Stochastic and Perturbed Parameter Representations of Model Uncertainty in Convection Parameterization*

H. M. CHRISTENSEN

Atmospheric, Oceanic and Planetary Physics, University of Oxford, Oxford, United Kingdom

I. M. MOROZ

Oxford Centre for Industrial and Applied Mathematics, University of Oxford, Oxford, United Kingdom

T. N. PALMER

Atmospheric, Oceanic and Planetary Physics, University of Oxford, Oxford, United Kingdom

(Manuscript received 19 August 2014, in final form 9 January 2015)

ABSTRACT

It is now acknowledged that representing model uncertainty in atmospheric simulators is essential for the production of reliable probabilistic forecasts, and a number of different techniques have been proposed for this purpose. This paper presents new perturbed parameter schemes for use in the European Centre for Medium-Range Weather Forecasts (ECMWF) convection scheme. Two types of scheme are developed and implemented. Both schemes represent the joint uncertainty in four of the parameters in the convection parameterization scheme, which was estimated using the Ensemble Prediction and Parameter Estimation System (EPPES). The first scheme developed is a fixed perturbed parameter scheme, where the values of uncertain parameters are varied between ensemble members, but held constant over the duration of the forecast. The second is a stochastically varying perturbed parameter scheme. The performance of these schemes was compared to the ECMWF operational stochastic scheme, stochastically perturbed parameterization tendencies (SPPT), and to a model that does not represent uncertainty in convection. The skill of probabilistic forecasts made using the different models was evaluated. While the perturbed parameter schemes improve on the stochastic parameterization in some regards, the SPPT scheme outperforms the perturbed parameter approaches when considering forecast variables that are particularly sensitive to convection. Overall, SPPT schemes are the most skillful representations of model uncertainty owing to convection parameterization.

1. Introduction

Convection is parameterized in both weather and climate models. The physics included in a parameterization is a compromise between accuracy and computing constraints: parameterization is a large source of

uncertainty in forecasts. Convection is generally acknowledged to be the parameterization to which weather and climate models are most sensitive (Knight et al. 2007), so it is important to represent the uncertainty originating in the parameterization of convection. One proposed representation of uncertainty is the class of perturbed parameter schemes. When developing a parameterization scheme, parameters are introduced to represent complex physical processes: because of the crudeness of typical convection schemes, such physical parameters are poorly constrained by observations. However, the evolution of convective clouds and the impact on weather and climate are very sensitive to the value of these parameters (Sanderson et al. 2008). In a perturbed parameter ensemble, the values of a selected set of parameters are sampled from a distribution

Denotes Open Access content.

*Supplemental information related to this paper is available at the Journals Online website: <http://dx.doi.org/10.1175/JAS-D-14-0250.s1>.

Corresponding author address: H. M. Christensen, Atmospheric, Oceanic and Planetary Physics, Clarendon Laboratory, Parks Road, Oxford OX1 3PU, United Kingdom.
E-mail: h.m.christensen@atm.ox.ac.uk

DOI: 10.1175/JAS-D-14-0250.1

representing the uncertainty in their values, with each ensemble member assigned a different set of parameters. These parameters are fixed globally and for the duration of the integration, sampling the “knowledge uncertainty” in the optimal value of the physical parameters. The parameter distribution is usually determined through “expert elicitation,” whereby scientists with the required knowledge and experience of using the parameterization suggest upper and lower bounds for the parameter (Stainforth et al. 2005). This method does not include information about the relationships between parameters in the distribution (and produces results that are dependent on the expert in question), though unrealistic simulations can be removed from the ensemble later (Stainforth et al. 2005).

A related problem is that the poorly constrained nature of the parameters in an atmospheric model can have adverse effects on the accuracy of high-resolution integrations. Tuning the many hundreds of parameters in models is a difficult, costly process, usually performed by hand. An attractive alternative is the use of a Bayesian parameter estimation approach. This estimates the probability distribution of parameters and provides a framework for using new data from forecasts and observations to update prior knowledge about the parameter distribution (Beck and Arnold 1977). One specific technique is the Ensemble Prediction and Parameter Estimation System (EPPES) (Järvinen et al. 2012; Laine et al. 2012), which runs online in conjunction with an operational ensemble forecasting system. As well as producing an improved estimate of the value of the parameters, this procedure also produces an estimate of the uncertainty in the values of these parameters in the form of a joint probability distribution. EPPES is of central interest to this study, as this distribution can be used to develop a perturbed parameter scheme without the need for expert elicitation.

An alternative technique for representing uncertainty in convection parameterization is the development of stochastic convection parameterization schemes. These use random numbers to represent the difference between a deterministic parameterization scheme and the true atmosphere, accounting for the unresolved subgrid-scale variability associated with convective clouds. Many such schemes have been proposed in recent years. Several schemes propose a stochastic perturbation to the input to a deterministic scheme, such as by using a stochastic representation of convective available potential energy (CAPE) or convective inhibition (CIN) (Lin and Neelin 2000; Majda and Khouider 2002; Khouider et al. 2003), using a stochastic parameterization of mass flux at cloud base (Craig and Cohen 2006; Plant and Craig 2008) or by using stochastic cellular automata to determine the updraft fraction at cloud base

(Bengtsson et al. 2013). By perturbing only the input to the deterministic scheme, these schemes assume the vertical structure produced by the deterministic parameterization scheme is satisfactory. An alternative is to perturb the output of a deterministic scheme, such as by including an additive noise term for the temperature at each vertical level (Lin and Neelin 2003) or by multiplying the parameterized convective tendencies by a random number (Palmer et al. 2009).

A criticism of some stochastic parameterization schemes is the ad hoc way in which stochasticity is introduced into the scheme. On the other hand, perturbed parameter ensembles provide an attractive way to include stochasticity into a parameterization scheme in a physically motivated way. Since physical parameters represent various complex physical processes, they may not have a single value valid at all times (Khouider and Majda 2006). Stochastically perturbing uncertain parameters can represent this parameter uncertainty and provides a physically motivated method for developing a stochastic parameterization scheme. An example of such a scheme is the random parameters (RP) scheme used in the Met Office Global and Regional Ensemble Prediction System (MOGREPS) (Bowler et al. 2008). A subset of the model parameters are varied globally following a first-order autoregressive process. The parameters in this scheme are bounded, and the maximum and minimum permitted values of the parameters are set by experts in the respective parameterization schemes.

In this paper we compare the performance of stochastic and perturbed parameter representations of uncertainty in the European Centre for Medium-Range Weather Forecasts (ECMWF) convection scheme. Both fixed and stochastically varying perturbed parameter schemes have been considered. The uncertainty in the parameters of interest was estimated using the EPPES algorithm (Ollinaho et al. 2013a). The stochastic scheme considered here is the stochastically perturbed parameterization tendencies (SPPT) scheme, used operationally at ECMWF. In sections 2 and 3 we describe the ECMWF model and SPPT, respectively. The new perturbed parameter representations of uncertainty are discussed in section 4, and in section 5 we outline the experimental procedure. The results are presented in section 6 and discussed in section 7, before some conclusions are drawn in section 8.

2. The Integrated Forecasting System

The experiments in this study are carried out using the Integrated Forecasting System (IFS), the operational global weather forecasting model of ECMWF. The IFS consists of a spectral atmospheric model and uses

persistent SSTs instead of a dynamical ocean out to day 10. The Ensemble Prediction System (EPS) samples initial condition uncertainty using perturbations derived from an ensemble of data assimilations (EDA) (Isaksen et al. 2010), which are weakly blended with perturbations from the leading singular vectors (Buizza et al. 2008). Stochastic parameterizations are used to represent uncertainty in the EPS due to model deficiencies. Two stochastic parameterization schemes are used. SPPT (Palmer et al. 2009) uses multiplicative noise to perturb the total parameterized tendencies about the average value that a deterministic scheme represents, thus addressing model uncertainty due to the physical parameterization schemes. The second scheme, stochastic kinetic energy backscatter (SKEB) (Berner et al. 2009), represents a physical process absent from the IFS deterministic parameterization schemes. It uses random streamfunction perturbations to represent upscale kinetic energy transfer, counteracting the kinetic energy loss from excessive dissipation in the numerical integration schemes.

The convection parameterization scheme in the IFS, described in Bechtold et al. (2014), is based on the mass-flux scheme of Tiedtke (1989). The scheme describes three types of convective cloud: deep, shallow, and midlevel. The convective clouds in a column are represented by a pair of entraining and detraining plumes of a given convective type, which describe updraft and downdraft processes, respectively. The choice of convective type determines certain properties of the cloud (such as the entrainment formulation). The mass flux at cloud base for deep convection is estimated by assuming that deep convection acts to reduce CAPE over some specified (resolution dependent) time scale. Midlevel convection occurs at warm fronts. The mass flux at cloud base is set to be the large-scale vertical mass flux at that level. For shallow convection, the mass flux at cloud base is derived by assuming that the moist static energy in the subcloud layer is in equilibrium.

3. Uncertainty in convection: Generalized SPPT

The SPPT scheme addresses model uncertainty due to the physics parameterization schemes by perturbing the physics tendencies using multiplicative noise; the word “tendency” refers to the change in a variable over a time step. SPPT perturbs the sum of the parameterized tendencies:

$$\frac{\partial X}{\partial t} = D + K + (1 + e) \sum_{i=1}^5 P_i, \quad (1)$$

where $\partial X/\partial t$ is the total tendency in X , D is the tendency from the dynamics, K is horizontal diffusion, P_i is the

tendency from the i th physics scheme, and e is the zero mean random perturbation. The SPPT scheme acts on tendencies from the five main parameterization schemes in the IFS: radiation (RDTT), turbulence and gravity wave drag (TGWD), nonorographic gravity wave drag (NOGW), convection (CONV), and large-scale water processes (or clouds—LSWP). Therefore, in particular, SPPT constitutes the operational representation of uncertainty in the IFS convection scheme.

SPPT perturbs the tendency for four variables: T , U , V , and q . Each variable tendency is perturbed using the same random number field. The perturbation field, e , is generated using a spectral pattern generator. The pattern at each time step is the sum of three independent random fields with horizontal correlation scales of 500, 1000, and 2000 km. These fields are evolved in time using an AR(1) process on time scales of 6 h, 3 days, and 30 days, respectively, and the fields have standard deviations of 0.52, 0.18, and 0.06, respectively. It is expected that the smallest scale (500 km and 6 h) will dominate at a 10-day lead time—the larger-scale perturbations are important for monthly and seasonal forecasts.

SPPT does not distinguish between the different parameterization schemes. However, the parameterization schemes likely have very different error characteristics, so this assumption may not be valid. This study is concerned with testing alternative, perturbed parameter representations of model uncertainty in convection. A generalized version of SPPT was developed in which multiplicative noise is applied separately to the tendencies from each physics parameterization scheme,

$$\frac{\partial X}{\partial t} = D + K + \sum_{i=1}^5 (1 + e_i) P_i, \quad (2)$$

such that the stochastic field, e_i , for the convection tendency can be set to zero and replaced with a perturbed parameter representation of uncertainty.

To detect an improvement in the representation of uncertainty in the convection scheme, the uncertainty in the other four parameterization schemes must be well represented. In this study, SPPT is used to represent uncertainty in the other four schemes, applying the same stochastic perturbation to each scheme. The SKEB scheme represents a process that is otherwise missing from the model, so it will be used in these experiments. Previous studies indicate that a multiplicative stochastic parameterization scheme is a skillful representation of model uncertainty in simple systems and is significantly more skillful than other simple stochastic parameterizations (e.g., additive noise schemes) (Arnold et al. 2013; Christensen et al. 2015b). Using SPPT and SKEB to represent convective uncertainty is therefore a good

benchmark when testing the perturbed parameter schemes outlined below.

4. Perturbed parameter approach to uncertainty in convection

a. Perturbed parameters and the EPPES

The EPPES has been recently proposed as a technique for constraining parameters in numerical weather prediction (NWP) models (Järvinen et al. 2012; Laine et al. 2012; Ollinaho et al. 2013b,a). Instead of manually undergoing the lengthy tuning process, the EPPES runs online to optimize the value of a selected set of parameters, with minimal cost to operational centers. The mean and covariance of a set of parameters are estimated using a Bayesian updating procedure. A summary of the EPPES algorithm can be found in the online supplemental material.

The EPPES was implemented using IFS cycle CY37R3, at a resolution of T159 by Ollinaho et al. (2013a) for forecasts initialized in May–August 2011. The SPPT and SKEB stochastic physics schemes were used to represent model uncertainty. The cost function used was the root-mean-square error (RMSE) in 500-hPa geopotential height (Z500), evaluated globally at a lead time of 3 and 10 days. The two lead times were scaled to contribute the same order of magnitude to the cost function. Z500 is chosen for the cost function because this variable is commonly used in NWP model

development for model tuning purposes. While rich in information in the extratropics, in the tropics Z500 shows little variability. However, it is useful for constraining uncertainty in tropical convection for two reasons. First, it is able to detect biases introduced by poor values of the physical parameters, and second, the 3–10-day delay in verification allows errors to propagate from the tropics to the extra tropics, where a larger signal is observed. Four of the parameters in the convection scheme were considered: ENTRORG, ENTSALP, DETRPEN, and RPRCON.

- ENTRORG represents organized entrainment for positively buoyant deep convection, with a default value of $1.75 \times 10^{-3} \text{ m}^{-1}$.
- ENTSALP $\times$ ENTRORG represents shallow entrainment, and the default value for ENTSALP is 2.
- DETRPEN is the average detrainment rate for penetrative convection and has a default value of $0.75 \times 10^{-4} \text{ m}^{-1}$.
- RPRCON is the coefficient for determining the conversion rate from cloud water to rain and has a default value of 1.4×10^{-3} .

The posterior distribution of these four parameters was determined using EPPES in terms of a mean vector, $\mathbf{M}(i)$, and covariance matrix with elements Σ_{ij} , where $i = 1$ represents ENTRORG, $i = 2$ represents ENTSALP, $i = 3$ represents DETRPEN, and $i = 4$ represents RPRCON:

$$\mathbf{M} = \begin{pmatrix} 0.1828 \times 10^{-2} \\ 0.2146 \times 10^1 \\ 0.7783 \times 10^{-4} \\ 0.1513 \times 10^{-2} \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} 9.7 \times 10^{-8} & -2.1 \times 10^{-5} & -4.2 \times 10^{-10} & -1.8 \times 10^{-8} \\ -2.1 \times 10^{-5} & 9.3 \times 10^{-2} & 1.3 \times 10^{-6} & -3.6 \times 10^{-5} \\ -4.2 \times 10^{-10} & 1.3 \times 10^{-6} & 5.2 \times 10^{-11} & -1.1 \times 10^{-9} \\ -1.8 \times 10^{-8} & -3.6 \times 10^{-5} & -1.1 \times 10^{-9} & 4.9 \times 10^{-8} \end{pmatrix}.$$

The resultant optimized value of each parameter was tested in the high-resolution deterministic forecast model, and many of the verification metrics were found to improve when compared to using the default values (Ollinaho et al. 2013a). This is very impressive, since the operational version of the IFS is already a highly tuned system. By comparison with the default values, the $\mathbf{M}$ vector indicates the magnitude by which the parameters should be changed to optimize the forecast.

The posterior covariance matrix estimated using the EPPES approach is also useful for model tuning as it can reveal parameter correlations and can thus identify redundant parameters. However, the covariance matrix also gives an indication of the uncertainty in the parameters and is used in this study to develop a perturbed parameter representation of uncertainty for the

ECMWF convection scheme. The off-diagonal terms in the Σ matrix indicate the covariance between the parameters, which is not negligible. This highlights one of the problems with using “expert elicitation” to define parameter distributions—such distributions contain no information about parameter interdependencies.

b. Fixed perturbed parameters

The usual method used in perturbed parameter experiments is a fixed perturbed parameter ensemble (Murphy et al. 2004; Sanderson 2011; Stainforth et al. 2005; Yokohata et al. 2010). Each ensemble member is assigned a set of parameter values that are held constant spatially and over the duration of the integration. Such ensembles are traditionally used for climate-length integrations. It will be interesting to see how well such an

ensemble performs at representing uncertainty in weather forecasts in the IFS.

The multivariate normal distribution specified by the EPPES was sampled to give n_{ens} sets of the four parameters, where the number of ensemble members $n_{\text{ens}} = 50$. The fixed perturbed parameter ensemble (“TSCF”—see Table 1 for explanation) considered here uses the same 50 sets of four parameters for all starting dates. Sampling of the parameters is performed offline using Latin hypercube sampling, ensuring the joint distribution is fully explored. The covariance of the resultant sample is checked against the EPPES covariance matrix; 10 000 iterations found a sample whose covariance matrix differed by less than 5% from the true matrix. The sampled parameter values are given in Arnold (2013).

c. Stochastically varying perturbed parameters

Khouider and Majda (2006) recognize that a problem with many deterministic parameterization schemes is the presence of parameters that are “nonphysically kept fixed/constant and spatially homogeneous.” An alternative to the fixed perturbed parameter ensemble described above is a stochastically varying perturbed parameter ensemble (“TSCV”), where the parameter values are varied spatially and temporally following the EPPES distribution. However, the EPPES does not estimate the correct spatial and temporal scales on which to vary the parameters. From the definition of the cost function, the set of parameters must perform well over 10 days to produce a skillful forecast, indicating that 10 days could be a suitable temporal scale. The cost function evaluates the skill of the forecast using Z500. The likelihood function will therefore focus on the mid-latitudes, where Z500 has high variability. A suitable spatial-scale could therefore be ~ 1000 km.

For the TSCV ensemble, the SPPT spectral pattern generator is used to generate four independent spatially and temporally correlated fields of random numbers. A three-scale composite pattern is used with the same spatial and temporal correlations as used in SPPT. These settings vary the parameters faster and on smaller spatial scales than the scales to which EPPES is sensitive, as estimated above. However, it will still be useful as a first test and, when combined with the fixed perturbed parameter ensemble (varying on an infinite spatial and temporal scale), it can provide bounds on the skill of such a representation of model uncertainty. The standard deviations of the independent patterns are 0.939 (smallest scale), 0.325, and 0.108 (largest scale) to give a total standard deviation of 1, before the fields are transformed to introduce the correct covariance structure Σ . The four covarying fields are used to define the

TABLE 1. Proposed experiments for investigating the representation of uncertainty in the ECMWF convection parameterization scheme. The experiment identifier indicates which method is used to represent uncertainty in the convection parameterization scheme (Cx) and that all experiments use SPPT to represent uncertainty in the other tendencies (TS).

Acronym	Description	
	Other tendencies	Convection tendency
TSCZ	Stochastic	Zero
TSCS	Stochastic	Stochastic
TSCF	Stochastic	Fixed perturbed parameters
TSCV	Stochastic	Varying perturbed parameters

values of the four convection parameters as a function of position and time.

5. Experimental procedure

Parameter estimation was carried out with the EPPES system using IFS model version CY37R3 for 45 dates between 12 May and 8 August 2011 (Ollinaho et al. 2013a). The same IFS model version is used to test the perturbed parameter schemes for consistency. The schemes are tested using 10-day hindcasts initialized every 5 days between 14 April and 6 September 2012: start dates are taken from the same time of year as for the EPPES, since the covariance pdfs may be seasonally dependent, but the test is “out of sample.” The parameterization schemes are tested at T159 (1.125°) using a 50-member ensemble forecast. The high-resolution ECMWF four-dimensional variational data assimilation (4DVar) operational analysis is used for verification.

Four experiments are proposed to investigate the representation of model uncertainty in the convection scheme in the IFS (Table 1). In each, the uncertainty in the other four parameterization tendencies (radiation, turbulence and gravity wave drag, nonorographic gravity wave drag, large-scale water processes) is represented by SPPT (“TS”). In the first experiment, there is no representation of uncertainty in the convection tendency (“CZ”). In the second, SPPT is used to represent uncertainty in the convection tendency (“CS”—this experiment is equivalent to the operational SPPT parameterization scheme). In the final two, uncertainty in the convection tendency is represented by a fixed perturbed parameter ensemble (“CF”) and by a stochastically varying perturbed parameter ensemble (“CV”). While the EPPES parameter information is used in the fixed and stochastically varying perturbed parameter models, the standard operational parameter values are used in the TSCZ and TSCS experiments.

To compare the different representations of convection model uncertainty, the SPPT scheme must correctly

account for uncertainty in the other four tendencies. Therefore, verification will be performed in a two stage process. First, the calibration of the ensemble will be checked in a region where forecasts have little uncertainty as a result of convection, (i.e., where there is little convective activity). The four experiments in Table 1 should perform very similarly in this region as they have the same representation of uncertainty in the other four physics tendencies. Second, a region where convection is the dominant process will be selected to test the different uncertainty schemes. Given that model uncertainty has been accounted for in the other four parameterizations using SPPT, and that a region has been selected where the model uncertainty is dominated by deep convection, a scheme that accurately represents uncertainty in deep convection will give a reliable forecast in this region, and any detected improvements in forecast skill can be attributed to an improvement in representation of uncertainty in the convection scheme.

a. Definition of verification regions

The regions of interest are defined using the Year of Tropical Convection (YOTC) dataset archived at ECMWF. YOTC was a joint World Climate Research Program and World Weather Research Programme/The Observing System Research and Predictability Experiment (WWRP/THORPEX) project designed to focus research efforts on the problem of organized tropical convection. The ECMWF YOTC dataset consists of high-resolution analysis and forecast data for May 2008–April 2010. In particular, the IFS parameterization tendencies were archived at every time step out to a lead time of 10 days.

The 24-h cumulative temperature tendencies at 850 hPa for each parameterization scheme are used. Forecasts initialized from 30 dates between 14 April and 6 September 2009 are selected, with subsequent start dates separated by 5 days. To identify regions where convection is the dominant process, the ratio between the magnitude of the convective tendency and the sum of the magnitudes of all tendencies is calculated and is shown in Fig. 1. This diagnostic can be used to define regions where there is little convection (the ratio is close to 0) or where convection dominates (the ratio is greater than 0.5). Since the forecasting skill of the IFS is strongly latitudinally dependent, both the regions with little convection and with significant convection are defined in the tropics (25°S–25°N). Both regions are approximately the same size and cover areas of both land and sea. Any differences in the forecast performance between these two regions will be predominantly due to convection.

b. Chosen diagnostics

Three variables of interest have been selected which will be used to verify the forecasts. Temperature and zonal wind at 850 hPa (T850 and U850, respectively) correspond to fields at approximately 1.5-km altitude, which falls above the boundary layer in many places. The zonal wind at 200 hPa (U200) is particularly interesting when considering convection. This is because 200 hPa falls close to the tropopause, where deep convection stops. Convective outflow often occurs at this level, which can be detected in U200. A fourth variable Z500 was also initially considered. However, the results from verifying this variable provided no additional information so are not presented here.

In convecting regions, precipitation (PPT) and total column water vapor (TCWV) are also considered. PPT is a product of the convection and large-scale water processes parameterization schemes, so an improvement to the convection scheme should be detectable by studying this variable. The convection scheme effectively redistributes and removes moisture from the atmosphere, so an improvement in TCWV could be indicative of an improvement in the convection scheme.

For each variable, the impact of the schemes will be evaluated using the following diagnostics.

1) BIAS

The bias of a forecast is defined as the systematic error in the ensemble mean:

$$\text{BIAS} = \langle m - z \rangle, \quad (3)$$

where z is the analysis, m is the ensemble mean, and the average indicated by the angle brackets is taken over the region of interest and all start dates.

2) SPREAD–ERROR RELATIONSHIP

The reliability of an ensemble forecast is tested through the spread–error relationship (Leutbecher and Palmer 2008; Leutbecher 2010). For a statistically consistent ensemble, the ensemble variance, s^2 , and squared ensemble mean error, $\langle m - z \rangle^2$, are related:

$$\begin{aligned} & \frac{n_{\text{ens}}}{n_{\text{ens}} - 1} \langle \text{estimate ensemble variance} \rangle \\ &= \frac{n_{\text{ens}}}{n_{\text{ens}} + 1} \langle \text{squared ensemble mean error} \rangle, \end{aligned} \quad (4)$$

where the variance and mean squared error have been estimated by averaging over many forecast–verification pairs. As described below, the average indicated by the angle brackets can be evaluated for the entire dataset (as in the bias calculation) or for subsets of the data. For large ensemble size, $n_{\text{ens}} \sim 50$, we can consider the correction factor to be close to 1.

FIG. 1. Convection diagnostic (color) derived from the IFS tendencies calculated as part of the YOTC project (see text for details). (a) Regions where the diagnostic is close to zero (bounded by white boxes), indicating there is little convection. (b) Regions where the diagnostic is large (bounded by white box), indicating convection is the dominant process.

This measure of reliability will be assessed in two ways. First, the RMSE and RMS ensemble spread (RMSS) are evaluated as a function of time for the entire sample of cases, and the two are compared. This gives a good summary of the forecast calibration. However, Eq. (4) can be used in a stricter sense—the equation should also be satisfied for subsamples of the forecast cases conditioned on the spread. This diagnoses the ability of the forecasting system to make flow-dependent uncertainty estimates. This measure can be assessed visually by binning the cases into subsamples of increasing RMSS and plotting against the average RMSE in each bin. The plotted points should lie

on the diagonal (“the RMS error–spread graphical diagnostic”).

3) SKILL SCORES

Three proper scores will be used to establish the skill of the probabilistic forecasts.

The ranked probability score (RPS) is a scoring rule used to evaluate a multicategory forecast. The RPS is defined as the squared sum of the difference between forecast and observed cumulative probabilities (Wilks 2006, 255–336). Defining the forecast (observed) probability of event j to be y_j (o_j), and the total number of event categories to be J ,

$$Y_m = \sum_{j=1}^m y_j, \quad O_m = \sum_{j=1}^m o_j; \quad m = 1, 2, \dots, J, \quad (5)$$

and

$$\text{RPS} = \sum_{m=1}^J (Y_m - O_m)^2. \quad (6)$$

A smaller RPS represents a better forecast.

The ignorance score (IGN) evaluates a forecast based on the information it contains (Roulston and Smith 2002). As for the RPS, define J event categories and consider N forecast–observation pairs. The forecast probability that the k th verification will be event i is defined to be $f(k)_i$ (where $i = 1, 2, \dots, J$ and $k = 1, 2, \dots, N$). If the corresponding outcome event was $j(k)$,

$$\text{IGN} = -\frac{1}{N} \sum_{k=1}^N \log_2 f(k)_{j(k)}, \quad (7)$$

where the score has been averaged over the N forecast–verification pairs. A smaller IGN represents a better forecast.

The error–spread score (ES) was proposed by Christensen et al. (2015a) as a score particularly suitable for evaluation of ensemble forecasts. It is formulated with respect to the moments of the forecast and is a proper score that is sensitive to both resolution and reliability. The ES is written

$$\text{ES} = \frac{1}{N} \sum_{k=1}^N (s_k^2 - e_k^2 - e_k s_k g_k)^2, \quad (8)$$

where the difference between the verification z_k and the ensemble mean m_k is the error in the ensemble mean,

$$e_k = (m_k - z_k), \quad (9)$$

s_k is the ensemble standard deviation, and g_k is the ensemble skewness. The ES is calculated by averaging over many forecast–verification pairs k both from different grid point locations and from different starting dates. A smaller average value of ES indicates a better set of forecasts.

Each of the above scores S may be converted into a skill score (SS) by comparison with the score evaluated for a reference forecast S_{ref} :

$$\text{SS} = \frac{(S - S_{\text{ref}})}{(S_{\text{perf}} - S_{\text{ref}})}. \quad (10)$$

For all the scoring rules above the perfect score, S_{perf} is zero, and the skill score can be expressed as

$$\text{SS} = 1 - \frac{S}{S_{\text{ref}}}, \quad -\infty < \text{SS} \leq 1. \quad (11)$$

For each skill score (RPSS, IGNSS, and ESS, respectively), the higher the skill score, the better the forecast, up to a perfect skill score of 1.

6. Verification of forecasts

a. Verification in nonconvecting regions

First, the impact of the different representations of model uncertainty will be considered in the non-convecting regions defined in Fig. 1a. Figures 2a–c show the bias of forecasts for regions of little convection. The bias is given as a fraction of the bias for the TSCZ forecasts: a fractional bias with magnitude less than one indicates an improvement over the TSCZ forecast model. Bias can indicate the presence of systematic errors in a forecast. A small change is observed in the bias when different uncertainty schemes are used; in particular, the TSCV scheme performs well for all variables considered. For U850 and U200, the TSCZ scheme outperforms the TSCS scheme—a result that will be discussed in section 7.

The impact of the uncertainty schemes on the calibration of the ensemble can be summarized by evaluating the RMSE and the RMSS as a function of time for the region of interest, which should be equal for a well-calibrated ensemble. Figures 2d–f show this diagnostic for the regions with little convection. Forecasts are slightly underdispersive for all variables. The RMS spread and error curves were also considered for operational ensemble forecasts at T639 (not shown). At T639, the ensemble spread is similar, but the RMSE is smaller. The low resolution (T159) used here is responsible for the higher RMSE and therefore contributes to the underdispersive nature of the ensemble.

Considering the RMS error–spread diagnostic gives a more comprehensive understanding of the calibration of the ensemble. The forecast–verification pairs for each spatial point and start date are collated. These pairs are ordered according to their forecast variance and divided into 30 equally populated bins. The RMSS and RMSE are evaluated for each bin and displayed on scatterplots. Figure 3 shows this diagnostic for regions with little convection at lead times of 1, 3, and 10 days. The scattered points should lie on the one-to-one diagonal, shown in black, for a statistically consistent ensemble. If the scattered points fall above the line, the ensemble is

FIG. 2. Forecast diagnostics as a function of time in tropical regions with little convection. Forecast bias for (a) T850, (b) U850, and (c) U200. Results are shown for each experiment: TSCS (black square), TSCF (dark gray upward triangle), and TSCV (midgray downward triangle). The bias is calculated following (3) and given as a fraction of the bias for TSCZ: $\text{BIAS}_{\text{TSCX}}/\text{BIAS}_{\text{TSCZ}}$. Root-mean-square ensemble spread (dashed lines) and root-mean-square error (solid lines) for (d) T850, (e) U850, and (f) U200. Results are shown for each experiment: TSCZ (light gray circle), TSCS (black square), TSCF (dark gray upward triangle), and TSCV (midgray downward triangle).

underdispersive—the ensemble spread is systematically smaller than the RMS error in the ensemble mean. Conversely, if the scattered points fall below the line, the ensemble is overdispersive. As expected, the results from the four experiments are very similar and show moderately underdispersive but otherwise well-calibrated forecasts. Figure 3 indicates a large degree of flow-dependent spread in the ensemble forecasts in non-convecting regions—the forecast ensemble spread is a good predictor of the observed RMS error in the ensemble mean, and the scattered points lie close to the one-to-one line. The T850 forecasts are particularly well calibrated and the spread of the U850 and U200 forecasts are also a skillful indicator of the expected error for all four experiments.

b. Verification in convecting regions

By considering nonconvecting regions, the previous section indicates that the model uncertainty from the other four tendencies is well represented by SPPT. To evaluate the impact of the new uncertainty schemes, this section considers forecasts for the strongly convecting

regions defined in Fig. 1b. Figures 4a–c show the bias for forecasts of T850, U850, and U200 in this region for the four different schemes considered. The bias is similar for all schemes; no one scheme is systematically better or worse than the others.

Figures 4d–f show the average RMS error and spread as a function of time for the region of interest. The RMSE is similar for all experiments—the perturbed parameter ensembles have not resulted in an increase in error over the operational scheme, except for a slight increase for T850. However, the fixed perturbed parameter ensemble has resulted in an increase in spread over the operational TSCS forecast. This is especially large for T850, where the observed increase is 25% at long lead times. Interestingly, the TSCZ “deterministic convection” forecasts for T850 also result in an increase in ensemble spread over TSCS. This is a counterintuitive result, as it is expected that a stochastic scheme would increase the spread of the ensemble. This result will be discussed in section 8. As is the case in regions with little convection, some of the ensemble underdispersion at T159 is due to an increased forecast RMSE compared to

FIG. 3. Root-mean-square error–spread diagnostic for tropical regions with little convection for (a)–(c) T850, (d)–(f) U850, and (g)–(i) U200 at lead times of (left) 1, (middle) 3, and (right) 10 days for each variable. Results are shown for each experiment: TSCZ (light gray circle), TSCS (black square), TSCF (dark gray upward triangle), and TSCV (midgray downward triangle). For a well-calibrated ensemble, the scattered points should lie on the one-to-one diagonal shown in black, following (4).

the operational T639 forecasts (not shown), though the forecasts are significantly underdispersive at both resolutions.

Figure 5 shows the RMS error–spread graphical diagnostic for the four forecast models in regions with significant convection. The impact of the different schemes is slight but greater than in regions with little

convection (cf. Fig. 3). All schemes remain well calibrated and do not show large increases in error compared to the operational TSCS forecasts. The fixed perturbed parameter scheme has larger spread than the other schemes, which is most apparent for T850 in Figs. 5a–c. For U850 and U200, TSCS has the most underdispersive ensemble at long lead times, though is

FIG. 4. As in Fig. 2, but for regions with significant convection.

better calibrated than the other experiments at short lead times. The TSCV experiment has intermediate spread, improving on TSCS but underdispersive compared to the TSCF experiments.

The skill of the forecasts is evaluated using the RPS, IGN, and ES as a function of lead time and the skill scores evaluated with respect to the TSCZ forecast for the convecting region. The results are shown in Fig. 6 for each variable of interest. The TSCF scheme scores highly for a range of variables according to each score: it performs significantly better than the other forecasts for T850 according to RPSS and IGNSS and for U850 at later lead times according to IGNSS and ESS (see appendix for details of significance testing). For U200, the TSCS forecasts are significantly better than the other forecasts, and the TSCZ forecasts are significantly poorer. However for the other variables, TSCS performs comparatively poorly and often produces significantly the worst forecasts. This is probably due to the poorer forecast ensemble spread.

1) PRECIPITATION FORECASTS

The impact of the different model uncertainty schemes on forecasts of convective precipitation is a

good indicator of improvement in the convection scheme. The Global Precipitation Climatology Project (GPCP) dataset is used for verification of precipitation forecasts. The GPCP, established by the World Climate Research Programme (WCRP), combines information from a large number of satellite- and ground-based sources to estimate the global distribution of precipitation. The dataset used here is the One-Degree Daily (1DD) product (Huffman et al. 2001), which has been conservatively regridded onto a T159 reduced Gaussian grid to allow comparison with the IFS forecasts.

Figure 7 shows the RMS error–spread diagnostic for convective precipitation. All forecasts are underdispersive, and the different uncertainty schemes have only a slight impact on the calibration of the ensemble. Figure 8b indicates more clearly the impact of the different schemes on the ensemble spread and error. On average, the TSCZ scheme is significantly the most underdispersive and has a significantly larger RMSE. The two stochastic schemes, TSCS and TSCV, have significantly the smallest error. TSCS has significantly the largest spread at short lead times, and TSCF has significantly the largest spread at later lead times. Figure 8a shows the bias in forecasts of convective precipitation. The stochastic schemes, TSCS and TSCV,

FIG. 5. As in Fig. 3, but for regions with significant convection.

have the smallest bias over the entire forecasting window. Figures 8c–e show the forecast skill scores for convective precipitation. TSCF produces significantly the most skillful forecasts at later lead times according to RPS and IGN. TSCZ is significantly the poorest according to RPS (i.e., all schemes show a significant improvement in skill over TSCZ), and ES and IGN also score TSCZ as significantly the worst at early lead times.

The spatial distribution of cumulative precipitation (convective plus large scale) was also considered for the different forecast models. All schemes performed equally well (not shown). When compared to the GPCP data, all showed too much precipitation over the ocean, and in particular forecast intensities of rain in the intertropical and South Pacific convergence zones that were higher than observed. The results were indistinguishable by eye—the difference between forecast

FIG. 6. Ensemble forecast skill scores calculated for tropical regions with significant convection. (a),(d),(g) Ranked probability skill score. (b),(e),(h) Ignorance skill score. (c),(f),(i) Error-spread skill score. (a)–(c) T850, (d)–(f) U850, and (g)–(i) U200. Skill scores are shown for each experiment, calculated with respect to the skill of the TSCZ experiment: TSCS (black square), TSCF (dark gray upward triangle), and TSCV (midgray downward triangle).

and observations is far greater than the differences between forecasts of different models.

2) TOTAL COLUMN WATER VAPOR

The impact of the different model uncertainty schemes on forecasts of TCWV is also a good indicator of improvement in the convection scheme. Figure 9 shows the RMS error–spread diagnostic for TCWV. The

forecasts for this variable are poorly calibrated when compared to convective precipitation. The RMSE is systematically larger than the spread, and the slope of the scattered points is too shallow. This shallow slope indicates that the forecasting system is unable to distinguish between cases with low and high predictability for this variable—the expected error in the ensemble mean is poorly predicted by the ensemble spread. The

FIG. 7. RMS error–spread diagnostic for cumulative convective precipitation for the 24-h window before a lead time of (a) 1, (b) 3, and (c) 10 days. The diagnostic is calculated for tropical regions with significant convection. Results are shown for each experiment: TSCZ (light gray circle), TSCS (black square), TSCF (dark gray upward triangle), and TSCV (midgray downward triangle). For a well-calibrated ensemble, the scattered points should lie on the one-to-one diagonal shown in black.

different forecast schemes show a larger impact than for forecasts of precipitation—the TSCS model produces forecasts that are underdispersive compared to the other forecasts. Figure 10 shows the bias (Fig. 10a), the RMSE and spread as a function of time (Fig. 10b), and the forecast skill scores for each experiment (Figs. 10c–e). Figure 10b shows that the TSCV forecasts have significantly the largest spread at lead times of 24 h and greater. The TSCS forecasts have significantly the smallest spread at later lead times and significantly the largest error at all lead times. Figure 10a shows the bias is also largest for the TSCS forecasts. This increase in bias and RMSE and, in particular, the reduction in spread compared to the TSCZ forecasts, results in a large reduction in TSCS forecast skill at long lead times, as shown in Figs. 10c–e. Possible causes for the observed reduction in spread in TSCS forecasts are discussed in section 7.

An early version of SPPT was found to dry out the tropics and resulted in a decrease in TCWV of approximately 10% (M. Leutbecher 2013, personal communication). This was corrected in a later version of the SPPT scheme. It is possible that TCWV could be sensitive to the proposed perturbed parameter representations of model uncertainty. The average TCWV between 20°N and 20°S is averaged over all start dates separately for each ensemble member and is shown in Fig. 11. Initially, all experiments show a drying of the tropics of approximately 0.5 kg m^{-2} over the first 12 h, indicating a spinup period in the model. The TSCZ, TSCS, and TSCV forecasts then stabilize. However, each ensemble member in the TSCF model has vastly different behavior, with some

showing systematic drying, and others showing systematic moistening over the 10-day forecast.

7. Discussion

The results indicate that the perturbed parameter schemes have a small positive impact on the IFS. Introducing the TSCF scheme does not lead to increased bias in T850, U850, or U200, indicating that systematic errors in these fields have not increased. An increase in ensemble spread is observed when the perturbed parameter schemes are used to represent uncertainty in convection instead of SPPT, and the TSCF forecasts have significantly the largest spread for T850 and U850 forecasts, which Fig. 5 indicates is flow dependent. The perturbed parameter schemes produce significantly the most skillful forecasts of T850 and U850 as ranked by the RPSS, IGNSS, and ESS.

Initially, it appears that these results indicate that using a fixed perturbed parameter ensemble instead of SPPT improves the representation of uncertainty in convection. However, the fixed perturbed parameter ensembles remain underdispersive. While an increase in spread is observed for TSCF compared to TSCS, a substantial proportion of this increase is also observed in Fig. 4 when SPPT is switched off for the convection scheme for TSCZ. Since SPPT is switched off for convection in TSCF, the parameter perturbations are contributing only slightly to the spread of the ensemble and much of the spread increase can be attributed to decoupling the convection scheme from SPPT. The small

FIG. 8. Forecast diagnostics for 24-h cumulative convective precipitation (prior to the indicated lead time) in tropical regions with significant convection. (a) Forecast bias as a fraction of the bias for TSCZ: TSCS (black square), TSCF (dark gray upward triangle), and TSCV (midgray downward triangle). (b) Temporal evolution of RMS ensemble spread (dashed lines) and error (solid lines) averaged over the region: TSCZ (light gray circle), TSCS (black square), TSCF (dark gray upward triangle), and TSCV (midgray downward triangle). (c) Ranked probability skill score, (d) ignorance skill score, and (e) error–spread skill score. The skill scores are calculated with reference to the skill of the TSCZ forecast for each forecast experiment: TSCS (black square), TSCF (dark gray upward triangle), and TSCV (midgray downward triangle).

impact of the perturbed parameter scheme indicates that the schemes are not fully capturing the uncertainty in the convection parameterization. This is surprising as the parameter uncertainty has been explicitly measured and used to develop the scheme.

The TSCV scheme had a positive impact on the skill of the weather forecasts and significantly improved over the TSCZ and TSCS forecasts for many diagnostics. The impact on spread and skill was smaller than the static perturbed parameter schemes. It is possible that the parameter perturbations vary on too fast a time scale for a significant impact to be observed—if the parameters varied more slowly, a larger, cumulative effect could be observed in the forecasts. It would be interesting to test the TSCV scheme using a longer correlation time scale to test this hypothesis.

The two types of perturbed parameter scheme presented here represent fundamentally different error

models. Fixed perturbed parameter schemes are based on the assumption that there exists some optimal (or “correct”) value of the parameters in the parameterization scheme. The optimal parameters cannot be precisely known, so a perturbed parameter ensemble samples from a set of likely parameter values. The fixed perturbed parameter ensembles tested in this paper were underdispersive and did not fully capture the uncertainty in the forecasts. This indicates that fixed parameter uncertainty is not the only source of model uncertainty and that fixed perturbed parameter ensembles cannot be used alone to represent model uncertainty in an atmospheric simulation. While parameter uncertainty could account for systematic errors in the forecast, the results indicate that some component of the error cannot be captured by a deterministic uncertainty scheme. In particular, perturbed parameter ensembles are unable to represent structural uncertainty owing to

FIG. 9. RMS error–spread diagnostic for total column water vapor for lead times of (a) 1, (b) 3, and (c) 10 days. The diagnostic is calculated for tropical regions with significant convection. Results are shown for each experiment: TSCZ (light gray circle), TSCS (black square), TSCF (dark gray upward triangle), and TSCV (midgray downward triangle). For a well-calibrated ensemble, the scattered points should lie on the one-to-one diagonal shown in black.

the choices made when developing the parameterization scheme, and a different approach is required to represent uncertainties because of the bulk formula assumption.

The second error model recognizes that in atmospheric modeling there is not necessarily a “correct” value for the parameters in the physics parameterization schemes. Instead there exists some optimal distribution of the parameters in a physical scheme. Since in many cases the parameters in the physics schemes have no direct physical interpretation, but represent a group of interacting processes, it is likely that their optimal value may vary from day to day, or from grid box to grid box, or they may be seasonally or latitudinally dependent. A stochastically perturbed parameter ensemble represents this parameter uncertainty. The stochastically perturbed parameter scheme also underestimated the error in the forecasts. Even generalized to allow varying parameters, parameter uncertainty is not the only source of model uncertainty in weather forecasts. Not all subgrid-scale processes can be accurately represented using a statistical parameterization scheme, and some forecast errors cannot be represented using the phase space of the parameterized tendencies.

The EPPES indicated that the uncertainty in the convection parameters was moderate and smaller than expected (H. Järvinen 2013, personal communication). The results presented here also indicate larger parameter perturbations could be necessary to capture the uncertainty in the forecast from the convection scheme. However, the average tropical total column water vapor indicates that even these moderate perturbations are

sufficient for biases to develop in this field over the 10-day forecast period. The different fixed parameter settings result in vastly different behaviors, with some ensemble members showing a systematic drying and others a moistening in this region. This is concerning. The fact that this problem develops noticeably over 10 days indicates that this could be a serious problem in climate prediction, where longer forecasts could result in larger biases developing. This result supports previous work considering the climate of perturbed parameter ensembles (Christensen et al. 2015b): individual perturbed parameter ensemble members were observed to have vastly different climatological regime behavior in the context of the Lorenz-96 “toy model” of the atmosphere. The TSCV forecasts did not develop biases in this way, as the parameter sets for each ensemble member varied over the course of the forecast: stochastically perturbed parameter ensembles could be an attractive way of including parameter uncertainty into weather and climate forecasts.

An interesting result is that removing the stochastic perturbations from the convection tendency increased the forecast spread for some variables. This is observed for T850, U850, and TCWV in both regions considered and for U200 in nonconvecting regions. In fact, this is as expected from the formulation of the IFS. SPPT does not represent uncertainty in individual tendencies but assumes uncertainty is proportional to the total tendency. In the IFS, convection acts to reduce the sum of the tendencies: the deep convection scheme calculates the warming effects of unresolved convection, but the evaporative cooling of

FIG. 10. As in Fig. 8, but for total column water vapor.

moist, detrained air is calculated by the LSWP scheme. Therefore in the TSCZ forecast model, the tendencies perturbed by SPPT tended to be larger than in the TSCS model, so a larger forecast spread was observed.

Despite the reduced spread, the TSCS scheme outperforms the TSCZ scheme according to other forecast diagnostics. The T850 RMSE is reduced using the TSCS scheme, increasing the skill scores, and TSCS is significantly more skillful than TSCZ out to day 3 for U850. Additionally, TSCS results in an increase of spread for U200 and precipitation compared to TSCZ. It is important to note that the parameterization tendencies are vectors corresponding to the tendency at different vertical levels and that SPPT uses the same stochastic perturbation field at each level.¹ The convection scheme is sensitive to the vertical distribution of temperature and humidity, and it is possible that the parameterized convective tendencies act to damp or excite the scheme at subsequent time steps. Therefore, perturbing the

convective (vector) tendency using SPPT could lead to an increased variability in convective activity between ensemble members through amplification of this excitation process. Since U200 and convective precipitation are directly sensitive to the convection parameterization scheme, these variables can detect this increased variability, resulting in an increase in ensemble spread. In fact, TSCS has significantly the most skillful forecasts between days 3 and 10 for U200 and shows a significant reduction in forecast bias and RMSE for convective precipitation. T850 and U850 are less sensitive to convection than U200 and precipitation. Since in general the total perturbed tendency is reduced for TSCS compared to TSCZ, this could lead to the reduction in ensemble spread observed for these variables.

The experiments presented in this study have used the IFS at a resolution of T159. Nevertheless, the experiments give a good indication of what the impact of the different schemes would be on the skill of the operational IFS at a resolution of T639—when certain experiments were repeated at T639, similar behavior was observed. Therefore, the low-resolution runs presented here can be used to indicate the expected results of the models at T639 and can

¹ The perturbation is constant vertically except for tapering in the boundary layer and the stratosphere.

FIG. 11. Average TCWV between 20°S and 20°N as a function of time. The spatial average is calculated for each ensemble member averaged over all start dates, and the averages for each of the 50 ensemble members are shown. Results are shown for the four experiments: (a) TSCZ, (b) TSCS, (c) TSCF, and (d) TSCV.

suggest whether it would be interesting to run further experiments at higher resolution.

8. Conclusions

This paper presents new representations of model uncertainty for the ECMWF convection scheme. The joint uncertainty in four of the parameters in the convection scheme, estimated using the EPPES (Ollinaho et al. 2013a), was used to construct a fixed (TSCF) and a stochastically varying (TSCV) perturbed parameter scheme. The performance of these schemes was compared to the ECMWF operational stochastic scheme (TSCS) and to a model that does not represent uncertainty in convection (TSCZ).

Forecasts made using the fixed perturbed parameter ensemble showed a significant improvement in forecast skill for the dynamical variables considered. However,

this does not necessarily indicate that such schemes represent uncertainty in convection better than the operational stochastic scheme. In fact, much of the skill improvement was observed on decoupling convection from the other schemes when applying SPPT (TSCZ was significantly more skillful than TSCS). The additional impact on forecast skill due to the perturbed parameter scheme was small, despite estimating the uncertainty in the parameters using a Bayesian approach. The stochastically perturbed parameter scheme tested in the IFS also had a small positive impact on skill. However, the temporal and spatial noise correlations were not estimated or tuned for this scheme, and the standard SPPT values were used instead. It is likely that the optimal noise parameters would be different for a perturbed parameter scheme than for SPPT, and using measured noise parameters could result in further improvement in the forecast skill.

The average total column water vapor in the tropics was considered as a function of forecast lead time. The fixed perturbed parameter scheme resulted in undesirable biases in moisture in the tropics. Conversely, the stochastically varying perturbed parameter scheme did not show such biases and could be an attractive technique for including information about parameter uncertainty into a forecast model.

Verification of variables that are particularly sensitive to convection was also performed. For these variables, the TSCS scheme performed well, significantly reducing the bias and RMSE in forecasts of convective precipitation and producing significantly the most skillful forecasts for U200 at longer lead times, though an improvement was not observed for TCWV. This indicates that SPPT is a skillful way of representing model uncertainty due to convection parameterization.

The ensemble forecasts produced by the perturbed parameter schemes were underdispersive. This indicates that parameter uncertainty is not the only source of model uncertainty in atmospheric models. In fact, assumptions and approximations made when constructing the parameterization scheme also result in errors that cannot be represented by varying uncertain parameters. It would be interesting to combine the two types of schemes investigated here, using a perturbed parameter scheme to represent the parameter uncertainty while a stochastic parameterization scheme could be used to represent other sources of model uncertainty (e.g., unresolved subgrid variability or structural uncertainty). For completeness, the characteristics of the stochastic term should also be measured and included in the forecast model. This will be the subject of a future study.

Acknowledgments. The authors thank Pirkka Ollinaho for providing the joint parameter uncertainties used to construct the perturbed parameter schemes presented in this paper. Thanks also to Heikki Järvinen and Peter Bechtold for helpful discussion regarding the EPPEs and the ECMWF convection scheme, respectively. The research of H.M.C. was supported by a Natural Environment Research Council studentship, and the research of T.N.P. was supported by European Research Council Grant 291406.

APPENDIX

Significance Testing

To state that one parameterization is better than another, it is necessary to know how significantly different one skill score is from another. A Monte Carlo technique

is used to evaluate the significance of the difference, D , between two skill scores (SS), assuming the null hypothesis that the two forecasts have equal skill. Consider two vectors, $\mathbf{A}$ and $\mathbf{B}$, which contain the values of the skill score evaluated for each forecast–verification pair for forecast models A and B, respectively. The vectors are each of length n , where n is the number of forecast–verification pairs.

If the forecasts have equal skill, the elements of $\mathbf{A}$ and $\mathbf{B}$ are interchangeable: any apparent difference in skill of the forecast systems is due to chance. To test this, the elements of $\mathbf{A}$ and $\mathbf{B}$ are shuffled and the skill of the shuffled vectors calculated. Since the difference between the skill of forecast system A under predictable and unpredictable flow conditions is likely to be greater than the difference between forecast system A and B for the same conditions, the shuffling is performed pairwise. For the tropical region with significant convection, spatial correlations were measured to drop below 0.5 for all variables by 1500–2000 km (approximately 15° at the equator). Therefore, to preserve the spatial correlation in the dataset to a large degree, the skill scores for each forecast are split into blocks $15^\circ \times 15^\circ$, which are then treated independently.

The difference in skill for the shuffled vectors, $D_{\text{shuf}} = \text{SS}(\mathbf{A}_{\text{shuf}}) - \text{SS}(\mathbf{B}_{\text{shuf}})$, is evaluated and the vectors reshuffled. If the proportion of D_{shuf} smaller than D is greater than (less than) 0.95 (0.05), A is considered significantly more (less) skillful than B at the 95% level.

REFERENCES

- Arnold, H. M., 2013: Stochastic parametrisation and model uncertainty. Ph.D. thesis, University of Oxford, 238 pp.
- , I. M. Moroz, and T. N. Palmer, 2013: Stochastic parameterizations and model uncertainty in the Lorenz '96 system. *Philos. Trans. Roy. Soc. London*, **A373**, doi:10.1098/rsta.2011.0479.
- Bechtold, P., N. Semane, P. Lopez, J.-P. Chaboureau, A. Beljaars, and N. Bormann, 2014: Representing equilibrium and non-equilibrium convection in large-scale models. *J. Atmos. Sci.*, **71**, 734–753, doi:10.1175/JAS-D-13-0163.1.
- Beck, J. V., and K. J. Arnold, 1977: *Parameter Estimation in Engineering and Science*. Wiley, 522 pp.
- Bengtsson, L., M. Steinheimer, P. Bechtold, and J.-F. Geleyn, 2013: A stochastic parameterization for deep convection using cellular automata. *Quart. J. Roy. Meteor. Soc.*, **139**, 1533–1543, doi:10.1002/qj.2108.
- Berner, J., G. J. Shutts, M. Leutbecher, and T. N. Palmer, 2009: A spectral stochastic kinetic energy backscatter scheme and its impact on flow dependent predictability in the ECMWF ensemble prediction system. *J. Atmos. Sci.*, **66**, 603–626, doi:10.1175/2008JAS2677.1.
- Bowler, N. E., A. Arribas, K. R. Mylne, K. B. Robertson, and S. E. Beare, 2008: The MOGREPS short-range ensemble

- prediction system. *Quart. J. Roy. Meteor. Soc.*, **134**, 703–722, doi:[10.1002/qj.234](https://doi.org/10.1002/qj.234).
- Buizza, R., M. Leutbecher, and L. Isaksen, 2008: Potential use of an ensemble of analyses in the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **134**, 2051–2066, doi:[10.1002/qj.346](https://doi.org/10.1002/qj.346).
- Christensen, H. M., I. M. Moroz, and T. N. Palmer, 2015a: Evaluation of ensemble forecast uncertainty using a new proper score: Application to medium-range and seasonal forecasts. *Quart. J. Roy. Meteor. Soc.*, doi:[10.1002/qj.2375](https://doi.org/10.1002/qj.2375), in press.
- , —, and —, 2015b: Simulating weather regimes: Impact of stochastic and perturbed parameter schemes in a simple atmospheric model. *Climate Dyn.*, doi:[10.1007/s00382-014-2239-9](https://doi.org/10.1007/s00382-014-2239-9), in press.
- Craig, G. C., and B. G. Cohen, 2006: Fluctuations in an equilibrium convective ensemble. Part I: Theoretical formulation. *J. Atmos. Sci.*, **63**, 1996–2004, doi:[10.1175/JAS3709.1](https://doi.org/10.1175/JAS3709.1).
- Huffman, G. J., R. F. Adler, M. M. Morrissey, D. T. Bolvin, S. Curtis, R. Joyce, B. McGavock, and J. A. Susskind, 2001: Global precipitation at one-degree daily resolution from multi-satellite observations. *J. Hydrometeorol.*, **2**, 36–50, doi:[10.1175/1525-7541\(2001\)002<0036:GPAODD>2.0.CO;2](https://doi.org/10.1175/1525-7541(2001)002<0036:GPAODD>2.0.CO;2).
- Isaksen, L., M. Bonavita, R. Buizza, M. Fisher, J. Haseler, M. Leutbecher, and L. Raynaud, 2010: Ensemble of data assimilations at ECMWF. European Centre for Medium-Range Weather Forecasts Tech. Rep. 636, 39 pp.
- Järvinen, H., M. Laine, A. Solonen, and H. Haario, 2012: Ensemble prediction and parameter estimation system: The concept. *Quart. J. Roy. Meteor. Soc.*, **138**, 281–288, doi:[10.1002/qj.923](https://doi.org/10.1002/qj.923).
- Khouider, B., and A. J. Majda, 2006: A simple multilayer parameterization for convectively coupled tropical waves. Part I: Linear analysis. *J. Atmos. Sci.*, **63**, 1308–1323, doi:[10.1175/JAS3677.1](https://doi.org/10.1175/JAS3677.1).
- , —, and M. A. Katsoulakis, 2003: Coarse-grained stochastic models for tropical convection and climate. *Proc. Natl. Acad. Sci. USA*, **100**, 11 941–11 946, doi:[10.1073/pnas.1634951100](https://doi.org/10.1073/pnas.1634951100).
- Knight, C. G., and Coauthors, 2007: Association of parameter, software, and hardware variation with large-scale behavior across 57,000 climate models. *Proc. Natl. Acad. Sci. USA*, **104**, 12 259–12 264, doi:[10.1073/pnas.0608144104](https://doi.org/10.1073/pnas.0608144104).
- Laine, M., A. Solonen, H. Haario, and H. Järvinen, 2012: Ensemble prediction and parameter estimation system: The method. *Quart. J. Roy. Meteor. Soc.*, **138**, 289–297, doi:[10.1002/qj.922](https://doi.org/10.1002/qj.922).
- Leutbecher, M., 2010: Diagnosis of ensemble forecasting systems. *Seminar on Diagnosis of Forecasting and Data Assimilation Systems*, Reading, United Kingdom, ECMWF, 235–266.
- , and T. N. Palmer, 2008: Ensemble forecasting. *J. Comput. Phys.*, **227**, 3515–3539, doi:[10.1016/j.jcp.2007.02.014](https://doi.org/10.1016/j.jcp.2007.02.014).
- Lin, J. W.-B., and J. D. Neelin, 2000: Influence of a stochastic moist convective parameterization on tropical climate variability. *Geophys. Res. Lett.*, **27**, 3691–3694, doi:[10.1029/2000GL011964](https://doi.org/10.1029/2000GL011964).
- , and —, 2003: Towards stochastic deep convective parameterization in general circulation models. *Geophys. Res. Lett.*, **30**, 1162, doi:[10.1029/2002GL016203](https://doi.org/10.1029/2002GL016203).
- Majda, A. J., and B. Khouider, 2002: Stochastic and mesoscopic models for tropical convection. *Proc. Natl. Acad. Sci. USA*, **99**, 1123–1128, doi:[10.1073/pnas.032663199](https://doi.org/10.1073/pnas.032663199).
- Murphy, J. M., D. M. H. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, M. Collins, and D. A. Stainforth, 2004: Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, **430**, 768–772, doi:[10.1038/nature02771](https://doi.org/10.1038/nature02771).
- Ollinaho, P., P. Bechtold, M. Leutbecher, M. Laine, A. Solonen, H. Haario, and H. Järvinen, 2013a: Parameter variations in prediction skill optimization at ECMWF. *Nonlinear Processes Geophys.*, **20**, 1001–1010, doi:[10.5194/npg-20-1001-2013](https://doi.org/10.5194/npg-20-1001-2013).
- , M. Laine, A. Solonen, H. Haario, and H. Järvinen, 2013b: NWP model forecast skill optimization via closure parameter variations. *Quart. J. Roy. Meteor. Soc.*, **139**, 1520–1532, doi:[10.1002/qj.2044](https://doi.org/10.1002/qj.2044).
- Palmer, T. N., R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G. J. Shutts, M. Steinheimer, and A. Weisheimer, 2009: Stochastic parametrization and model uncertainty. European Centre for Medium-Range Weather Forecasts Tech. Rep. 598, 42 pp.
- Plant, R. S., and G. C. Craig, 2008: A stochastic parameterization for deep convection based on equilibrium statistics. *J. Atmos. Sci.*, **65**, 87–104, doi:[10.1175/2007JAS2263.1](https://doi.org/10.1175/2007JAS2263.1).
- Roulston, M. S., and L. A. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Mon. Wea. Rev.*, **130**, 1653–1660, doi:[10.1175/1520-0493\(2002\)130<1653:EPFUIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(2002)130<1653:EPFUIT>2.0.CO;2).
- Sanderson, B. M., 2011: A multimodel study of parametric uncertainty in predictions of climate response to rising greenhouse gas concentrations. *J. Climate*, **24**, 1362–1377, doi:[10.1175/2010JCLI3498.1](https://doi.org/10.1175/2010JCLI3498.1).
- , C. Piani, W. J. Ingram, D. A. Stone, and M. R. Allen, 2008: Towards constraining climate sensitivity by linear analysis of feedback patterns in thousands of perturbed-physics GCM simulations. *Climate Dyn.*, **30** (2–3), 175–190, doi:[10.1007/s00382-007-0280-7](https://doi.org/10.1007/s00382-007-0280-7).
- Stainforth, D. A., and Coauthors, 2005: Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature*, **433**, 403–406, doi:[10.1038/nature03301](https://doi.org/10.1038/nature03301).
- Tiedtke, M., 1989: A comprehensive mass flux scheme for cumulus parameterization in large-scale models. *Mon. Wea. Rev.*, **117**, 1779–1800, doi:[10.1175/1520-0493\(1989\)117<1779:ACMFSF>2.0.CO;2](https://doi.org/10.1175/1520-0493(1989)117<1779:ACMFSF>2.0.CO;2).
- Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. International Geophysics Series, Vol. 91, Elsevier, 648 pp.
- Yokohata, T., M. J. Webb, M. Collins, K. D. Williams, M. Yoshimori, J. C. Hargreaves, and J. D. Annan, 2010: Structural similarities and differences in climate responses to CO₂ increase between two perturbed physics ensembles. *J. Climate*, **23**, 1392–1410, doi:[10.1175/2009JCLI2917.1](https://doi.org/10.1175/2009JCLI2917.1).