Stochastic Parametrisations and Model Uncertainty in the Lorenz '96 System

By H. M. Arnold[1], I. M. Moroz[2] and T. N. Palmer[1,3]

(1) Atmospheric, Oceanic and Planetary Physics, University of Oxford, OX1 3PU. (2) Oxford Centre for Industrial and Applied Mathematics, University of Oxford, OX1 3LB. (3) European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, RG2 9AX.

Simple chaotic systems are useful tools for testing methods for use in numerical weather simulations due to their transparency and computational cheapness. The Lorenz (1996) system was used here; the full system is defined as "truth", while a truncated version is used as a testbed for parametrisation schemes. Several stochastic parametrisation schemes were investigated, including additive and multiplicative noise. The forecasts were started from perfect initial conditions, eliminating initial condition uncertainty. The stochastically generated ensembles were compared to perturbed parameter ensembles and deterministic schemes.

The stochastic parametrisations showed an improvement in weather and climate forecasting skill over deterministic parametrisations. Including a temporal autocorrelation resulted in a significant improvement over white noise, challenging the standard idea that a parametrisation should only represent sub-gridscale variability. The skill of the ensemble at representing model uncertainty was tested; the stochastic ensembles gave better estimates of model uncertainty than the perturbed parameter ensembles. The forecasting skill of the parametrisations was found to be linked to their ability to reproduce the climatology of the full model. This is important in a seamless prediction system, allowing the reliability of short term forecasts to provide a quantitative constraint on the accuracy of climate predictions from the same system.

Keywords: model uncertainty; stochastic parametrisations; reliability; seamless prediction; ensemble prediction

1. Introduction

The central aim of any atmospheric parametrisation scheme must be to improve the forecasting skill of the atmospheric model in which it is embedded and to represent better our beliefs about the future state of the atmosphere, be this the weather in five days time or the climate in 50 years time. One aspect of this goal is the accurate representation of uncertainty: a forecast should skilfully indicate the confidence the forecaster can have in his or her prediction. There are two main sources of error in atmospheric modelling; errors in the initial conditions and errors in the model's representation of the atmosphere. The ensemble forecast generated should explore these uncertainties, and a probabilistic forecast issued to the user [21]. A probabilistic forecast is of great economic value to the user as it allows reliable assessment of the risks associated with different decisions, which cannot be achieved using a deterministic forecast [22].

Uncertainties in initial conditions, due to limited spatial and temporal distribution of atmospheric measurements, are represented by perturbing the initial conditions of different ensemble members, for example using singular vectors [6], and tracking their subsequent evolution. Model error arises due to the computational representation of the equations describing the evolution of the atmosphere. The atmospheric model has a finite resolution, and sub-grid scale processes must be represented through schemes which often grossly simplify the physics involved. For each state of the resolved, macroscopic variables there are many possible states of the unresolved variables, so this parametrisation process is a significant source of model uncertainty. The large-scale equations must also be discretised in some way, which is a secondary source of error. If only initial condition uncertainty is represented, the forecast ensemble is underdispersive, i.e. it does not accurately represent

the error in the ensemble mean (e.g. Stensrud et al. [33]). The verification frequently falls outside of the range of the ensemble; model uncertainty must be included for a skilful forecast.

Accurate representation of model and initial condition uncertainties will ensure that the resultant ensemble forecast is *reliable*. This refers to the consistency, when statistically averaged, of the forecast and measured probabilities of an event [36]. If a forecasting system is reliable then, from the very definition of reliability, it must have the correct climatology [5]. This motivates the use of reliability of shorter term forecasts as a test for climatological prediction skill. Palmer et al. [25] use the reliability of seasonal forecasts to calibrate climate change projections, particularly concerning precipitation. They argue that it is a necessary requirement that a climate model gives a reliable weather forecast, though this requirement is not sufficient as numerical weather prediction (NWP) models do not include the longer timescale physics of the cryosphere and biosphere which are nevertheless important for accurate climate prediction.

Several ways of representing model uncertainty have been proposed. Multi-model ensembles (MMEs), such as the Coupled Model Intercomparison Project, phase 3 (CMIP3) [18], allow for a pragmatic representation of model uncertainty. While MMEs have been shown to perform better than the best single model in the ensemble [34], they are not able to represent systemic errors which are potentially common to all deterministically parametrised models, and the superensemble remains underdispersive. Furthermore the individual models in a MME are not independent, and the effective number of models in the ensemble is far lower than the total number of models [17, 27]. This lack of diversity adversely affects how well a MME can represent model uncertainty.

A large source of forecast model error is the assumptions built into the parametrisation schemes. The model uncertainty from these assumptions can be explored by using several different parametrisation schemes (*multiparametrisation*) to generate an ensemble of forecasts [11], though such approaches likely suffer from systemic deficiencies. Alternatively, the perturbed parameter approach takes uncertain parameters in the parametrisation schemes and varies their values within their physical range. Perturbing parameters gives a greater control over the ensemble than the multiparametrisation approach [29], though this approach also does not explore structural or systemic errors, as a single base model is used for the experiment.

In this study, stochastic parametrisations are investigated as a way of accurately representing model uncertainty. In their seminal paper, Nastrom and Gage [20] show the presence of a shallow power-law slope in the kinetic energy spectrum in the atmosphere, with no observed scale separation. This lack of scale separation allows errors in the representation of unresolved, poorly constrained small scale processes to infect the larger scales [23]. Therefore, a one-to-one mapping of the large-scale on to the small-scale variables, as is the case in a deterministic parametrisation, seems unjustified. A stochastic scheme in which random numbers are included in the computational equations of motion acknowledges this limitation, and an ensemble generated by repeating a stochastic forecast gives an indication of the uncertainty inherent in the parametrisation process. A stochastic parametrisation must be viewed as providing possible realisations of the subgrid scale motion, whereas a deterministic parametrisation represents the average of all the possible sub-gridscale effects.

There has been significant interest in stochastic parametrisation schemes in recent years, and they have been shown to perfom well in simple systems such as the Lorenz '96 System [7, 13, 35]. Coarse graining studies [31] allow for the development of physically motivated schemes such as the Stochastically Perturbed Parametrisation Tendencies (SPPT) and Stochastic Kinetic Energy Backscatter (SKEB) schemes operationally in use in the European Centre for Medium-Range Weather Forecasts (ECMWF) NWP system [26]. Their inclusion has resulted in a reduction of model biases, and a significant improvement in forecast reliability [2, 3, 26]. A further example of a physically motivated scheme is the stochastic multicloud model for parametrisation of tropical convection proposed by Frenkel et al. [8]. The resultant cloud structure and wave features compare favourably with CRM simulations, especially when compared to the performance of a deterministic parametrisation. Stochastic schemes have also been found to be beneficial in climate simulation and analysis [24].

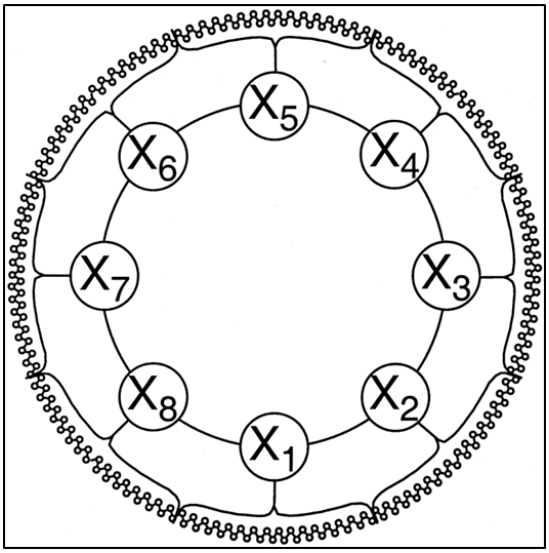

Figure 1: Schematic of the L96 system (taken from Wilks [35]). Each of the $K = 8$ large-scale X variables is coupled to $J = 32$ small-scale Y variables.

The need for stochastic parametrisations has been motivated by considering the requirement that a forecast ensemble should include an estimate of uncertainty due to errors in the forecast model. This experiment seeks to test the ability of a stochastic parametrisation scheme to represent this model uncertainty skilfully. Therefore, perfect initial conditions are used for all ensemble members, removing initial condition uncertainty. The only source of uncertainty in the forecast is model error from truncation and imperfect parametrisation of the unresolved 'Y' variables in the Lorenz '96 system, and from errors in the timestepping scheme. The spread in the forecast ensemble is generated purely from the stochastic parametrisation schemes, so the ability of such a scheme to represent model uncertainty can be rigorously tested. The ability to distinguish between model and initial condition uncertainty can only take place in an idealised "toy model" setting. However, in Wilks [35] and Crommelin and Vanden-Eijnden [7], model uncertainty is not distinguished from initial condition uncertainty in this way, so the ability of stochastic parametrisations to represent model uncertainty was not investigated. In Section 2, we describe the Lorenz '96 System used in this experiment, and in Section 3 we describe the stochastic schemes tested. Sections 4, 5 and 6 discuss the effects of stochastic schemes on short term weather prediction skill, reliability and climatological skill respectively. Section 7 discusses experiments with perturbed parameter ensembles and Section 8 draws some conclusions.

2. The Lorenz '96 System

There are many benefits of performing proof of concept experiments using simple systems before moving to a GCM or NWP model. Simple chaotic systems are transparent and computationally cheap but are able to mimic certain properties of the atmosphere. Crucially, they also allow for a robust definition of "truth", important for development and testing of parametrisations and verification of forecasts. The Lorenz '96 system used here was designed by Lorenz [16] to be a "toy model" of the atmosphere, incorporating the interaction of variables of different scales. It is therefore particularly suited as a testbed for new parametrisation methods which must represent this interaction of scales. The second model proposed in Lorenz [16], henceforth, the L96 system, describes a coupled system of equations for two types of variables arranged around a latitude circle (Figure 1) [16, 35]:

$$\frac{\mathrm{d}X_k}{\mathrm{d}t} = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F - \frac{hc}{b}\sum_{j=J(k-1)+1}^{kJ} Y_j; \qquad k = 1, ..., K \quad (2.1a)$$

$$\frac{\mathrm{d}Y_j}{\mathrm{d}t} = -cbY_{j+1}(Y_{j+2} - Y_{j-1}) - cY_j + \frac{hc}{b}X_{int[(j-1)/J]+1}; \qquad j = 1, ..., JK, \quad (2.1b)$$

Parameter	Symbol	Setting
Number of X variables	K	8
Number of Y variables per X variable	J	32
Coupling Constant	h	1
Forcing Term	F	20
Spatial Scale Ratio	b	10
Timescale Ratio	c	4 or 10

Table 1: Parameter settings used for the L96 system in this experiment

where the variables have cyclic boundary conditions; $X_{k+K} = X_k$ and $Y_{j+JK} = Y_j$. The X_k variables are large amplitude, low frequency variables, each of which is coupled to many small amplitude, high frequency Y_j variables. Lorenz suggested that the Y_j represent convective events, while the X_k could represent, for example, larger scale synoptic events. The interpretation of the other parameters, and the values in this study, are shown in Table 1. The scaling of the variables is such that one model time unit is approximately equal to five atmospheric days, deduced by comparing the error doubling time of the model to that observed in the atmosphere [16].

3. Description of the Experiment

A series of experiments was carried out using the L96 system. Each of the $K = 8$ low frequency, large amplitude X variables is coupled to $J = 32$ high frequency, small amplitude Y variables. The X variables are considered resolved and the Y variables unresolved, so must therefore be parametrised in a truncated model. The effects of different stochastic parametrisations were then investigated by comparing the truncated forecast model to the "truth", defined by running the full set of coupled equations.

Two different values of the timescale ratio, $c = 4$ and $c = 10$, were used in this experiment. The $c = 10$ case was also considered by Wilks [35]. This case has a large timescale separation so can be considered "easy" to parametrise. However, it has been shown that there is no such timescale separation in the atmosphere [20], so a second parameter setting of $c = 4$ was chosen, where parametrisation of the sub-grid is more difficult, but which closer represents the real atmosphere.

(*a*) *"Truth" model*

The full set of equations was run and the resultant time series defined as "truth". The equations were integrated using an adaptive fourth order Runge-Kutta timestepping scheme, with a maximum timestep of 0.001 model time units (MTU). Having removed the transients, the 300 initial conditions on the attractor were selected at intervals of 10 MTU, corresponding to 50 "atmospheric days". This interval was selected to ensure adjacent initial conditions are uncorrelated — the temporal autocorrelation of the X variables is close to zero after 10 MTU. A truth run was carried out from each of these 300 initial conditions.

(*b*) *Forecast model*

A forecast model was constructed by assuming that only the X variables are resolved, and parametrising the effect of the unresolved sub-gridscale Y variables in terms of the resolved X variables:

$$\frac{\mathrm{d}X_k^*}{\mathrm{d}t} = -X_{k-1}^*(X_{k-2}^* - X_{k+1}^*) - X_k^* + F - U_p(X_k^*);\ k = 1, ..., K, \tag{3.1}$$

where $X_k^*(t)$ is the forecast value of $X_k(t)$ and U_p is the parametrised subgrid tendency. Equation (3.1) was integrated using a piecewise deterministic, adaptive second order Runge-Kutta scheme, in which the stochastic noise term is held constant over the timestep. Such a stochastic Runge-Kutta scheme has been shown to converge to the true Stratonovich forward integration scheme, as long as the parameters in such a scheme are used in the

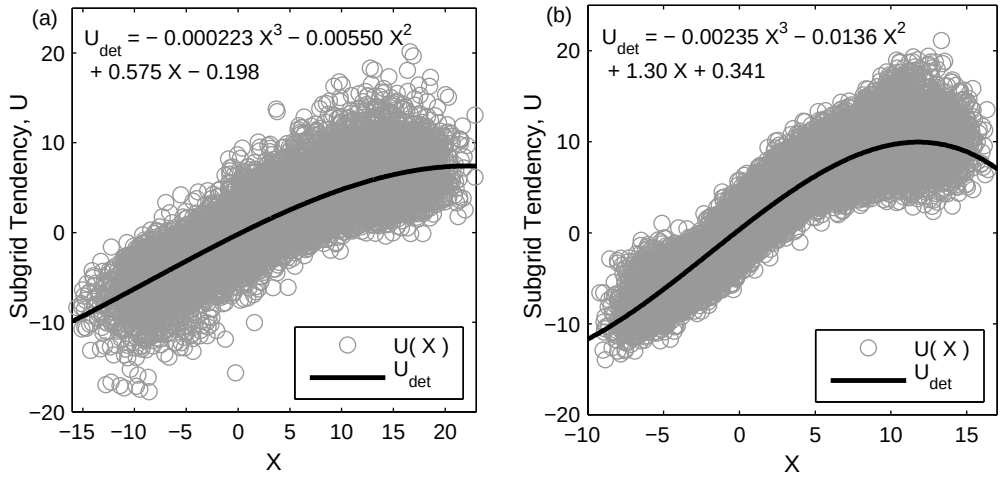

Figure 2: Measured subgrid tendency, U, as a function of the gridscale X variables for the (a) $c = 4$ and (b) $c = 10$ cases. For each case, the data was generated from a long truth integration of 3000 MTU $\approx$ 40 "atmospheric years", sampled at intervals of 0.125 MTU. The solid line on each graph is a cubic fit to the truth data, representing a deterministic parametrisation of the tendencies. There is considerable variability in the tendencies not captured by such a deterministic scheme.

same way they are estimated [9, 10]. This was verified for the different stochastic parametrisations tested. The parametrisations $U_p(X)$ approximate the true sub-grid tendencies,

$$U(X,Y) = \frac{hc}{b} \sum_{j=J(k-1)+1}^{kJ} Y_j, \tag{3.2}$$

which are estimated from the truth time series as

$$U(t) = [-X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F] - \left(\frac{X_k(t+\Delta t) - X_k(t)}{\Delta t} \right). \tag{3.3}$$

The forecast timestep was set to $\Delta t = 0.005$. The L96 system exhibits cyclic symmetry, so the same parametrisation is used for all X_k.

The true sub-grid tendency, U, was plotted as a function of the large-scale X variables for both $c = 4$ and $c = 10$ (Figure 2). This was modelled in terms of a deterministic parametrisation, U_{det}, where

$$U(X) = U_{det}(X) + r(t), \tag{3.4}$$

for

$$U_{det}(X) = b_0 + b_1 X + b_2 X^2 + b_3 X^3, \tag{3.5}$$

and the parameter values (b_0, b_1, b_2, b_3) were determined by a least squares fit to the (X, U) truth data. However, Figure 2 shows significant scatter about the deterministic parametrisation — the residuals $r(t)$ are non-zero. This variability can be taken into account by incorporating a stochastic component, $e(t)$, into the parametrised tendency, U_p.

A number of different stochastic parametrisations were considered. These use different statistical models to represent the subgrid-scale variability due to the truncated Y variables. The different noise models will be described below.

(i) *Additive Noise (A)*

This work builds on Wilks [35], where the effects of white and red additive noise were considered on the skill of the forecast model. The parametrised tendency is modelled as the deterministic tendency and an additive noise term, $e(t)$:

$$U_p = U_{det} + e(t) \tag{3.6}$$

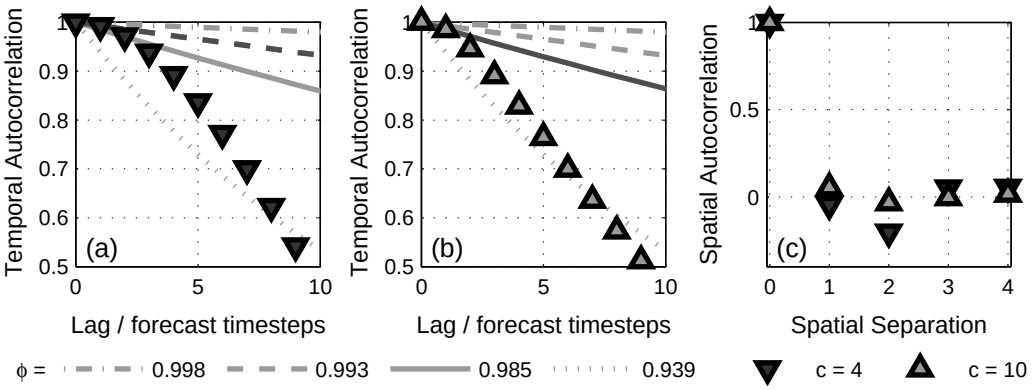

Figure 3: Temporal and spatial autocorrelation for the residuals, $r(t)$, measured from the truth data. Figures (a) and (b) show the measured temporal autocorrelation function for c = 4 and c = 10 respectively as grey triangles. Also shown is the temporal autocorrelation function for an AR(1) process with different values of the ϕ-parameter (grey dot–dash, dash, solid and dotted lines). The fitted AR(1) process is indicated by the darker grey line in each case. Figure (c) shows the measured spatial correlation for the residuals measured from the truth data. The spatial correlation is close to zero for lag $\neq 0$.

The stochastic term, $e(t)$, is designed to represent the residuals, $r(t)$. The temporal and spatial autocorrelation functions for the residuals are shown in Figure 3. The temporal autocorrelation is significant but the spatial correlation is small. Therefore, the temporal characteristics of the residuals are included in the parametrisation by modelling the $e(t)$ as a first order autoregressive (AR(1)) process. The stochastic tendencies for each of the X variables are assumed to be mutually independent. It is expected that in more complicated systems, including the effects of both spatial and temporal correlations will be important to accurately characterise the sub-gridscale variability. A second order autoregressive process was also considered, but fitting the increased number of parameters proved difficult, and the resultant improvement over AR(1) was slight, so it will not be discussed further here.

A zero mean AR(1) process can be written as [36]:

$$e(t) = \phi e(t - \Delta t) + \sigma_e (1 - \phi^2)^{1/2} z(t), \tag{3.7}$$

where ϕ is the first autoregressive parameter (lag-1 autocorrelation), σ_e^2 is the variance of the stochastic tendency and $z(t)$ is unit variance white noise: $z(t) \sim \mathcal{N}(0, 1)$. ϕ and σ_e can be fitted from the truth time series.

(ii) *State Dependent Noise (SD)*

A second type of noise was considered where the standard deviation of additive noise is dependent on the value of the X variable. This will be called *state dependent noise*. It can be motivated in the L96 system by studying Figure 2; the degree of scatter about the cubic fit is greater for large magnitude X values:

$$U_p = U_{det} + e(t), \tag{3.8}$$

where the state dependent standard deviation of $e(t)$ is modelled as

$$\sigma_e = \sigma_1 |X(t)| + \sigma_0. \tag{3.9}$$

As Figure 3 shows a large temporal autocorrelation, it is unlikely that white state dependent noise will adequately model the residuals. Instead, $e(t)$ will be modelled as an AR(1) process:

$$e(t) = \frac{\sigma_e(t)}{\sigma_e(t - \Delta t)} \phi \, e(t - \Delta t) + \sigma_e(t)(1 - \phi^2)^{1/2} z(t), \tag{3.10}$$

where the time dependency of the standard deviation and the requirement that $e(t)$ must be a stationary process have motivated the functional form.

The parameters σ_1 and σ_0 can be estimated by binning the residuals according to the magnitude of X and calculating the standard deviation in each bin. The lag-1 autocorrelation was estimated from the residual time series.

(iii) *Multiplicative Noise (M)*

Multiplicative noise has been successfully implemented in the ECMWF NWP model using the SPPT scheme, and has been shown to improve the skill of the forecasting system. Therefore it is of interest whether a parametrisation scheme involving multiplicative noise could give significant improvements over additive stochastic schemes in the L96 system.

The parametrisation proposed is

$$U_p = (1 + e(t))\, U_{det}, \tag{3.11}$$

where $e(t)$ is modelled as an AR(1) process, given by Equation 3.7.

The parameters in this model can be estimated by forming a time series of the truth "residual ratio", R_k, we wish to represent:

$$R_k + 1 = U \,/\, U_{det}. \tag{3.12}$$

However whenever U_{det} approaches zero the residual ratio tends to infinity. Therefore, the time series was first filtered such that only sections away from $U_{det} = 0$ were considered, and the temporal autocorrelation and standard deviation estimated from these sections.

For multiplicative noise, it is assumed that the standard deviation of the true tendency is proportional to the parametrised tendency, such that when the parametrised tendency is zero the uncertainty in the tendency is zero. Figure 2 shows that multiplicative noise does not appear to be a good model for the uncertainty in the L96 system as the uncertainty in the true tendency is large even when U_{det} is zero. Nevertheless, it was investigated.

(iv) *Multiplicative and Additive Noise (MA)*

Figure 2 motivates a final stochastic parametrisation scheme for testing in the L96 system, which will include both multiplicative and additive noise terms. This represents the uncertainty in the parametrised tendency even when the deterministic tendency is zero. This type of uncertainty has been observed in coarse-graining studies. For example, Shutts and Palmer [31] observed that the standard deviation of the true heating in a coarse gridbox does not go to zero when $\bar{Q}$, the parametrised heating, is zero. This type of stochastic parametrisation can also be motivated by considering errors in the timestepping scheme, which will contribute to errors in the total tendency even if the sub-gridscale tendency is zero.

When formulating this parametrisation, the following points were considered:

1. In a toy model setting, random number generation is computationally cheap. However in a weather or climate prediction model, generation of spatially and temporally correlated fields of random numbers is comparatively expensive, and two separate generators must be used if two such fields are required. It is therefore desirable to use only one random number per timestep so that the parametrisation could be further developed for use in an atmospheric model.

2. The fewer parameters there are to fit, the less complicated the methodology required to fit them, the easier it would be to apply this method to a more complex system such as an atmospheric model. This also avoids overfitting.

We consider the most general form of additive and multiplicative noise:

$$U_p = (1 + \epsilon_m)U_{det} + \epsilon_a = U_{det} + (\epsilon_m U_{det} + \epsilon_a). \tag{3.13}$$

This can be written as pure additive noise:

$$U_p = U_{det} + e(t), \tag{3.14}$$

where

$$e(t) = \epsilon_m(t)U_{det} + \epsilon_a(t). \tag{3.15}$$

Following point (1) above, we assume $\epsilon_m(t)$ and $\epsilon_a(t)$ are the same random number, appropriately scaled:

$$e(t) = \epsilon(t)\,(\sigma_m U_{det} + \sigma_a)\,. \tag{3.16}$$

Full Name	Abbreviation	Functional Form	Measured Parameters	
			c = 4	c = 10
Additive AR(1)	AR1	$U_p = U_{det} + e(t)$ $e(t) = \phi e(t - \Delta t) + \sigma(1 - \phi^2)^{1/2} z(t)$	$\phi = 0.993$ $\sigma = 2.12$	$\phi = 0.986$ $\sigma = 1.99$
State Dependent	SD	$U_p = U_{det} + e(t)$ $e(t) = \left({}^{\sigma_t}/_{\sigma_{t-\Delta t}}\right) \phi e(t - \Delta t) + \sigma_t(1 - \phi^2)^{1/2} z(t)$ where $\sigma_t = \sigma_1 \lvert X(t) \rvert + \sigma_0$	$\phi = 0.993$ $\sigma_0 = 1.62$ $\sigma_1 = 0.0780$	$\phi = 0.989$ $\sigma_0 = 1.47$ $\sigma_1 = 0.0873$
Multiplicative	M	$U_p = (1 + e(t))\, U_{det}$ $e(t) = \phi e(t - \Delta t) + \sigma(1 - \phi^2)^{1/2} z(t)$	$\phi = 0.950$ $\sigma = 0.746$	$\phi = 0.940$ $\sigma = 0.469$
Multiplicative and Additive	MA	$U_p = U_{det} + e(t)$ $e(t) = \epsilon(t)\,(\sigma_m \lvert U_{det} \rvert + \sigma_a)$ where $\epsilon(t) = \epsilon(t - \Delta t)\phi + (1 - \phi^2)^{1/2} z(t)$	$\phi = 0.993$ $\sigma_m = 0.177$ $\sigma_a = 1.55$	$\phi = 0.988$ $\sigma_m = 0.101$ $\sigma_a = 1.37$

Table 2: Stochastic parametrisations of the sub-grid tendency, U, used in this experiment, and the values of the model parameters fitted from the truth time series.

In the current form, Equation (3.16) is not symmetric about the origin with respect to U_{det}. The standard deviation of the stochastic tendency is zero when $\sigma_m U_{det} = -\sigma_a$. Therefore, U_{det} in the above equation will be replaced with $|U_{det}|$:

$$e(t) = \epsilon(t) \left(\sigma_m |U_{det}| + \sigma_a\right), \tag{3.17}$$

where

$$\epsilon(t) = \epsilon(t - \Delta t)\phi + (1 - \phi^2)^{1/2} z(t). \tag{3.18}$$

This does not change the nature of the multiplicative noise as $\epsilon(t)$ is zero mean, but it forces the additive part of the noise to act always in the same direction as the multiplicative. $\epsilon(t)$ is modelled as a AR(1) process of unit variance. The parameters will be fitted from the residual time series.

The different stochastic parametrisations used in this experiment are summarised in Table 2, together with the parameters measured from the truth time series.

4. Weather Forecasting Skill

The stochastic parametrisation schemes were first tested on their ability to predict the "weather" of the L96 system, and represent the uncertainty in their predictions correctly. An ensemble of 40 members was generated for each of the 300 initial conditions on the attractor. Each ensemble member is initialised from the perfect initial conditions defined by the "truth" time series.

Each stochastic parametrisation involves two or more tunable parameters which may be estimated from the "truth" timeseries. Many parameter settings were considered, and the skill of the parametrisation evaluated for each setting using scalar skill scores. The Ranked Probability Score (RPS) evaluates a multi-category forecast. In this experiment, ten categories were defined as the ten deciles of the climatological distribution. The Ranked Probability Skill Score (RPSS) was also calculated with respect to the climatology [36]. The Ignorance score (IGN) evaluates a continuous forecast. Roulston and Smith [30] suggest defining $N + 1$ categories for an ensemble forecast of N members, and approximating the probability density function (PDF) as a uniform distribution between consecutive ensemble members. This approximation is used here. The Ignorance Skill Score (IGNSS) was calculated with respect to climatology. The Brier Score (BS) is used when considering a dichotomous event [4, 36]. The event was defined here as "the X variable is in the upper tercile of the climatological distribution". The Brier Score is mathematically related to the RPS [36], and gave very similar results to the RPS, so is not shown here for brevity. The forecast skill scores are calculated at a lead time of 0.6 units, equivalent to 3 atmospheric days, for each case.

Figure 4 shows the calculated skill scores for a forecast model with an additive AR(1) stochastic parametrisation, for both the $c = 4$ and $c = 10$ cases. The shape of peak in forecasting skill is qualitatively similar in each case, but is lower for the $c = 4$ case. The closer timescale separation of the $c = 4$ case is harder to parametrise, so a lower skill is to be expected.

IGNSS shows a different behaviour to RPSS. Ignorance heavily penalises an underdispersive ensemble, but does not heavily penalise an overdispersive ensemble. This asymmetry is observed in the contour plot for IGNSS — the peak is shifted upwards compared to the peak for RPSS, and deterministic parametrisations have negative skill (are worse than climatology). The very large magnitude and high autocorrelation noise parametrisations are not penalised, but score highly, despite being overdispersive.

The RPSS may be decomposed explicitly into reliability, resolution and uncertainty components [36]. This decomposition demonstrates that the deterministic and low amplitude noise parametrisations score highly on their resolution, but poorly for their reliability, and the converse is true for the large amplitude, highly autocorrelated noise parametrisations. The peak in skill according to the RPSS corresponds to parametrisations which score reasonably well on both accounts.

A number of important parameter settings can be identified on these plots. The first corresponds to the deterministic parametrisation, which occurs on the x-axis where the standard deviation of the noise is zero. The second corresponds to white noise, which

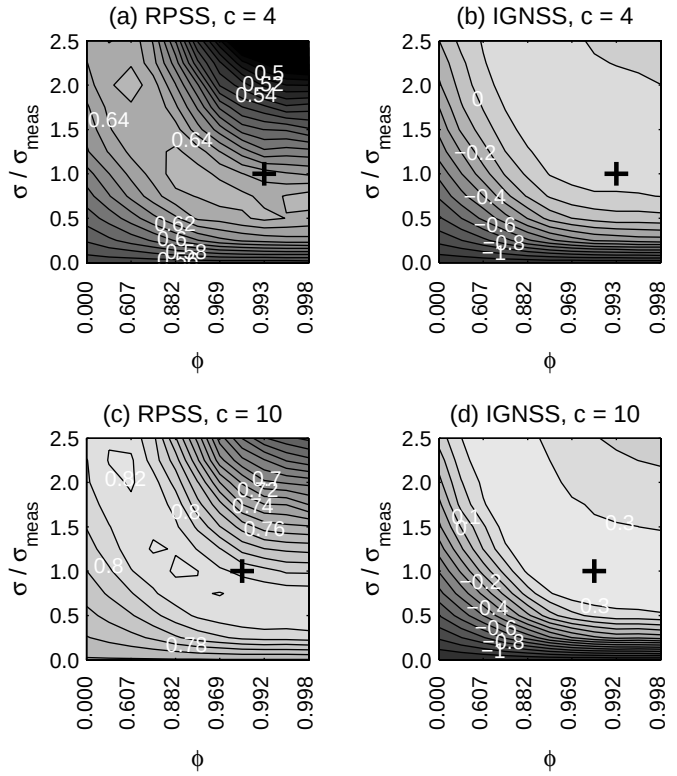

Figure 4: Weather Skill Scores (RPSS and IGNSS) for a forecast model with an additive AR(1) stochastic parametrisation for (a) and (b) the $c = 4$ case, and (c) and (d) the $c = 10$ case. The skill scores were evaluated as a function of the tuneable parameters in the model; the lag-1 autocorrelation, ϕ, and the standard deviation of the noise, σ. The black crosses indicate the measured parameter values. All skill scores are calculated at a lead time of 0.6 model time units (3 atmospheric days).

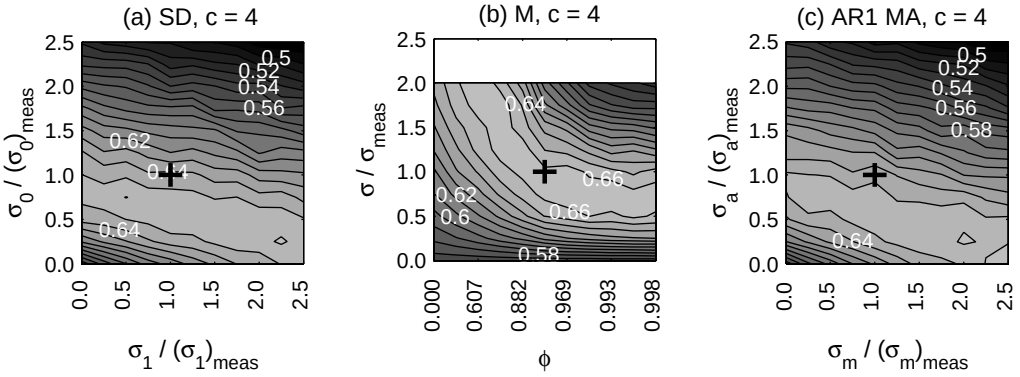

Figure 5: Comparing the RPSS for the $c = 4$ case for the different stochastic parametrisations. The $c = 10$ results are qualitatively similar and are not shown here. These figures are included in the online supplementary material (Supplementary Figure 1). (a) The skill of the state dependant (SD) additive parametrisation is shown for different values of the noise standard deviations, σ_1 and σ_0, with the lag-1 autocorrelation set to $\phi = \phi_{meas}$. (b) The skill of the pure multiplicative (M) noise is shown for different values of the lag-1 autocorrelation, ϕ, and magnitude of the noise, σ. The parametrisation scheme was found to be numerically unstable for high $\sigma > 2\sigma_{meas}$. (c) The skill of the additive and multiplicative (MA) parametrisation is shown for different values of the noise standard deviations, σ_m and σ_a, with the lag-1 autocorrelation set to $\phi = \phi_{meas}$. In all cases, the measured parameters are indicated by the black cross. The skill scores were evaluated at a lead time of 0.6 model time units (3 atmospheric days).

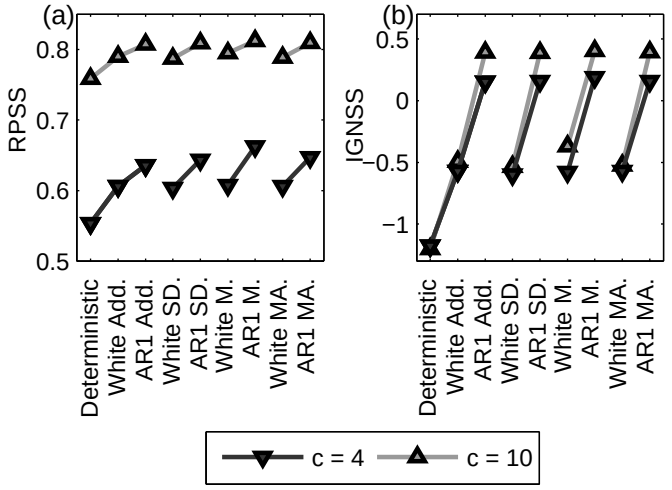

Figure 6: The skill of the different parametrisations according to (a) RPSS and (b) IGNSS are compared for the $c = 10$ and $c = 4$ cases. The parameters in each parametrisation have been estimated from the truth time series. In each case, "White" indicates that the measured standard deviations have been used, but the autocorrelation parameters set to zero.

occurs on the y-axis where the autocorrelation parameter is set to zero. In particular, $(\phi, \sigma/\sigma_{meas}) = (0, 1)$ corresponds to additive white noise with a magnitude fitted to the truth time series. The third setting is the measured parameters, marked by a black cross. Comparing the skill of these three cases shows an improvement over the deterministic scheme as first white noise, then red noise is included in the parametrisation.

The RPSS calculated for the other stochastic parametrisations for the $c = 4$ case is shown in Figures 5. The contour plots for IGNSS are comparable, so are not shown for brevity. The forecasts are more skilful for the $c = 10$ case, but qualitatively similar. They have been included in the supplementary online material (Supplementary Figure 1). For all cases considered, including a stochastic term in the parametrisation scheme results in an improvement in the skill of the forecast over the deterministic scheme. This result is robust to error in the measurement of the parameters — a range of parameters in each forecast model gave good skill scores. This is encouraging, as it indicates that stochastic parametrisations could be useful in modelling the real atmosphere, where temporally and spatially limited, noisy data restrict how accurately these parameters may be estimated.

The results are summarised in Figure 6. For each parametrisation, the value for the measured parameters is shown when both no temporal autocorrelation ("white" noise) and the measured temporal autocorrelation characteristics are used. The significance of the difference between pairs of parametrisations was estimated using a Monte-Carlo technique. Please see the supplementary online material for more details, but for example, there is no significant difference between the RPS for AR(1) multiplicative and additive noise and for AR(1) state dependent noise for the $c = 10$ case, but AR(1) multiplicative noise gave a significant improvement over both of these.

The stochastic parametrisations are significantly more skilful than the deterministic parametrisation in both the $c = 4$ and $c = 10$ cases. For the $c = 4$ case, the more complicated parametrisations show a significant improvement over simple additive, especially the multiplicative noise. For the closer timescale separation, the more accurate the representation of the sub-gridscale forcing, the higher the forecast skill. For the $c = 10$ case, the large timescale separation allows the deterministic parametrisation to have reasonable forecasting skill, and a simple representation of sub-grid variability is sufficient to represent the uncertainty in the forecast model; the more complicated stochastic parametrisations show little improvement over simple additive AR(1) noise.

Traditional deterministic parametrisation schemes are a function of the grid-scale variables at the current time step only. If a stochastic parametrisation also needs only to represent the sub grid- and time-scale variability, the white noise schemes would be adequate. However, for both timescale separations we observe a significant improvement in the skill of stochastic parametrisations which include a temporal autocorrelation over those which use white noise. This challenges the standard idea that a parametrisation should

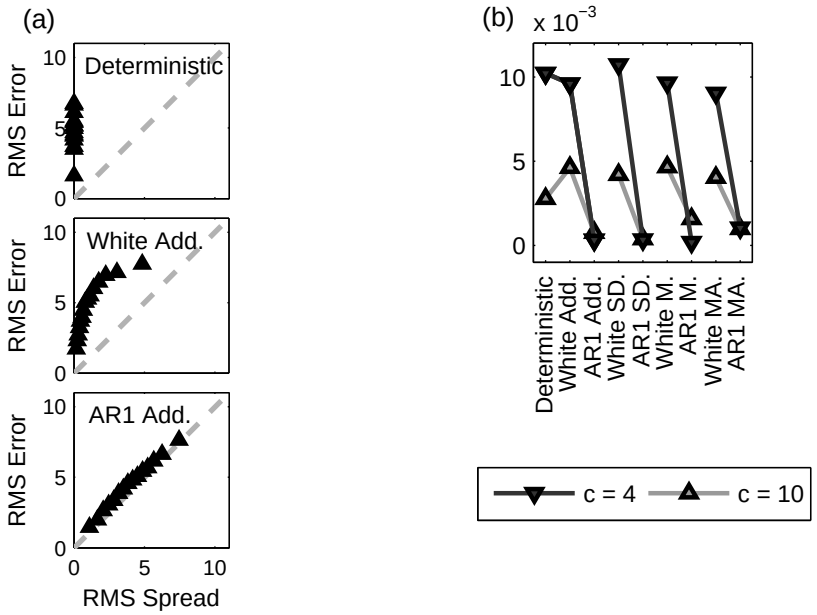

Figure 7: (a) RMS Error-Spread Diagnostic for the $c = 4$ case. The forecast-verification pairs are ordered according to their forecast variance, and divided into 15 equally populated bins. The mean square spread and mean square error are evaluated for each bin, before the square root is taken. Equation 5.1 shows the points will lie on the diagonal if the ensemble is statistically consistent. The diagnostic is considered for a deterministic forecast model, the parametrisation scheme using additive white noise, and finally including the measured temporal autocorrelation in the stochastic term. (b) REL, the reliability component of the Brier Score, is calculated for the different parametrisations and compared for the $c = 4$ and $c = 10$ cases. The smaller the REL, the more reliable the forecast. The parameters in each parametrisation have been estimated from the truth time series. In each case, "White" indicates that the measured standard deviations have been used, but the autocorrelation parameters set to zero.

only represent sub-gridscale and sub-timestep variability: including temporal autocorrelation accounts for the effects of the sub-gridscale at time scales greater than the model timestep. In the L96 system, the spatial correlations are low. However in an atmospheric situation, it is likely that spatial correlations will be significant, and a stochastic parametrisation must account for the effects of the sub-grid at scales larger than the spatial discretisation scale.

5. Representation of Model Uncertainty

The full forecast PDF represents our uncertainty about the future state of the system. Ideally, the ensemble forecast should be a random sample from that PDF. The consistency condition is that the verification also behaves like a sample from that PDF [1].

In order to meet the consistency condition, the ensemble must have the correct second moment. If it is under-dispersive, the verification will frequently fall as an outlier. Conversely, if the ensemble is over-dispersive the verification may fall too often towards the centre of the distribution. The reliability of an ensemble forecast can be tested through the spread-error relationship [14, 15]. The expected squared error of the ensemble mean can be related to the expected ensemble variance by assuming the ensemble members and the truth are independently identically distributed random variables with variance σ^2. Assuming the ensemble is unbiased, this gives the following requirement for a statistically consistent ensemble:

$$\frac{M}{M-1}\,\overline{\text{estimate ensemble variance}} = \frac{M}{M+1}\,\overline{\text{squared ensemble mean error}}, \qquad (5.1)$$

where the variance and mean error have been estimated by averaging over many forecast-verification pairs. For large ensemble size, $M \sim 40$, we can consider the correction factor to be close to 1.

Evaluating the average ensemble error as a function of forecast ensemble spread is a useful measure of how well the forecast model represents uncertainty, and is tested through diagnostic plots of binned root mean square (RMS) Spread against the average RMS Error in each bin. Figure 7(a) shows this diagnostic for a selection of the parametrisation schemes tested for the $c = 4$ case. For a well calibrated ensemble, the points should lie on the diagonal. A clear improvement over the deterministic scheme is seen as first white, then red additive noise is included in the parametrisation scheme. Visual forecast verification measures are limited when comparing many different models as they do not give an unambiguous ranking of the performance of these models. Therefore, we also consider the Reliability component of the Brier Score [19], REL, which is a scalar scoring rule testing how reliable an ensemble forecast is when predicting an event. We define the event to be "in the upper tercile of the climatological distribution". The smaller the REL, the closer the forecast probability is to the average observed frequency, the more reliable the forecast.

The results are summarised in Figure 7(b). The REL score indicates the different AR(1) noise terms all perform similarly. A significant improvement is observed when temporal autocorrelation is included in the parametrisation, particularly for the $c = 4$ case.

6. Climatological Skill and Seamless Prediction

The climatology of the L96 system is defined to be the PDF of the X variables, averaged over a long run (10,000 model time units $\sim$ 140 "atmospheric years"), and the forecast climatology is defined in an analogous way. The skill at predicting the climatology can then be quantified by measuring the difference between these two PDFs, which may be evaluated in several ways. The Kolmogorov-Smirnov (KS) statistic, D, has been used in this context in several other studies [13, 35], where

$$D = \max_{X_k} |P(X_k) - Q(X_k)| \,. \tag{6.1}$$

Here P is the forecast cumulative PDF, and Q is the verification cumulative PDF. A second measure, the Hellinger Distance, H, was also calculated for each forecast model:

$$H^2(p, q) = \frac{1}{2} \int \left(\sqrt{p(x)} - \sqrt{q(x)} \right)^2 \mathrm{d}x, \tag{6.2}$$

where $p(x)$ is the forecast PDF, and $q(x)$ is the verification PDF [28]. Similarly, the Kullback-Leibler (KL) divergence, is defined as

$$KL(p, q) = \int p(x) \ln \left(\frac{p(x)}{q(x)} \right) \mathrm{d}x, \tag{6.3}$$

motivated by information theory [12]. For all these measures, the smaller the measure, the better the match between forecast and verification climatologies.

The Hellinger Distance was found to be a much smoother measure of climatological skill than the KS statistic as it integrates over the whole PDF. Therefore, the KS statistic has not been considered further here. Figure 8 shows that the Hellinger distance and Kullback-Leibler divergence give very similar results, so for these reasons only the Hellinger Distance will be considered. Pollard [28] shows that the two measures are linked, so this similarity is not suprising.

The Hellinger Distance is evaluated for the different parametrisations for the two cases, $c = 4$ and $c = 10$. Figure 9 shows the results for $c = 4$ when the tuneable parameters were varied as for the weather skill scores. Qualitatively, the shape of the plots is similar to those for the weather skill scores. If a parametrisation performs well in forecasting mode, it performs well at reproducing the climate. The peak is shifted up and to the right compared to the RPSS, but is in a similar position to IGNSS and REL. The climatological skill for the different parametrisations is summarised in Figure 10. As for the weather forecasting skill, a significant improvement in the climatological skill is observed when temporal autocorrelation is included in the parametrisations.

The climatological skill, as measured by the Hellinger distance, can be compared to the weather skill scores using scatter diagrams (Figure 11). This is of interest, as the

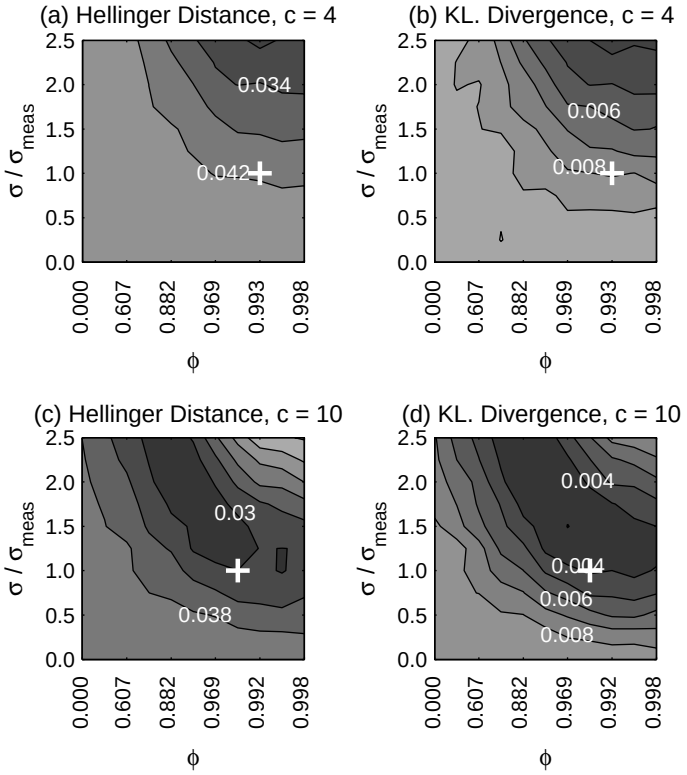

Figure 8: Comparing the Hellinger distance and the Kullback-Leibler divergence as measures of climatological skill. The measures are shown for the additive AR(1) noise parametrisation, as a function of the tuneable parameters, for both (a) and (b) the $c = 4$ case, and (c) and (d) the $c = 10$ case. The crosses indicate the measured parameters.

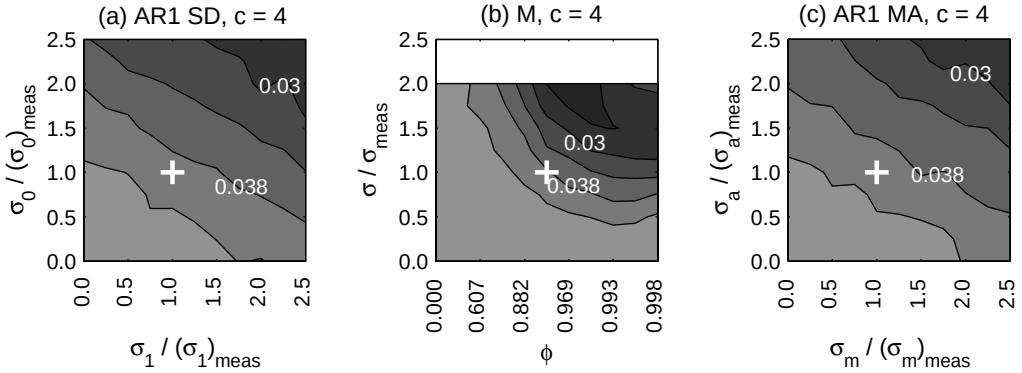

Figure 9: The climatological skill of different stochastic parametrisations for the $c = 4$ case, as a function of the tuneable parameters in each parametrisation. The crosses indicate the measured parameters in each case. The Hellinger distance was calculated between the truth and forecast probability density functions (PDFs) of the X variables. The smaller the Hellinger distance, the better the forecast PDF. Please see the online Supplementary Figure 2 for the c = 10 case.

seamless prediction paradigm suggests that climate models could be verified by evaluating the model in weather forecasting mode. Figures 11(a) and 11(d) show the relationship between RPSS and the Hellinger distance. For the $c = 10$ case, there appears to be strong negative correlation between the two. However, the peak in RPSS is offset slightly from the minimum in Hellinger Distance giving two branches in the scatter plot. The $c = 4$ case can be interpreted as being positioned at the joining point of the two branches, and shows how using the RPSS as a method to verify a model's climatological skill could be misleading.

Figures 11(b) and 11(e) compare Ignorance with the Hellinger Distance. The upper branch in (e) corresponds to the large magnitude high temporal autocorrelation para-

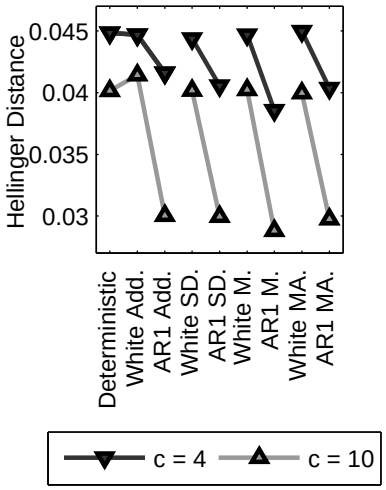

Figure 10: The Hellinger Distance for the different parametrisations is compared for the $c = 4$ and $c = 10$ cases. The smaller the Hellinger Distance, the better the climatological "skill". The parameters in each parametrisation have been estimated from the truth time series. In each case, "White" indicates that the measured standard deviations have been used, but the autocorrelation parameters set to zero.

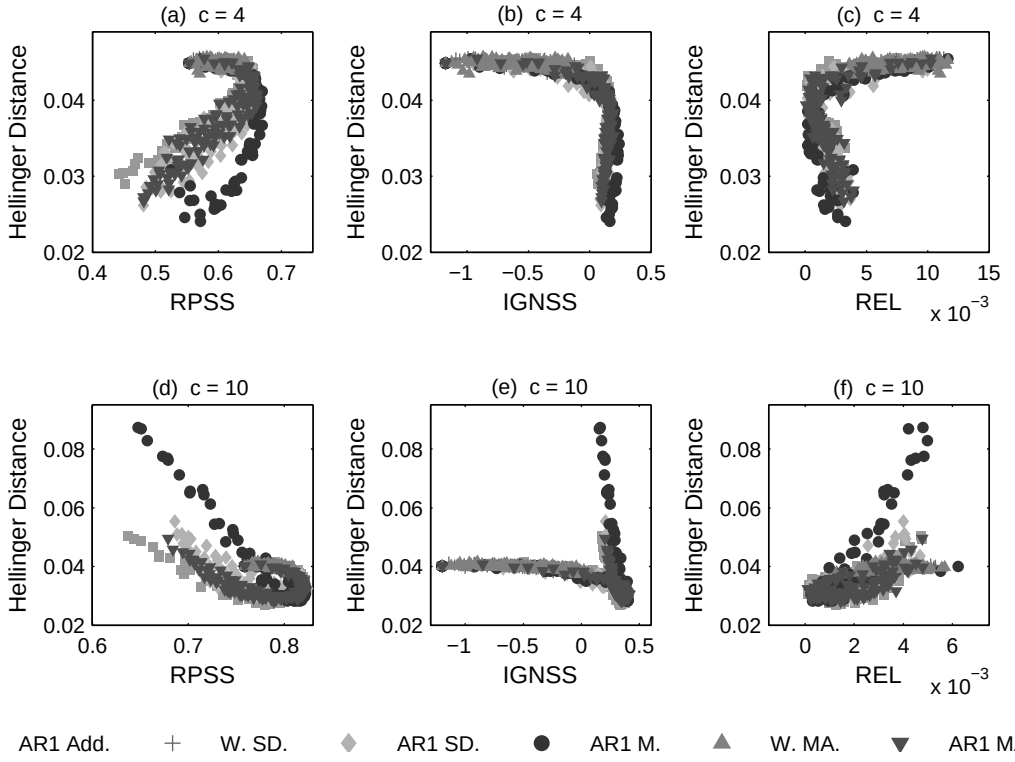

Figure 11: Scores (RPSS, IGNSS, REL) calculated when the forecast model is in weather mode are compared to the climatological skill of the forecast model, as measured by the Hellinger Distance. Figures (a)–(c) are for the $c = 4$ case, and Figures (d)–(f) are for the $c = 10$ case. The greater the forecast skill score (RPSS, IGNSS), the better the forecast. The lower the reliability score (REL), the more reliable the forecast. The lower the Hellinger Distance, the closer the match between the forecast climatology and the true climatology of the system. The symbols represent the different parametrisations - the legend below corresponds to all figures.

metrisations which Ignorance scores highly, but which have poor climatological skill. This makes Ignorance unsuitable as an evaluation method for use in seamless prediction.

Figures 11(c) and 11(f) show the results for the Reliability component of the Brier Score. There is a strong correlation between REL and the Hellinger Distance, indicating

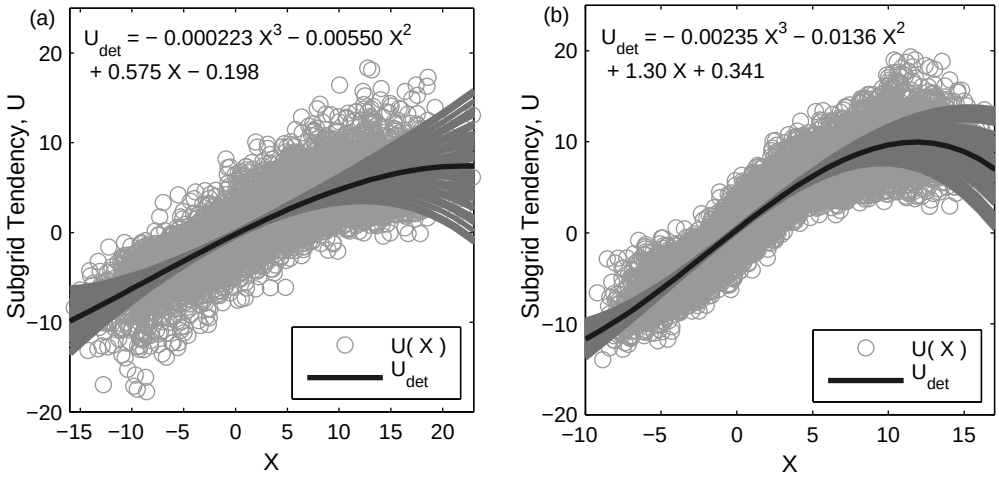

Figure 12: The ensemble of deterministic parametrisations used to represent model uncertainty in the perturbed parameter ensemble. The degree to which the parameters are perturbed has been estimated from the truth timeseries in each case ($S = 1$).

	$c = 4$	$c = 10$
b_0^{meas}	-0.198	0.341
$\sigma(b_0^{samp})$	0.170	0.146
b_1^{meas}	0.575	1.30
$\sigma(b_1^{samp})$	0.0464	0.0381
b_2^{meas}	-0.00550	-0.136
$\sigma(b_2^{samp})$	0.00489	0.00901
b_3^{meas}	-0.000223	-0.00235
$\sigma(b_3^{samp})$	0.000379	0.000650

Table 3: Measured parameters defining the cubic polynomial, $(b_0^{meas}, b_1^{meas}, b_2^{meas}, b_3^{meas})$, and the variability of these parameters, $\sigma(b_i^{samp})$, calculated by sampling from the truth time series, for the $c = 4$ and $c = 10$ cases.

this is a suitable score for use in seamless prediction. It is not surprising that the REL is well suited for this task, as it is particularly sensitive to the reliability of an ensemble, which is the characteristic of a forecast important for climate prediction [25]. The other weather skill scores studied put too much weight on resolution to be used for this purpose. Additionally, Figure 11(c) indicates that reliability in weather forecasting mode is a necessary but not a sufficient requirement of a good climatological forecast, as was suggested by Palmer et al. [25].

7. Perturbed Parameter Ensembles in the Lorenz '96 System

It is of interest to determine whether a perturbed parameter ensemble can also provide a reliable measure of model uncertainty in the Lorenz '96 system. The four measured parameters $(b_0^{meas}, b_1^{meas}, b_2^{meas}, b_3^{meas})$ defining the cubic polynomial will be perturbed to generate a 40 member ensemble. The skill of this representation of model uncertainty will be evaluated as for the stochastic parametrisations.

Following Stainforth et al. [32], each of the four parameters will be set to one of three values: low (L), medium (M) or high (H). The degree to which the parameters should be varied was estimated from the truth time series. The measured $U(X)$ was split into sections 3 MTU long, and a cubic polynomial fitted to each section. The measured variability in each of the parameters was then calculated as the standard deviation of the parameters fitted to each section $\sigma(b_i^{samp})$. The measured standard deviations are shown in Table 3.

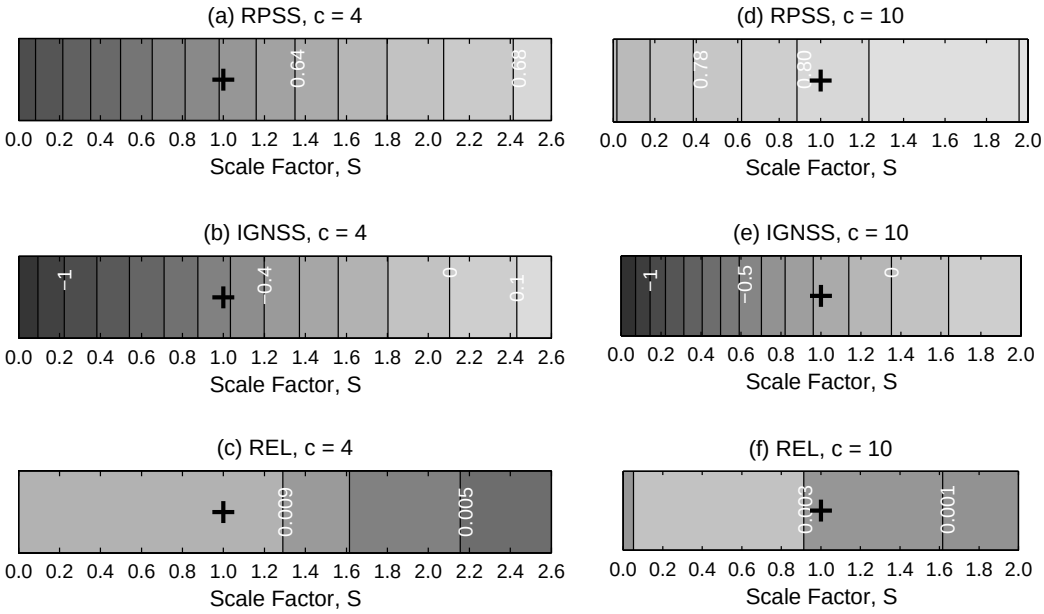

Figure 13: Weather forecasting skill scores for the perturbed parameter model as a function of the scale factor, S. The skill scores were calulated at a lead time of 0.6 model time units in each case.

The low, medium and high values of the parameters are given by:

$$\begin{aligned} L &= b_i^{meas} - S\sigma(b_i^{samp}) \\ M &= b_i^{meas} \\ H &= b_i^{meas} + S\sigma(b_i^{samp}), \end{aligned} \tag{7.1}$$

where the scale factor, S, can be varied to test the sensitivity of the scheme. There are $3^4 = 81$ possible permutations of the parameter settings, from which a subset of 40 permutations was selected to sample the uncertainty. This allows for a fair comparison to be made with the stochastic parametrisations, which also use a 40 member ensemble.

The same "truth" model will be used as for the stochastic parametrisations, and the forecast model will be constructed in an analogous way: only the X variables are assumed resolved, and the effects of the unresolved sub-gridscale Y variables are represented by an ensemble of deterministic parametrisations:

$$U_{pp}(X_k) = b_0^p + b_1^p X_k + b_2^p X_k^2 + b_3^p X_k^3, \tag{7.2}$$

where the values of the perturbed parameters, b^p, vary between ensemble members. The scale factor, S, in Equation (7.1) will be varied to investigate the effect on the skill of the forecast. The ensemble of deterministic parametrisations is shown in Figure 12 where the degree of parameter perturbation has been measured from the truth timeseries (i.e. $S = 1$). The truncated model will be integrated using an adaptive second order Runge-Kutta scheme.

(a) *Weather Prediction Skill*

The skill of the ensemble forecast is evaluated using the RPSS and IGNSS at a lead time of 0.6 model time units for both the $c = 4$ and $c = 10$ cases. The results are shown in Figure 13. The measured parameter perturbations are indicated with a black cross in each case.

Both RPSS and IGNSS indicate the measured perturbed parameter ensemble is significantly less skilful than the stochastic ensemble for both the $c = 4$ and $c = 10$ case. Figures 13(c) and (f) show the reliability component of the Brier Score calculated for the perturbed parameter ensembles. Comparing these figures with Figure 7(b), REL for the perturbed parameter schemes is greater, indicating that the perturbed parameter ensemble forecasts are less reliable than the stochastic parametrisation forecasts, so the ensemble is

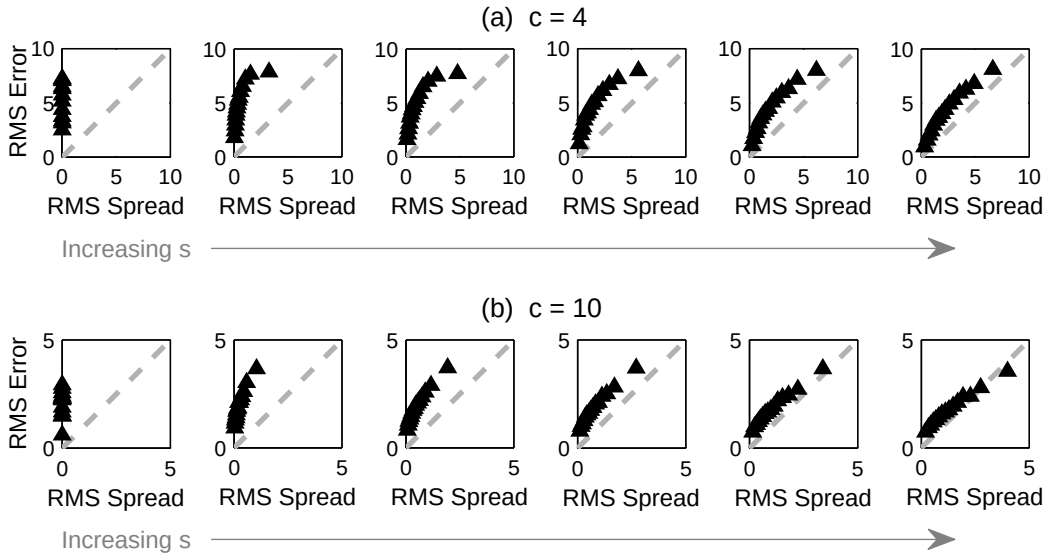

Figure 14: RMS Spread vs. RMS Error plots for the perturbed parameter ensemble as a function of the scale factor, S. The separate figures correspond to $S = [0.0, 0.4, 0.8, 1.2, 1.6, 2.0]$.

a poorer representation of model uncertainty. The significance of the difference between skill scores for the the measured stochastic parametrisation schemes and the perturbed parameter schemes with the measured perturbations ($S = 1$) are shown in the significance tables in the online material.

The reliability of the forecast ensemble is also considered using the RMS spread–error diagnostic as a function of the scale factor in Figure 14. For small-scale factors, the ensemble is systematically underdispersive for both the $c = 4$ and $c = 10$ cases. For larger scale factors for the $c = 10$ case, the ensemble is systematically overdispersive for large errors, and underdispersive for small errors. Comparing Figure 14 with Figure 7(a) shows that none of the perturbed parameter ensembles are as reliable as the AR1 additive stochastic parametrisation, reflected in the poorer REL score for the perturbed parameter case.

(*b*) *Climatological Skill*

The climatology of the perturbed parameter ensemble must include contributions from each of the 40 ensemble members. Therefore, the climatology is defined as the PDF of the X variables, averaged over the same total number of model time units as the stochastic parametrisations (10,000); each of the 40 ensemble members is integrated for 250 model time units. The Hellinger Distance between the truth and forecast climatologies can then be calculated as a function of the scale factor (Figure 15). The climatology of the measured perturbed parameter ensemble is worse than the measured parameter stochastic parametrisations. This is as predicted by the "seamless prediction" paradigm; the perturbed parameter ensembles are less reliable than the stochastic parametrisations, and so predict a less accurate climatology.

8. Conclusions

Several different stochastic parametrisation schemes were investigated using the Lorenz '96 system. All showed an improvement in weather and climate forecasting skill over deterministic parametrisations. This result is robust to error in measurement of the parameters — scanning over parameter space indicated a wide range of parameter settings gave good skill scores.

Stochastic parametrisations have been shown to represent the uncertainty in a forecast due to model deficiencies accurately. This increase in forecast reliability is reflected by the increase in the weather prediction skill and improved climatology of the forecast model.

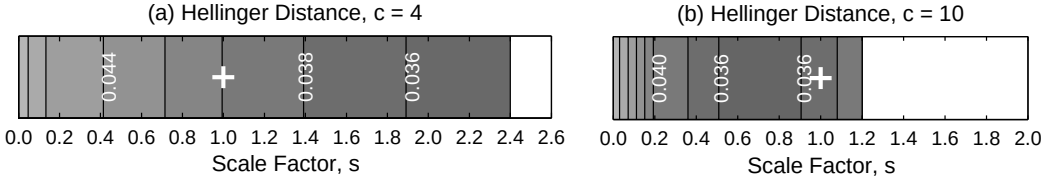

Figure 15: The Hellinger Distance between the truth and forecast climatologies of the perturbed parameter model as a function of the scale factor, s. The smaller the Hellinger Distance, the better the predicted climatology. The $S > 2.4$ and $S > 1.2$ forecast models for the $c = 4$ and $c = 10$ cases respectively are numerically unstable over long integrations, so a climatology could not be calculated.

A significant improvement in the skill of the forecast models was observed when the stochastic parametrisations included temporal autocorrelation in the noise term. This challenges the notion that a parametrisation scheme should only represent sub-gridscale (both temporal and spatial) variability. The coupling of scales in a complex system means a successful parametrisation must represent the effects of the sub-grid acting on spatial and time scales greater than the truncation level.

The correlation between the performance of the forecast model in weather prediction mode, and its ability to reproduce the climatology of the Lorenz '96 system provides support for the "Seamless Prediction" paradigm. This provides a method of verifying climate predictions: the climate model can be run in weather forecasting mode, and its short term predictive skill analysed.

Stochastic representations of model uncertainty are shown to outperform perturbed parameter ensembles in the L96 system. They have higher short term forecasting skill and are more reliable than perturbed parameter ensembles. They also predict the climatology of the L96 system better.

The Lorenz '96 system is an excellent tool for testing developments in stochastic parametrisations. These ideas can now be applied to numerical weather prediction models and tested on the atmosphere.

9. Acknowledgements

Thank you very much to the following for interesting and useful discussions and suggestions: Jochen Bröcker, Andrew Dawson, Chris Ferro, Frank Kwasniok, Martin Leutbecher, Cecile Penland, Dan Rowlands and Paul Williams. This work was supported by a NERC studentship.

References

[1] J. L. Anderson. The impact of dynamical constraints on the selection of initial conditions for ensemble predictions: Low-order perfect model results. *Mon. Weather Rev.*, 125(11):2969–2983, 1997.

[2] J. Berner, G. J. Shutts, M. Leutbecher, and T. N. Palmer. A spectral stochastic kinetic energy backscatter scheme and its impact on flow dependent predictability in the ECMWF ensemble prediction system. *J. Atmos. Sci.*, 66(3):603–626, 2009.

[3] J. Berner, T. Jung, and T. N. Palmer. Systematic model error: The impact of increased horizontal resolution versus improved stochastic and deterministic parameterizations. *J. Climate*, 2012.

[4] G. W. Brier. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.*, 78(1):1–3, 1950.

[5] J. Bröcker. Reliability, sufficiency, and the decomposition of proper scores. *Q. J. Roy. Meteor. Soc.*, 135(643):1512–1519, 2009.

[6] R. Buizza and T. N. Palmer. The singular-vector structure of the atmospheric global circulation. *J. Atmos. Sci.*, 52(9):1434–1456, 1995.

[7] D. Crommelin and E. Vanden-Eijnden. Subgrid-scale parametrisation with conditional markov chains. *J. Atmos. Sci.*, 65(8):2661–2675, 2008.

[8] Y. Frenkel, A. J. Majda, and B. Khouider. Using the stochastic multicloud model to improve tropical convective parametrisation: A paradigm example. *J. Atmos. Sci.*, 69(3):1080–1105, 2012.

[9] J. A. Hansen and C. Penland. Efficient approximation techniques for integrating stochastic differential equations. *Mon. Weather Rev.*, 134(10):3006–3014, 2006.

[10] J. A. Hansen and C. Penland. On stochastic parameter estimation using data assimilation. *Physica D*, 230(1–2):88–98, 2007.

[11] P. L Houtekamer, L. Lefaivre, and J. Derome. A system simulation approach to ensemble prediction. *Mon. Weather Rev.*, 124(6):1225–1242, 1996.

[12] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Stat.*, 22 (1):79–86, 1951.

[13] F. Kwasniok. Data-based stochastic subgrid-scale parametrisation: an approach using cluster-weighted modelling. *Phil. Trans. R. Soc. A*, 370(1962):1061–1086, 2012.

[14] M. Leutbecher. Diagnosis of ensemble forecasting systems. In *Seminar on Diagnosis of Forecasting and Data Assimilation Systems, 7 - 10 September 2009*, pages 235–266, Shinfield Park, Reading, 2010. ECMWF.

[15] M. Leutbecher and T. N. Palmer. Ensemble forecasting. *J. Comput. Phys.*, 227(7): 3515–3539, 2008.

[16] E. N. Lorenz. Predictability – a problem partly solved. In *Proceedings, Seminar on Predictability ECMWF*, volume 1, pages 1–18, 1996.

[17] D. Masson and R. Knutti. Climate model genealogy. *Geophys. Res. Let.*, 38, 2011.

[18] G. A. Meehl, C. Covey, T. Delworth, M. Latif, B. McAvaney, J. F. B. Mitchell, R. J. Stouffer, and K. E. Taylor. The WCRP CMIP3 multimodel dataset. *B. Am. Meteorol. Soc.*, 88(9):1383–1394, 2007.

[19] A. H. Murphy. A new vector partition of the probability score. *J. Appl. Meteorol.*, 12(4):595–600, 1973.

[20] G. D. Nastrom and K. S. Gage. A climatology of atmospheric wavenumber spectra of wind and temperature observed by commercial aircraft. *J. Atmos. Sci.*, 42(9): 950–960, 1985.

[21] T. N Palmer. A nonlinear dynamical perspective on model error: A proposal for non-local stochastic-dynamic parametrisation in weather and climate prediction models. *Q. J. Roy. Meteor. Soc.*, 127(572):279–304, 2001.

[22] T. N Palmer. The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Q. J. Roy. Meteor. Soc.*, 128(581):747–774, 2002.

[23] T. N. Palmer. Towards the probabilistic earth-system simulator: A vision for the future of climate and weather prediction, 2012.

[24] T. N. Palmer and P. Williams. *Stochastic Physics and Climate Modelling*. Cambridge University Press, Cambridge, U.K., 2010.

[25] T. N. Palmer, F. J. Doblas-Reyes, A. Weisheimer, and M. J. Rodwell. Towards seamless prediction: Calibration of climate change projections using seasonal forecasts. *B. Am. Meteorol. Soc.*, 89(4):459–470, 2008.

[26] T. N. Palmer, R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G. J. Shutts, M. Steinheimer, and A. Weisheimer. Stochastic parametrization and model uncertainty. Technical Report 598, European Centre for Medium-Range Weather Forecasts, 2009.

[27] C. Pennell and T. Reichler. On the effective number of climate models. *J. Climate*, 24(9):2358–2367, 2011.

[28] D. Pollard. *A User's Guide to Measure Theoretic Probability*. Cambridge University Press, Cambridge, U.K. and New York, NY, U.S.A., 2002.

[29] J. Rougier, D. M. H. Sexton, J. M. Murphy, and D. Stainforth. Analyzing the climate sensitivity of the HadSM3 climate model using ensembles from different but related experiments. *J. Climate*, 22:3540–3557, 2009.

[30] M. S. Roulston and L. A. Smith. Evaluating probabilistic forecasts using information theory. *Mon. Weather Rev.*, 130(6):1653–1660, 2002.

[31] G. J. Shutts and T. N. Palmer. Convective forcing fluctuations in a cloud-resolving model: Relevance to the stochastic parameterization problem. *J. Climate*, 20(2):187–202, 2007.

[32] D. A Stainforth, T. Aina, C. Christensen, M. Collins, N. Faull, D. J. Frame, J. A. Kettleborough, S. Knight, A. Martin, J. M. Murphy, C. Piani, D. Sexton, L. A. Smith, R. A Spicer, A. J. Thorpe, and M. R Allen. Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature*, 433(7024):403–406, 2005.

[33] D. J. Stensrud, J.-W. Bao, and T. T. Warner. Using initial condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Mon. Weather Rev.*, 128(7):2077–2107, 2000.

[34] A. P. Weigel, M. A. Liniger, and C. Appenzeller. Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Q. J. Roy. Meteor. Soc.*, 134(630):241–260, 2008.

[35] D. S. Wilks. Effects of stochastic parametrizations in the Lorenz '96 system. *Q. J. Roy. Meteor. Soc.*, 131(606):389–407, 2005.

[36] D. S. Wilks. *Statistical Methods in the Atmospheric Sciences*, volume 91 of *International Geophysics Series*. Elsevier, second edition, 2006.