, expressed as a random variable, is given by

$$Y_{clim} \sim \mathcal{N}\left(0, \frac{\sigma_\epsilon^2}{1-\varphi^2}\right). \tag{15}$$

Hence, the forecast distribution is state (and time) independent. It is important to distinguish here between the climatological forecasts and the randomly resampled forecasts since they are constructed differently, and are effectively evaluated with two different (i.e. serially dependent and independent) series of target data.

The sampling variances of the IGN estimates for the time dependent forecasts, climatological forecasts, and randomly resampled forecasts (natural measure) over increasing sample sizes are shown in Fig. 8. The inflationary effect on the sampling variance of the climatological forecast skill statistics is clearly visible, and is attributable to the fact that a time independent forecast PDF is being evaluated with serially dependent data, resulting in serial dependence in the score time series ($r_1(IGN_{clim}) \approx 0.81$).

Respective demonstrations of accurate (serial independence of the time dependent forecast skill scores) and inaccurate (serial dependence of the time independent climatological forecast skill scores) estimates of forecast skill with a single data-generating system highlight the importance of understanding how serial dependence present in forecast target data may or may not be likewise present in corresponding time series of scores. Both the data-generating system and the forecast system should be considered when determining whether serial dependence will have an impact on the accuracy of forecast skill estimates.

The results in this section demonstrate that there are forecasting scenarios where serial dependence in forecast target data does not result in biased estimates of forecast skill. Even with an arbitrarily high degree of lag 1 autocorrelation in the time series (illustrated here with $r_1(s) \approx 0.9$), there is no significant autocorrelation in the scoring rule time series ($r_1(IGN) \approx 0$), and no induced inflation of the score sampling variance. The absence of serial dependence in forecast skill scores, of course, can also be shown for a nonlinear stochastic process.

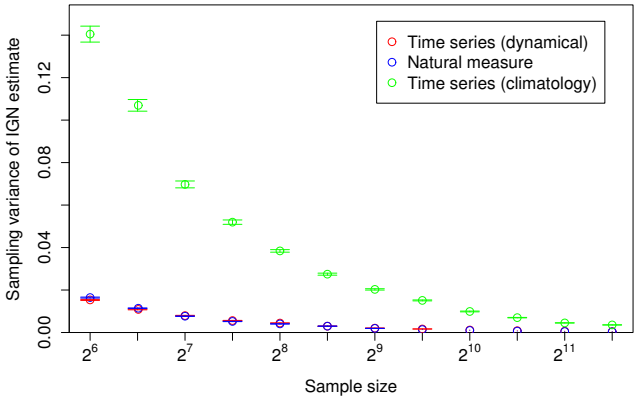

Figure 8. Results for AR(1) process: sampling variances of 2^8 ignorance estimates computed from dynamical (red circles) and climatological (green circles) forecasts - both evaluated with the serially dependent time series of observations ($r_1(s) \approx 0.9$) - and an IID randomly resampled series of scores ($r_1(IGN) \approx 0$; blue circles). All sampling variances are plotted with $5\% - 95\%$ uncertainty intervals. There is a clear inflation of the climatological forecast ignorance sampling variance ($r_1(IGN_{clim}) \approx 0.81$) which is explained by a time independent forecast PDF being evaluated with autocorrelated observations.

3.3. *Case study 3: logistic map*

The misleading impact of serial dependence on forecast skill estimation is not restricted to scenarios where the forecast target exhibits linear correlation. Temporal dependence in a forecast target time series with no linear correlation can lead to linear correlation of the corresponding scores (see Table 1; bottom left). Establishing the statistical significance of skill estimates in such a case is problematic given that serial dependence need not be reflected in the autocorrelation function.

The logistic map is a 1-dimensional nonlinear dynamical system with zero autocorrelation at all lags. It was first proposed by Ulam (1947) as a pseudo-random number generator, and popularised by May (1976) as an educational tool and a simple ecological model of population dynamics. The mathematical form of the logistic map is expressed as

$$x_t = f(x_{t-1}) \tag{16}$$

$$= a x_{t-1}(1 - x_{t-1}), \tag{17}$$

where $x \in (0, 1)$ represents the state of the map. x_t is delta correlated in time. Consider a simple truncated noise forecast

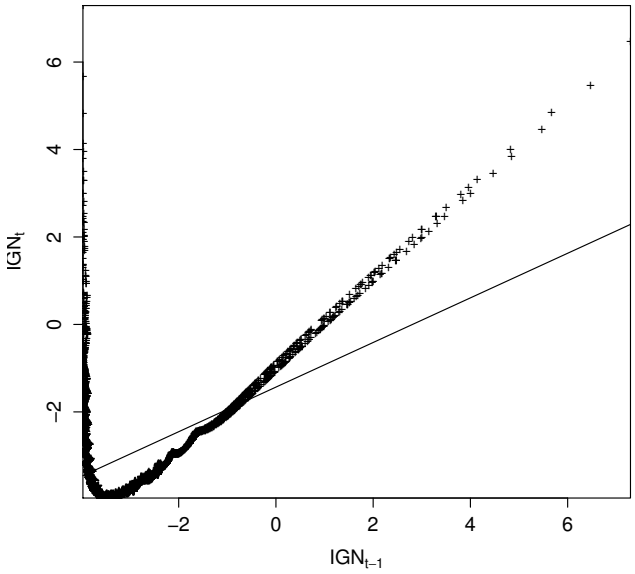

Figure 9. Results for the logistic map: empirical ignorance of 2^{12} forecasts at t and $t-1$ evaluated on x ($\alpha = 4.0$). The ignorance scores are lag 1 autocorrelated ($r_1(IGN) \approx 0.52$) while the true values, x, are not ($r_1(x) \approx 0$). A linear fit is also shown indicating a degree of linear correlation in the score time series.

system for x where state estimation, as in Section 3.1, is subject to observational uncertainty, but the observation, or initial condition for the forecast, at time t is a quantised approximation of the truth x_t, so that

$$s_t = \lfloor x_t \cdot 10^2 \rfloor / 10^2 + 0.005. \tag{18}$$

A forecast system for the logistic map using a singleton ensemble is used[8]. Let a 1-step ahead singleton member probabilistic dynamical forecast of the system state at time t be Gaussian distributed, and defined as

$$p_t(x) = \frac{1}{\sqrt{2\sigma_f^2 \pi}} e^{-\frac{(x - f(s_{t-1}))^2}{2\sigma_f^2}}. \tag{19}$$

Truncation of the true value allows an analytical approach to forecast evaluation, constraining the target data, and increasing linear correlation in the forecast skill scores.

[8]This system could no doubt be improved, but is sufficient for its purpose here.

Figure 9 shows a 1-step delay plot of IGN scores of single iteration forecasts of the logistic map with parameter value $a = 4.0$ evaluated against the unconditional climatological distribution (see Eqn. (8)), defined by the natural measure of the logistic map as

$$p_{clim}(x) = \frac{1}{\pi \sqrt{x(1-x)}}. \tag{20}$$

The kernel width is $\sigma_f = 0.04$, and empirical ignorance is evaluated here on the true value, x. The initial state is uniformly sampled from the support of the logistic map $x \in (0, 1)$, and the transient is discarded.

Plotting the ignorance as a function of the target (that is, IGN_t vs x_t) and ignorance as a function of the initial conditions (that is IGN_t as a function of x_{t-1}) demonstrates that the points with IGN_{t-1} small and IGN_t large in Fig. 9 all pass near $x \sim 0.5$. This is a result of the fixed kernel width in the forecast system. It is these points in Fig. 9 which decrease the linear correlation of IGN scores, the serial dependence clearly remains very high. Improving the forecast system by allowing a variable kernel width (allowing the width to decrease in the regions where the map is contracting) would yield a forecast system with both a lower (better) ignorance score and higher linear correlation.

Linear correlation in the score time series is evident from the linear fit and the lag 1 ACF value $r_1(IGN) \approx 0.52$ computed from a time series of 2^{12} iterations of the map. The effect of the correlation is to inflate the corresponding sampling variances of ignorance estimates computed from the zero autocorrelated time series of the target variable with respect to the random resampled IID series, shown in Fig. 10.

4. Effective Sample Size

An important contribution by Wilks (2010) is the derivation of empirical *effective sample size* (ESS) corrections to account for the combined inflationary effects of serial dependence, forecast skill, and forecast calibration (and event frequency in a binary predictand scenario) on the sampling variances of scoring rules in one case of interest. ESS corrections are formulated from the ratio of the analytical sampling variance to the empirical sampling variance of the Brier score (and Brier skill score). The analytical solutions of the sampling variances are derived under the assumption that forecast-observation pairs are IID (Bradley et

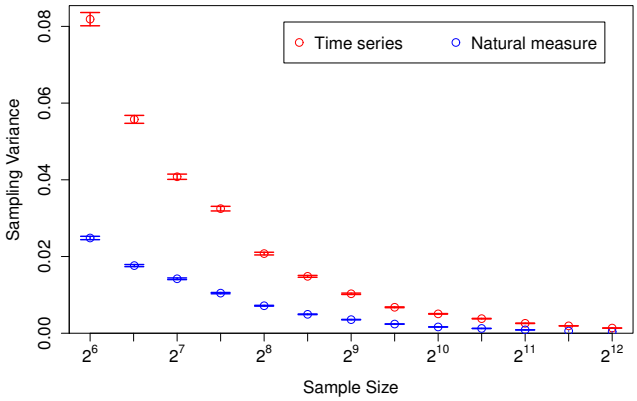

Figure 10. Results for the logistic map: sampling variances of 2^8 IGN estimates computed from forecasts of a time series of logistic map ($a = 4.0$) target data ($r_1(IGN) \approx 0.52$; red circles) and an IID randomly resampled series of scores ($r_1(IGN) \approx 0$; blue circles), both with $5\% - 95\%$ uncertainty intervals. There is a clear deflation of the sampling variances of the score time series up to at least a sample size of 2^{12} showing the case where there is no linear correlation in the target data yet linear correlation in the scores.

al., 2008), and can be used to measure inflation of the empirical sampling variances under serial dependence. Wilks (2010) utilises the decomposition of the Brier score sampling variance into the moments of the joint distribution of forecasts and target data which are expressible in terms of the parameters of the "linear-calibration/beta-refinement" (LCBR) probability model, allowing for derivation of analytical expressions for effective sample size (ESS) corrections. Derivation of such analytical expressions has not been possible in this study because neither is the IGN sampling variance dependent on the moments of the joint distribution, nor are these moments expressible in terms of the parameters of the Lorenz63 system, AR(1) process, or logistic map (or, of course, in any real-world system since the precise parameters are unknown).

In practice, a series of steps can be undertaken to determine approximate ESS corrections, and make reliable estimates of forecast skill and the statistical confidence of that skill. Firstly, to detect whether serial dependence is inducing inflation of the sampling variance, empirical estimates of the sampling variance made from the score time series can be compared with those for a serially independent series constructed using a random resampling method as in Sections 3.1, 3.2, and 3.3 (see also Efron, 1981; Wilks, 2011). These two sets of sampling variances can be plotted for a range of sample sizes, as in Fig. 3. Secondly, the ratio of the sample sizes of the time series and randomly resampled series of scores which correspond to a given score sampling variance is equal to $\frac{N'}{N}$, and the actual ESS correction is given by the difference $N - N'$. For example, referring to Fig. 3, a sampling variance of $s^2_{\widehat{IGN_e}} \approx 0.08$ corresponds to a sample size $N' \approx 2^5$ for the time series and $N \approx 2^4$ for the natural measure. This indicates a required increase in sample size of $\Delta N \approx 16$ to achieve accurate estimation of the sampling variance, and hence, significantly improved probability coverage of confidence intervals.

Where reliable estimates of forecast skill are the ultimate aim, the convergence of the sampling variance of the serially dependent scores on that of the serially independent scores determines the sample size required. Referring again to Fig. 3, the $5\% - 95\%$ uncertainty intervals for sampling variances of the forecast ignorance estimates do not quite overlap sample size $N = 2^6$ so a larger sample size is required to be certain of obtaining correct estimates under serial dependence in this case. At the point at which the uncertainty intervals do overlap, the forecast skill estimates under serial dependence and serial independence can be considered to converge indicating that the estimates are accurate[9]. The above procedure is summarised as follows:

1. construct independent series via random resample method
2. compare the score sampling variances from the time series with those from independent draws as a function of sample size
3. check for inflation of the sampling variance of the score estimates against that of the independent series score estimates
4. if inflation is detected, and where sample size allows, determine which sample size is sufficient for convergence of the score sampling variance under serial dependence and serial independence to achieve accurate estimates of forecast skill

5. Discussion

Serial dependence is a longstanding challenge in the estimation of forecast skill. Inspired by Wilks (2010) demonstration of the effect in probability of precipitation forecasting, the impact of serial dependence has been shown to be nontrivial even in cases where it might not have been expected, given the properties of the predictand. This fact suggests testing for serial dependence in every estimate of forecast skill, and a simple, straightforward initial test has been demonstrated. Figures 3, 6, 8, and 10 illustrate the comparison of time series estimates (red) with estimates from independent sampling of the natural measure (blue). The difference of these (red and blue) estimates in Figures 3 and 10 clearly signal the presence of serial dependence. Three of the four possible cases of linear correlation either present or absent in the target data and forecast skill scores have been illustrated in these case studies. The results demonstrate not only how serial dependence in an observation time series can lead to a (lesser) degree of serial dependence in the corresponding score time series, resulting in inflation of the score sampling variance and misestimation of skill, but also explain how the presence of serial dependence in target data is not a sufficient condition for the effects to occur. The conclusions reached from the case studies are summarised below and in Table 1.

Linear correlation in forecast target and linear correlation in scores - Lorenz63 (Section 3.1)

The inflationary effect on the variances of the ignorance score's sampling distribution, previously examined by Wilks (2010), has been emulated here with the Lorenz63 system, and shown to increase with forecast skill. Of course, the effect materialises for any scoring rule and for any statistic computed from serially dependent data, but the results have demonstrated that the effects of serial dependence on ignorance are weaker than they are on the sample mean of the target data (for which the effective sample size is determined by $N' \simeq N(1 - r_1)/(1 + r_1)$ under a similar assumption about the correlation structure of the time series as noted by Wilks (2011)). Hence, the effects on skill estimation may not be so severe in real-world forecasting cases where the data are not highly serially dependent. Improvements in forecast systems over time (Bosart, 2003; Homar et al., 2006; Stuart, 2006; Ruth, 2009; Novak et al., 2014), and hence forecast skill, may be expected to lead to increases in the effect, however, which are substantial enough to warrant making sample size corrections.

Linear correlation in forecast target and no linear correlation in scores - AR(1) (Section 3.2)

A stochastic AR(1) process demonstrates how serial dependence in forecast target time series need not imply correlation in corresponding forecast scores. Score sampling variance inflation and the misleading effects on forecast skill estimation do not occur if the distribution of scores is not time-dependent. Conversely, the inflationary effect has been shown to occur for time-independent climatological forecast because the score realisations are only dependent on the observation s_t, which are serially dependent.

No linear correlation in forecast target and linear correlation in scores - logistic map (Section 3.3)

The misleading effect on forecast skill estimation has also been shown to occur in the case of the logistic map where the forecast target time series is delta correlated. In such a scenario, a forecaster may not even be aware that their estimates of forecast skill are inaccurate so a check for autocorrelation in the score time series (or comparison of serially dependent and serially independent score sampling variances) may be worthwhile.

The preceding results reveal a previously unreported complexity to the effects/non-effects of serial dependence on forecast skill estimation, and highlight how one might choose to exercise

[9] To find this point of convergence, extrapolating lines could be fitted to the two plots for example.

caution to avoid misidentifying-identifying the best forecast system. To compensate for the effects of serial dependence, effective sample size (ESS) corrections which are dependent on the ratio of analytical to empirical score sampling variances can be made to attain sufficient sample sizes for accurate skill estimation. Where analytical solutions for score sampling variances are available, one can employ the method of Wilks (2010) to determine ESS corrections and which sample sizes are sufficient for accurate estimates of forecast skill. In practice, these analytical solutions are generally not available, however, so some ad hoc procedure is required, such as comparing the sampling variances of score estimates made from the time series and a randomly resampled series. The suggestion proffered here is that, given detection of the effect, one can compensate by applying ESS corrections, but, without detection, one cannot be sure if there is an effect or not, and so may choose to either increase to larger sample sizes to detect the effect, or settle on their estimation of skill.

While the case studies presented in this paper provide new insights into the effects of dependence in forecast target data on forecast skill estimation, the investigations are limited to system-model configurations with single parameter sets, and hence, time-series structures. Future research might extend to techniques for the assessment of the effect of both higher-order serial dependence and spatial dependence of forecast skill estimation.

A. Dynamical Systems and Forecast Construction

A.1. Lorenz63 System

The Lorenz63 system (Lorenz, 1963) is a three dimensional dynamical system defined by a set of three ordinary differential equations (with respect to time) given as

$$\dot{x} = -\sigma x + \sigma y \tag{21}$$

$$\dot{y} = -xz + rx - y \tag{22}$$

$$\dot{z} = xy - bz, \tag{23}$$

where σ is the Prandtl number, r is the Rayleigh number, and b is the system parameter. The standard parameter values are: $\sigma = 10$, $r = 28$, and $b = 8/3$ (Sprott, 2003), and the initial conditions are set to $\{x_0 = 0, y_0 = -0.01, z_0 = 9\}$. Numerical solutions are obtained using a fourth order Runge-Kutta time stepping scheme (Press et al., 2007), with time step $h = 10^{-2}$.

The forecasts are constructed here from a core ensemble model using kernel dressing and blending (Bröcker and Smith, 2008). The kernel dressing approach here is to transform an ensemble of model simulations $\mathbf{x} = x_1, \ldots, x_M$ into a PDF $(y|\mathbf{x}, \sigma)$ by assigning a linear combination of kernels centred on each ensemble member x_j. The kernel dressed PDF is given as

$$\hat{p}(y|\mathbf{x}, \sigma) = \frac{1}{M\sigma} \sum_{j=1}^{M} K\left(\frac{y - x_j}{\sigma}\right), \tag{24}$$

where σ is the strictly positive bandwidth or smoothing parameter, and the kernel K is represented by a standard Gaussian density

$$K(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}. \tag{25}$$

Ideally, the optimal bandwidth is selected so that the divergence of the estimate $\hat{p}$ from the true p is minimised, that is $d(\hat{p}, p) = ||\hat{p} - p||$ where $d(\hat{p}, p)$ is some measure of the divergence. Obviously, measuring the divergence is not possible since p is unknown. The best alternative is to deploy an automated selection method such as K-fold cross-validation or "plug-in" selection (Hall, Marron, and Park, 1992). K-fold cross-validation is useful method for fitting and validating a model where datasets are limited in size (Picard and Cook, 1984; Hastie and Tibshirani, 2009). The data is partitioned into K roughly equal sized subsets which are, in turn, used to validate the model which has been fitted with the other $K - 1$ subsets. Leave-one-out cross-validation (i.e. K-fold cross-validation (CV) with $K = N$) is preferred where datasets are limited in size. Where larger synthetic datasets are available, as is the case here, 2-fold cross-validation is performed.

The optimised kernel width $\hat{\sigma}$ of the forecast PDF is obtained by minimising some cost function, ideally a proper probabilistic forecast scoring rule according to

$$(\hat{\sigma}) := \arg\min_{\sigma} -\frac{1}{N} \sum_{i=1}^{N} S(\hat{p}(Y_i; \sigma)), \tag{26}$$

where a scoring rule S is evaluated over a sufficiently large number N of target data Y_i.

Acknowledgements

This research was supported both by the London School of Economics Grantham Research Institute and the Economic and Social Research Council Centre for Climate Change Economics and Policy, funded by the Economic and Social Research Council and Munich Re. The authors wish to thank Erica Thompson and Hailiang Du for their useful critique. The Lorenz63 forecast-observation data utilised in Section 3.1 have been kindly provided by Ed Wheatcroft and Hailiang Du at the Centre for the Analysis of Time Series, London School of Economics. Both of the authors have no conflicts of interest to declare with the publication of this research.

References

Albers W. 1978. Testing the mean of a normal population under dependence. The Annals of Statistics. **6**:1337-1344.

Bosart LF. 2003. Whither the weather analysis and forecasting process? Weather and Forecasting, **18**:520-529.

Bradley AA, Schwartz SS, and Hashino T. 2008. Sampling uncertainty and confidence intervals for the brier score and brier skill score. Weather and Forecasting, **23**:992-1006.

Bröcker J and Smith LA. 2007. Scoring probabilistic forecasts: The importance of being proper. Weather and Forecasting, **22**:382-388.

Bröcker J and Smith LA. 2008. From ensemble forecasts to predictive distribution functions. Tellus A, **60**:663-678.

DeCarlo LT and Tyron WW. 1993. Estimating and testing autocorrelation with small samples: A comparison of the c-statistic to a modified estimator. Behaviour research and therapy, **31**:781-788.

Du H. 2009. Combining Statistical Methods with Dynamical Insight to Improve Nonlinear Estimation. PhD thesis, London School of Economics and Political Science.

Du H and Smith LA. 2012. Parameter estimation through ignorance. Physical Review E, **86**:016213.

Du H and Smith LA. 2014. Pseudo-orbit data assimilation. part i: The perfect model scenario. Journal of the Atmospheric Sciences, **71**:469-482.

Efron B. 1981. Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. Biometrika, **68**:589-599.

Ferro CAT. 2007. Comparing probabilistic forecasting systems with the brier score. Weather and Forecasting, **22**:1076-1088.

Fraser AM and Swinney HL. 1986. Independent coordinates for strange attractors from mutual information. Physical Review A, **33**:1134-1140.

Gneiting T and Raftery AE. 2007. Strictly proper scoring rules, prediction, and estimation. **102**:359-378.

Good IJ. 1952. Rational decisions. Journal of the Royal Statistical Society, **14**:107-114.

Good IJ. 1983. Good Thinking: The Foundations of Probability and Its Applications. University of Minnesota Press.

Hall P, Marron JS, and Park BU. 1992. Smoothed cross-validation. Probability Theory and Related Fields, **92**:1-20.

Hamill TM. 1999. Hypothesis tests for evaluating numerical precipitation forecasts. Weather and Forecasting, **14**:155-167.

Hastie TJ and Tibshirani RJ. 2009. The Elements of Statistical Learning. Springer, 2nd edition edition.

Homar V, Stensrud DJ, Levit JJ, and Bright DR. 2006. Value of human-generated perturbations in short-range ensemble forecasts of severe weather. Weather and Forecasting, **21**:347-363.

Jarman AS. 2014. On the Provision, Reliability, and Use of Hurricane Forecasts on all Timescales. PhD thesis, London School of Economics and Political Science.

Joliffe IT. 2007. Uncertainty and inference for verification measures. Weather and Forecasting, **22**:637-650.

Jones RH. 1975. Estimating the variance of time averages. Journal of Applied Meteorology and Climatology, **14**:159-163.

Judd K and Smith LA. 2004. Indistinguishable states ii. imperfect model scenario. Physica D, 196:224-242.

Leith CE. 1973. The standard error of time-average estimates of climatic means. Journal of Applied Meteorology and Climatology, **12**:1066-1069.

Lorenz EN. 1963. Deterministic nonperiodic flow. Journal of the Atmospheric Sciences, **20**:130-141.

May RM. 1976. Simple mathematical models with very complicated dynamics. Nature, 261:459-467.

Murphy AH and Winkler RL. 1987. A general framework for forecast verification. Monthly Weather Review, **115**:1330-1338.

Novak DR, Bailey C, Brill KF, Burke P, Wallace A, Hogsett WA, Rausch R, and Schichtel M. Precipitation and temperature forecast performance at the weather prediction center. 2014. Weather and Forecasting, **29**:489-504.

Ott E. 2002. *Chaos in Dynamical Systems*. Cambridge University Press, Cambridge, New York, 2nd edition.

Picard RR and Cook RD. Cross-validation of regression models. 1984. **79**: 575-583.

Pinson R, McSharry P, and Madsena H. Reliability diagrams for non-parametric density forecasts of continuous variables: Accounting for serial correlation. Quarterly Journal of the Royal Meteorological Society, **136**:77-90 (2010).

Press WH, Tuekolsky SA, Vetterling WT, and Flannery BP. 2007. *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, 3rd edition.

Roulston MS and Smith LA. 2002. Evaluating probabilistic forecasts using information theory. Monthly Weather Review, **130**:1653-1660.

Ruth DP, Glahn B, Dagostaro V, and Gilbert K. 2009. The performance of MOS in the digital age. Weather and Forecasting, **24**:504-519.

Seaman R, Mason I, and Woodcock F. 1996. Confidence intervals for some performance measures of yes/no forecasts. Australian Meteorological Magazine, **45**:4953.

Seaman R. 1992. Serial correlation considerations when assessing differences in predictive skill. Australian Meteorological Magazine, **40**:227237.

Smith LA. 1999. Uncertainty dynamics and predictability in chaotic systems. Quarterly Journal of the Royal Meteorological Society, **125**: 2855-2886.

Smith LA. 2001. Disentangling Uncertainty and Error: On the Predictability of Nonlinear Systems. In Alistair I. Mees, editor, *Nonlinear Dynamics and Statistics*, chapter 2, pages 31-64. Birkhuser Boston.

Smith LA. 2006. Predictability of Weather and Climate. In T. Palmer and R. Hagedorn, editors, *Predictability past, predictability present*, chapter 10, pages 219-242. Cambridge University Press.

Sprott JC. 2003. *Chaos and Time-Series Analysis*. Oxford university Press, 1st edition.

Stephenson DB, Hannachi A, and O'Neill A. On the existence of multiple climate regimes. Quarterly Journal of the Royal Meteorological Society, **130**: 583-605.

Stuart NA, Market PS, Telfeyan B, Lackmann GM, Carey K, Brooks HE, Nietfeld D, Motta BC, and Reeves K. 2006. The future of humans in an increasingly automated forecast process. Bulletin of the American Meteorological Society, **87**:14971502.

Thiébaux HJ and Zwiers FW. 1984. The interpretation and estimation of effective sample size. Journal of Applied Meteorology and Climatology, **23**:800811.

Trenberth KE. 1984. Some effects of finite sample size and persistence on meteorological statistics. part i: Autocorrelations. Monthly Weather Review, **112**:23592368.

Ulam SM and von Neumann J. 1947. On combination of stochastic and deterministic processes. Bulletin of the American Meteorological Society, 53(6):1120.

Wilks DS. 2010. Sampling distributions of the brier score and brier skill score under serial dependence. Quarterly Journal of the Royal Meteorological Society, **136**: 2109-2118.

Wilks DS. 2011. *Statistical Methods in the Atmospheric Sciences, volume 100 of International Geophysics*. Academic Press, 3rd edition .

Yule GU. 1927. On a method of investigating periodicities in disturbed series, with special reference to Wolfers sunspot numbers. Phil. Trans. R. Soc. A, **226**:267298.