TECHNICAL MEMORANDUM

784

Towards process-level representation of model uncertainties: Stochastically perturbed parametrisations in the ECMWF ensemble

Pirkka Ollinaho[1], Sarah-Jane Lock, Martin Leutbecher, Peter Bechtold, Anton Beljaars, Alessio Bozzo, Richard M. Forbes, Thomas Haiden, Robin J. Hogan, and Irina Sandu

Research Department

[1]Finnish Meteorological Institute, Finland

Submitted to the Quarterly Journal

September 2016

European Centre for Medium-Range Weather Forecasts
Europäisches Zentrum für mittelfristige Wettervorhersage
Centre européen pour les prévisions météorologiques à moyen terme

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:
http://www.ecmwf.int/en/research/publications

Contact: library@ecmwf.int

European Centre for Medium-Range Weather Forecasts
Shinfield Park, Reading, RG2 9AX, England

Abstract

Ensemble forecasts depend on representations of model uncertainties. Here, we introduce a model uncertainty representation where a novel approach is taken to the established methodology of perturbing model parameters. The Stochastically Perturbed Parametrisations (SPP) scheme applies spatially and temporally varying perturbations to 20 parameters and variables in the ECMWF IFS model. The perturbed quantities are chosen from the IFS parametrisations of (a) turbulent diffusion and subgrid orography, (b) convection, (c) clouds and large-scale precipitation, and (d) radiation. The perturbations are drawn from prescribed distributions.

Numerous configurations of SPP are compared in experiments with the ECMWF ensemble forecasts at T_L399 resolution up to 15 day lead times. Halving the standard deviations of the perturbations considerably reduces the ensemble spread. Smaller variations of the standard deviations lead to minor changes to the ensemble spread. Experiments with different space and time correlations for the perturbations suggest optimal correlation scales of 2,000 km and 72 h. SPP displays a lower skill for upper air variables in the medium range than the current operational model uncertainty scheme Stochastically Perturbed Parametrisation Tendencies (SPPT) for a given set of fixed initial state perturbations. However, in short ranges the two schemes display a similar skill. Moreover, verification against surface observations shows SPP is more skilful than SPPT in 2 m-temperature for the first couple of forecast days. We show that the direct perturbation of cloud (and radiation) processes in SPP has a greater impact on radiative fluxes than the indirect perturbation via SPPT. SPP also produces a better model climate for a range of variables when comparing long model integrations with the two schemes, indicating the potential advantage of a physically consistent model uncertainty representation. A comparison of the tendency perturbations introduced by SPP and SPPT suggests that the two schemes represent different aspects of model uncertainty.

1 Introduction

Meteorological services increasingly issue outputs from ensemble forecasts to indicate the level of confidence in a given forecast. An ensemble forecast comprises multiple individual forecasts (the ensemble members), where each ensemble member differs to some degree in its description of the atmospheric state (see e.g. ECMWF, 2015). The purpose of an ensemble forecast is to represent uncertainties in the underlying deterministic forecasting system. That uncertainty derives from errors in the initial conditions and from the model itself (see e.g. Leutbecher and Palmer, 2008). This present study focuses on the latter, the so-called "model uncertainty" (or "model error"), that derives from inaccuracies in the methods employed to integrate forward-in-time from a given state.

Model uncertainty arises due to deficiencies in the methods used to simulate the complex interactions occurring in the atmosphere. The associated model uncertainties comprise large-scale errors occurring due to simplifications made necessary by computational constraints (e.g. using climatological rather than prognostic aerosol distributions); but also errors in the very smallest scales due to unresolved processes that must be parametrised at the sub-grid scale (e.g. convection and the effect of gravity waves). Furthermore, the large-scale errors will influence the small scale ones, and vice versa. The processes that need to be parametrised depend on the resolution of the model. For a typical medium-range (3-15 days) global Numerical Weather Prediction (NWP) model, these include convection and cloud processes, turbulent mixing, radiative transfer, orographic and non-orographic gravity wave drag. The methods used to account for the uncertainties in these schemes can vary between different Ensemble Prediction Systems (EPS). For example, the Environment Canada Global EPS uses a "multi-physics" approach (Charron et al., 2010), in which different physical parametrisations for the same processes are used within an ensemble; a "multi-model" system, where the ensemble consists of multiple different models, is utilised in some limited area models (see e.g. Iversen et al., 2011); the UK Met Office EPS (Bowler et al.,

2008) includes "random parameters", where model uncertainties are represented by perturbing some of the physical parametrisations directly; and the ensembles from the European Centre for Medium-Range Weather Forecasts (ECMWF) use "stochastic parametrisations" (Palmer et al., 2009), that account for model uncertainties arising from errors in the net effect of the physical parametrisations. The two latter forecasting centres also apply a Stochastic Kinetic Energy Backscatter (SKEB; Berner et al., 2009) scheme to simulate upscale-propagating errors. By including some representation of model uncertainty, there is greater scope for the individual ensemble member forecasts to diverge from each other, yielding greater ensemble spread. Since ensemble forecasts with only a representation of initial condition uncertainties tend to be under-dispersive (and, thereby, over-confident), the increased spread from model uncertainty representations leads to improved forecast skill (e.g. Buizza et al., 1999). The real challenge, however, is to enhance ensemble spread (and the consequent forecast skill) through model uncertainty representations that retain physical consistency[1]. Work on model uncertainty representations that are physically consistent may ultimately lead to better understanding of the underlying model's shortcomings.

The ensemble forecasts from the ECMWF Integrated Forecasting System (IFS) include a representation of model uncertainty via the Stochastically Perturbed Parametrisation Tendencies (SPPT) scheme (Buizza et al., 1999; Palmer et al., 2009; Shutts et al., 2011). SPPT is designed to account for uncertainties associated with the sub-grid physics parametrisations — radiation, cloud and convection, diffusion and gravity wave drag schemes. At each timestep, all of the physical parametrisations change the prognostic model variables (winds, temperature and humidity) by a certain amount, that is generally defined as "tendency". The SPPT scheme randomly perturbs the net of these tendencies with multiplicative noise. Therefore, the scheme attributes the greatest uncertainty to the largest net tendencies while preserving the relative balances between the tendencies of different physical processes. Despite its relative simplicity, the SPPT scheme has yielded positive results. Analyses of the performance of the IFS ensembles — both from the ECMWF ensemble (ENS) for medium-range to sub-seasonal forecasts and System 4 (S4) for seasonal forecasts — demonstrate clear positive impacts due to the inclusion of SPPT (Shutts et al., 2011; Weisheimer et al., 2014).

The simplicity of the SPPT scheme comes at a cost. Applying multiplicative noise to the net physics tendencies can enhance or diminish the effect of the represented physical processes, but it is unable (in a single timestep) to trigger a new state, e.g. generation of a cloud layer, thereby, not directly capturing the large uncertainty associated with the timing and location of convection. Related, the dominant uncertainty in the radiation scheme arises from the presence (or lack) of clouds. By contrast, the radiative transfer process in clear-skies is well described by the radiation scheme (Pincus et al., 2003). In SPPT, tendencies due to both clear and cloudy skies radiative processes are perturbed alike. Indeed, in recognition of this shortcoming of SPPT, an exception is applied in the stratosphere: a tapering function reduces the SPPT perturbations to zero above 50 hPa, where the dominant contribution to the net physics tendencies is from clear-sky radiation. Another pragmatic choice is to apply a tapering function such that SPPT does not perturb the tendencies in the lowest 300 m, in this instance, in order to avoid numerical instabilities. An additional concern with SPPT relates to the inconsistency that arises between the perturbed physics tendencies and fluxes that are computed from the unperturbed tendencies: no correction is made to the top-of-the-atmosphere or surface fluxes after perturbing the atmospheric tendencies, so an energy imbalance is introduced into the system and individual ensemble members no longer conserve energy.

[1] We consider a physically consistent scheme one that conserves the fundamental aspects of the model physics (and dynamics), such as preservation of a local energy budget. For clarity, we want to stress that a physically consistent scheme should not be confused with a scheme that agrees completely with the true distribution of tendency errors.

A path towards a process-level stochastic representation of model uncertainties is to apply perturbations directly to poorly constrained parameters and variables within the parameterisation schemes. This leads to a model uncertainty representation that is closely linked to the existing parametrisations. Moreover, it can be directly related to known sources of model uncertainties associated with specific processes. By their very nature, sub-grid parametrisation schemes simulate processes that are either too fast or too small-scale to be resolved directly by the model. The schemes use assumptions and simplifications, as well as many poorly known parameters; these are then used in the parametrisations to yield bulk descriptions of processes that are otherwise either absent from or inadequately represented by the model (e.g. turbulent and convective mixing and transport, microphysical processes, surface exchanges).

In this study, individual IFS parameters and variables are chosen that are uncertain and that are known to have a significant impact on the model forecast skill. The latter knowledge comes from experience gained in the development of the deterministic model where global changes to these parameters have been explored. The new scheme will be referred to as the Stochastically Perturbed Parametrisations (SPP) scheme. The selected parameters are perturbed at each timestep with in-space varying noise derived from in-time evolving 2D random number fields. The scheme is designed such that it converges to the unperturbed (i.e. deterministic) forecast model in the limit of vanishing variance of the noise, a property shared by the current operational SPPT scheme. Thereby, the stochastic parameterisation samples a PDF that has a mode which is informed by a set of established physical parameterisations that play an important role in contributing to the skill of the forecasts.

Adding stochastic noise to a set of parameters is not a new idea: since 2008, the UKMO has been using a "random parameter" scheme, which currently perturbs a set of 16 parameters (Baker et al., 2014). However, their scheme introduces globally constant perturbations that vary only in time. Ollinaho et al. (2013b) envisaged using informed parameter covariances for applying parameter perturbations in ensemble forecasts. This was subsequently studied by Christensen et al. (2015) by perturbing four IFS convection parameters. They also experimented with in-space varying perturbations, but found only a small positive impact on the ENS skill. Although SPP is similar to this scheme, the methodologies for constructing the parameter perturbations differ significantly.

This study introduces the SPP scheme and describes numerical experimentation that will evaluate its probabilistic skill using ECMWF's medium-range ensemble forecasting system with a set of fixed initial state perturbations. The method for generating the stochastic perturbations is presented, alongside a description of the perturbed parameters (Section 2). A wide range of SPP configurations (Section 3) has been investigated in an effort to examine the sensitivity of the scheme to different (perturbation) configurations (Section 4). Differences between the SPP and SPPT schemes are studied through the vertical profiles of net physics tendencies (Section 5). Furthermore, differences between the schemes in medium-range forecasts are evaluated in terms of probabilistic skill scores for upper air and surface variables (Section 6). In addition, the response of the model climate is measured through a comparison with satellite observations and reanalysis data (Section 7). We also discuss various aspects of the SPP scheme (e.g. the configuration selection process and the computational cost) as well as envisage a possible setup for a future operational scheme within the ENS (Section 8), before concluding (Section 9).

2 Methodology

The current implementation of SPP allows simultaneous perturbations of up to 20 parameters and variables in the deterministic IFS parametrisations of (a) turbulent diffusion and subgrid orography, (b) convection, (c) cloud processes, and (d) radiation. These 20 parameters and variables have been chosen

based on expert insight into which parameters and variables in the parametrisations are known both to be uncertain (e.g. due to lack of direct observations), and to play a crucial role within the parametrisation, i.e. changing their values will have a notable impact on the forecasts. They are also chosen to describe the uncertainty across a range of meteorological regimes/phenomena and the entire atmosphere from boundary layer and free troposphere to the stratosphere.

2.1 Parameter selection

The 20 parameters and their role in the IFS parametrisations are summarized in Table 1. Below we also give a concise motivation for choosing these parameters for each of the parametrisations (a)–(d).

(a) Boundary layer processes are important for atmospheric models, because it is only through the turbulent transfer (apart from radiation) that an atmospheric model knows about the surface boundary condition. Most models use some form of Monin Obukhov Similarity (MOS), but MOS only applies to stationary homogeneous conditions. However, land surfaces tend to be heterogeneous and it is very

Table 1: The parameters and variables considered in this paper. The first column gives the parameter identifier and the second offers the explanation of the role of the parameter in the model. The next column (dist.) indicates the sampled distribution type: LN and N refer to the log-normal distribution and the normal distribution, respectively. The next columns present standard deviations of the two underlying Gaussian distributions $\sigma(1)$ and $\sigma(2)$, and whether the mean or median distribution was set to equal the unperturbed parameter value (mode).

Param./Variable	Description	dist.	$\sigma(1)$	$\sigma(2)$	mode
	TURBULENT DIFF. AND SUBGRID OROGRAPHY				
CFM_OC	transfer coefficient for momentum	LN	0.2	0.2	mean
(CFM_LA)	over ocean (over land)	LN	0.5	0.6	mean
TOFDC	coeff. in turb. orographic form drag scheme	LN	0.6	0.6	mean
HSDT	stdev. of subgrid orography	LN	0.4	0.4	mean
VDEXC_LEN	length scale for vert. mixing in stable boundary layer	LN	0.8	0.8	mean
	CONVECTION				
ENTRORG	entrainment rate	LN	0.3	0.3	median
ENTSHALP	shallow entrainment rate	LN	0.3	0.3	median
DETRPEN	detrainment rate for penetrative convection	LN	0.3	0.3	median
RPRCON	conversion coefficient cloud to rain	LN	0.4	0.4	median
CUDU	zonal conv. momentum transport	N	1.0	1.0	median
CUDV	meridional conv. mom. transport				
RTAU	adjustment time scale in CAPE closure	LN	0.6	1.0	median
	CLOUD AND LARGE-SCALE PRECIPITATION				
RAMID	RH threshold for onset of stratiform cond.	LN	0.1	0.2	median
RCLDIFF	diffusion coeff. for evap. of turb. mixing	LN	0.8	0.6	median
RCLCRIT	critical cloud water content	LN	0.8	1.0	median
RLCRITSNOW	threshold for snow autoconversion	LN	0.6	0.6	median
	RADIATION				
ZDECORR	cloud vert. decorrelation height in McICA	LN	0.6	0.4	median
ZSIGQCW	fractional stdev. of hor. distrib. of water content	LN	0.4	0.6	median
ZRADEFF	eff. radius of cloud water and ice	LN	0.6	1.0	median
ZHS_VDAERO	scale height of aerosol norm. vert. distrib.	LN	0.8	1.0	median
DELTA_AERO	optical thickness of aerosol	LN	0.6	0.6	median

difficult to characterize real terrain in terms of the surface drag that it exerts on the flow. This applies particularly to areas with orography, where all scales (resolved and subgrid) contribute to the surface drag. Although the homogeneous ocean is in many respects simpler than land, also for the ocean there are uncertainties related to lack of understanding of the interaction between turbulence and ocean waves. It is the combination of documented sensitivity of atmospheric circulation to surface drag (Sandu et al., 2015) and the uncertainty in land and ocean drag coefficients that motivates the choice to perturb surface drag related parameters. These parameters are part of the vertical diffusion scheme (Sandu et al., 2015), and the subgrid orography schemes (Lott and Miller, 1997; Beljaars et al., 2004).

(b) Convection in the IFS is represented by a bulk mass flux convection scheme (Bechtold et al., 2014). The most important and uncertain processes in the convection scheme are the mixing of the cloud with the environment, represented by the entrainment and detrainment rates, the estimation of the overall amount of convection which is related to the adjustment time scale, the autoconversion in the cloud microphysics and the convective momentum transport. The momentum transport is particularly uncertain as it depends on the organization of the convection meaning that convective momentum transport is on average downgradient (reducing the shear) but upgradient momentum transport can occur in mesoscale convective systems (Grubišić and Moncrieff, 2000).

(c) There are many uncertainties in the parametrization of cloud and precipitation microphysics, but parameters chosen here represent uncertainty across the different processes of cloud formation, cloud dissipation, rain formation and snow aggregation. The critical relative humidity parameter characterises the sub-grid heterogeneity of humidity in a grid box and defines the relative humidity at which condensation begins to form in the respective grid box. Other uncertain parameters include: the evaporation rate of water and ice due to turbulent mixing at cloud edges, which acts to dissipate the cloud; the critical cloud water content parameter which determines the onset of the warm-rain precipitation process; the critical ice water content parameter that characterises the onset of ice-to-snow aggregation and thereby represents a part of the uncertainty in the cold-phase precipitation process.

(d) The greatest source of instantaneous error in the IFS radiation scheme is the treatment of clouds and aerosols (Morcrette et al., 2008). The scheme represents the sub-grid cloud structure by using the Monte Carlo Independent Column Approximation (McICA), which constructs multiple realizations of the cloud profile from the prognostic variables cloud fraction and gridbox-mean cloud water content. Uncertainty arises due to the need to specify the vertical decorrelation length scale governing the degree to which clouds in adjacent layers overlap, and the fractional standard deviation of water content in a given model layer. Converting water content to the optical depth of a layer is parameterized as a function of effective radius, which for liquid clouds is calculated as a function of liquid water content (Martin et al., 1994), while for ice is a function of temperature. The aerosol optical depth for 5 different species is prescribed from a climatology derived from Tegen et al. (1997). The climatology varies as a function of longitude, latitude and month. To compute the 3D distribution of aerosol, concentrations are assumed to decrease exponentially with height according to a specified scale height for each species. SPP perturbs aerosol optical depth and this scale height independently.

2.2 Distribution

A number of choices need to be made to define the probability distribution from which the perturbed parameters are sampled. For describing this distribution, it is convenient to introduce a unified notation for the parameters that are perturbed. Thus, instead of referring to the parameters (TOFDC, ENTRORG, RTAU, RAMID, etc.), we will distinguish the different perturbed parameters by an integer index j ranging from 1 to K, with $K \leq 20$. Let $\widehat{\xi}_j$ denote the unperturbed value of the j-th parameter. This is the

value of the respective parameter used in the deterministic forecasts.

The perturbed parameters are referred to as ξ_j. Except for the convective momentum transport, all perturbed parameters sample a log-normal distribution:

$$\xi_j = \exp(\Psi_j)\,\widehat{\xi}_j, \quad \Psi_j \sim \mathcal{N}(\mu_j, \sigma_j^2) \tag{1}$$

Here, the perturbations Ψ_j sample a Gaussian distribution with a mean μ_j and a standard deviation σ_j, both determined individually for each perturbed parameter j. At present, two options are considered for determining the mean μ_j: (a) $\mu_j = -\frac{1}{2}\sigma_j^2$ and (b) $\mu_j = 0$. Option (a) implies that the mean of the distribution of the perturbed parameter ξ_j is equal to the unperturbed value $\widehat{\xi}_j$ while option (b) implies that the median of the distribution is equal to the unperturbed value. Table 1 summarizes the distribution standard deviations σ_j for each of the 20 parameters. Two sets of standard deviations $\sigma(1)$ and $\sigma(2)$ are given, which are used for testing the sensitivity of the SPP scheme in Section 4. The standard deviation values are based on expert estimates about the uncertainties of each parameter and variable. The transfer coefficients for momentum, CFM_OC and CFM_LA, have individually defined standard deviations over land and ocean areas, but they use the same standardised perturbation $\sigma_j^{-1}(\Psi_j - \mu_j)$.

The choice of a log-normal distribution stems from practical reasons: both smaller and larger perturbations ($\exp(\Psi_j)$) to the default parameter values ensure perturbed parameter values ξ_j always retain their original sign without any need to adjust or clip the perturbation values. An exception to this are the perturbations applied to the zonal and meridional components of the convective momentum transport, which are perturbed with multiplicative noise drawn from a bi-variate normal distribution with expected value of $(1,1)$. For reasons of numerical stability, the normal distribution is truncated such that the Euclidian norm of the perturbations does not exceed a value of 3. Note, that this approach permits perturbations leading to an upgradient momentum transport while the unperturbed momentum transport is down-gradient.

The perturbations $\exp(\Psi_j)$ vary in space and time using independent patterns for each parameter j, and for each ensemble member. Figures 1 and 2 illustrate the structure of the perturbation patterns $\exp(\Psi_j)$ from a randomly chosen ensemble member and model time step. Perturbation patterns for two parameters with standard deviations of 0.3 (Fig 1) and 0.8 (Fig 2) are shown. The corresponding distributions of the perturbation values are shown next to the geographically mapped perturbations (with matching bin colours). When drawing perturbations from a narrow distribution with standard deviation of 0.3 (Fig 1), 80% of the perturbations are confined within a rather conservative range of 0.67–1.5, and less than 1% of the perturbations are outside 0.33–3.0. When increasing the distribution standard deviation to 0.8 (Fig 2), the amount of grid points falling into the perturbation range of 0.67–1.5 drops to 45%. However, 87% of the perturbations still lie within the range 0.33–3.0. Note that the distributions diagnosed from instantaneous realisations will not exactly agree with the log-normal distribution they are sampling.

In order to maintain physical consistency within a grid-column, the same horizontal pattern is applied to all model levels. Temporal and spatial correlations are obtained through a first-order auto-regressive (AR(1)) process in spectral space (Berner et al., 2009; Palmer et al., 2009): Let $\hat{r}_{mn}(t)$ denote the spectral coefficients of the pattern Ψ at time t. Subscripts m and n refer to the zonal and total wavenumber, respectively. The coefficients evolve from time t to the next model timestep $t + \Delta t$ according to

$$\hat{r}_{mn}(t+\Delta t) = \phi\,\hat{r}_{mn}(t) + s_n \varepsilon_{mn}(t),$$

with variance depending on total wavenumber according to

$$s_n^2 = F_0^2 \exp\left(-L_\Psi^2 R_E^{-2} n(n+1)\right).$$

For a correlation timescale τ_Ψ, the correlation of the coefficients $\hat{r}_{mn}$ over a single timestep Δt is given by $\phi = \exp(-\Delta t/\tau_\Psi)$. Real and imaginary parts of $\varepsilon_{mn}(t)$ are independent random variables that sample a Gaussian distribution with unit variance and mean zero. The $\varepsilon_{mn}(t)$ are also independent for each m,n,t and each ensemble member. The ratio of the spatial correlation length scale L_Ψ to the Earth's radius R_E determines how quickly the spectral variance s_n^2 decays with wavenumber n. The pattern resulting from this representation has homogeneous isotropic correlations on the sphere that correspond to a Gaussian spatial correlation function (Weaver and Courtier, 2001). The random pattern in grid point space is obtained via a spectral transform of $\hat{r}$ at each model timestep. The factor F_0^2 is set to a value that yields the desired grid point variance σ_Ψ^2 of the random pattern.

The random patterns applied to different model parameters i and j are independent. This is achieved through using a different seed for the pseudo random number generator that supplies the ε_{mn} for each pattern. The initial implementation is parsimonious and uses a single correlation time scale $\tau_\Psi = 72\,\text{h}$ and a single correlation length scale $L_\Psi = 2000\,\text{km}$ for all perturbed parameters.

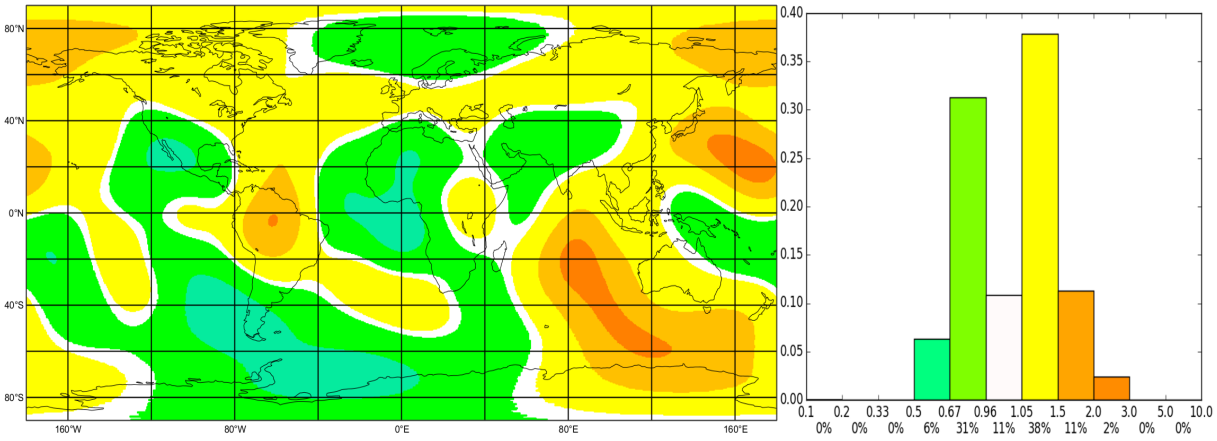

Figure 1: Lognormally distributed perturbation patterns with standard deviation of $\sigma = 0.3$ and median of the distribution equal to 1. Left shows a global map of the 2D-pattern and right a histogram of the relative number of grid points falling into bins indicated by the values on the x-axis. The bin colours match those of the 2D-pattern.

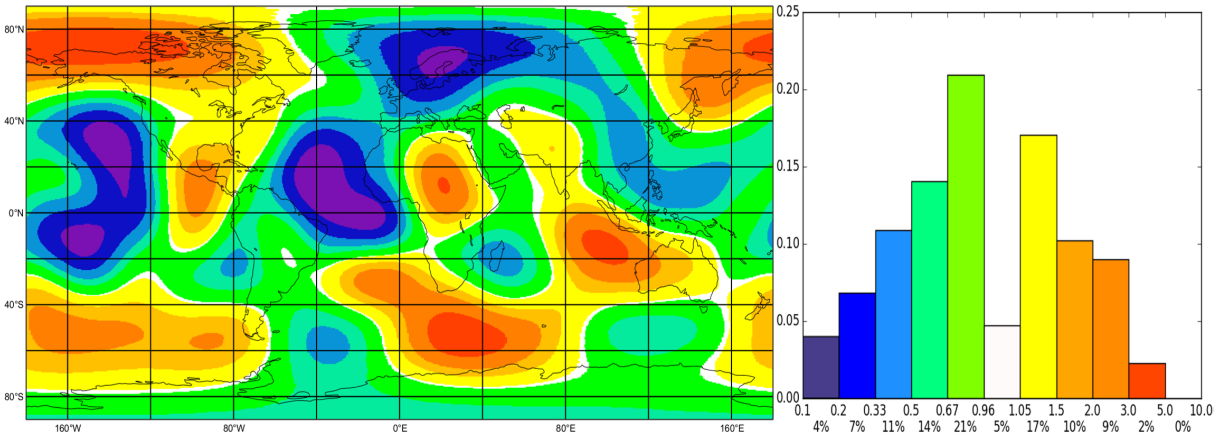

Figure 2: As in Fig 1, but with standard deviation of $\sigma = 0.8$.

A large set of standard deviations was screened during the initial testing. As expected, allowing for larger parameter perturbations increased the ensemble spread, but this usually came with a price of increasing the ensemble mean Root Mean Squared Error (RMSE). Two sets of standard deviations are presented:

$\sigma(2)$ was initially selected to increase the ensemble spread and improve the ensemble mean RMSE in the medium-range. Upon experimentation, further tuning yielded an improved model climate in long integrations from $\sigma(1)$. Selecting whether the unperturbed parameter value equals to the mean or median of the distribution also had a noticeable effect on the ensemble scores. Selecting the median mode led to overall better probabilistic scores for most parameters, but for some, it had a more mixed signal (improvements in most of the model variables, but notable deteriorations in one or two). For this reason, the boundary layer perturbations were set to the mean mode (Table 1)[2]. Extensive experimentation was also conducted to explore the impacts of the choice of horizontal and temporal correlations of the perturbation patterns. Increasing the horizontal correlation lengths leads to larger areas sharing and synchronizing information about the perturbation field, while longer temporal correlations mean the perturbation patterns are more persistent. The impacts of both changing the distribution standard deviations and altering the correlation scales are shown in Section 4, and further discussed in Section 8.

3 Numerical experiments

The SPP experimentation was conducted with model version CY41R1 of the IFS (operational from May 2015 to March 2016). Until March 2016, the operational ensemble consisted of 51 ensemble members run at T_L639 resolution (∼32 km grid spacing at equator) with 91 vertical levels. The forecasts are run at T_L639 resolution for the first 10 days, and at T_L319 resolution (∼65 km) for the following 5 days (i.e. from day 10 to day 15). Timesteps of 20 min and 45 min are applied to both the dynamical core and the model physics at T_L639 and T_L319 resolutions, respectively. An exception to this is radiation, which is updated in 3 h intervals. A combination of singular vectors (see e.g. Leutbecher and Lang, 2014) and ensemble data-assimiliation (Buizza et al., 2008) is used to account for the initial state uncertainties. In addition, model uncertainties are represented through the SPPT and SKEB schemes. Here, three major alterations were made to the operational ENS configuration: (i) a horizontal resolution of T_L399 (∼50 km) and a timestep of 30 min is applied throughout the 15 day forecast range, (ii) the ensemble was run with 21 members only, and (iii) the operational stochastic physics schemes (SPPT and SKEB) were disabled. Changes (i) and (ii) were done in order to reduce the computational resources needed for the experimentation, and (iii) in order to study the SPP scheme in isolation.

Five SPP experiments (see Table 3) were conducted with this system. SPP-L acts as a "reference" and uses a distribution with standard deviations $\sigma(1)$ (Table 1), with horizontal correlation scales of 2000 km and with a temporal correlation of 72 h. The impact of changing the parameter distribution is studied by halving $\sigma(1)$ values for all parameters (SPP-L(3)), and by using a slightly altered parameter distribution (SPP-L(2)) with standard deviations $\sigma(2)$ (Table 1). Furthermore, the effect of altering the horizontal and temporal correlation scales is explored by changing them to smaller scales/higher frequencies (SPP-S), and to globally constant/frozen-in-time (SPP-Inf). We perform separate tests for changing the distribution widths and the horizontal and temporal correlations scales in order to explore the importance of both configuration choices.

Model runs with only initial state perturbations active (IP_only) and initial state perturbations and SPPT active (SPPT) were done in order to benchmark the probabilistic skill of the SPP experiments. In addition, the SPPT and SPP-L experiments were repeated without initial state perturbations (SPPT^{noi}, $\text{SPP-L}^{\text{noi}}$) in order to study the differences in the perturbed tendencies produced by the two stochastic schemes.

Experiments IP_only, SPPT and SPP-L (SPP-L(2), SPP-L(3), SPP-S and SPP-Inf) cover one year starting from Dec 2013 and ending in Nov 2014; an ensemble is launched every 8 (16) days, resulting in 46 (23)

[2] Choosing the mean mode leads to a distribution shifted towards lower values than given by the median mode.

start dates. The experiments without initial state perturbations (SPPT^{noi}, SPP-L^{noi}) cover the boreal winter months Dec 2013 - Feb 2014, with an ensemble run every 16 days, totalling a sample of 6 ensemble start dates.

Table 2: Experiments conducted for this study. N indicates the number of start dates. L_Ψ and τ_Ψ indicate the horizontal and temporal correlation scales, respectively, and additional information about the experiment is given under Notes.

Name	N	L_Ψ	τ_Ψ	Notes
IP_only	46			Only initial state perturbations
SPPT	46	3 scales †	3 scales †	SPPT and initial state pert.
SPP-L	46	2000 km	72 h	$\sigma(1)$
SPP-L(2)	23	2000 km	72 h	$\sigma(2)$
SPP-L(3)	23	2000 km	72 h	stdev. $0.5\sigma(1)$
SPP-S	23	500 km	6 h	$\sigma(1)$
SPP-Inf	23	∞	∞	$\sigma(1)$
SPPT^{noi}	6	3 scales †	3 scales †	Copy of SPPT, no ini. state pert.
SPP-L^{noi}	6	2000 km	72 h	Copy of SPP-L, no ini. state pert.

†SPPT has a pattern composed of three independent random fields with horizontal correlation scales of 500, 1000 and 2000 km. The temporal correlation scales are 6 h, 72 h and 30 d. The standard deviations are 0.52, 0.18 and 0.06, respectively (Shutts et al., 2011).

4 Sensitivity to configuration

The different SPP configurations (Table 3) were first used to evaluate the sensitivity of the scheme to various aspects. All SPP configurations generate more spread than IP_only, but large differences exist among the experiments (Fig 3). Both changes in the perturbation distributions and in the correlation scales result in a similar range of variations in the ensemble spread. Experiment SPP-L produces the most spread among the 3 experiments using different correlation scales. Both decreasing (SPP-S) and increasing (SPP-Inf) the correlation scales results in decreasing spread. SPP-S produces a similar amount of ensemble spread to SPP-Inf in short forecast lead times. After forecast day 3, SPP-S has the least amount of spread among the 3 experiments with different correlation scales. A possible explanation for these different responses is offered in Section 8.

Halving the standard deviations $\sigma(1)$ of the distributions also leads to a large drop in the ensemble spread: SPP-L(3) produces the least ensemble spread of all SPP experiments in the first 6 forecast days. Thereafter, the amount of spread is similar to that of SPP-S. However, small alterations to the distributions (SPP-L(2)) show slightly larger spread than from SPP-L. Similar results are also observed in the extra-tropics (spread for Northern extra-tropics shown in Fig 4), with SPP-L(2) generating the most spread, closely followed by SPP-L.

The correlation scales of 2000 km and 72 h produce, in addition to the most spread, the best ensemble mean RMSE scores among the experiments with different correlation scales (Figs 5-6). They were therefore set as the "default" correlation scales. Using either $\sigma(1)$ or $\sigma(2)$ standard deviations results in very similar ensemble skill scores. $\sigma(1)$ was chosen as the "default" configuration based on the additional benefit to the model climate. We note that observed behaviour of both the ensemble spread and the mean RMSE discussed here are consistent with results for other model levels and variables.

In order to study further the ENS sensitivity to the different SPP configurations we will next apply

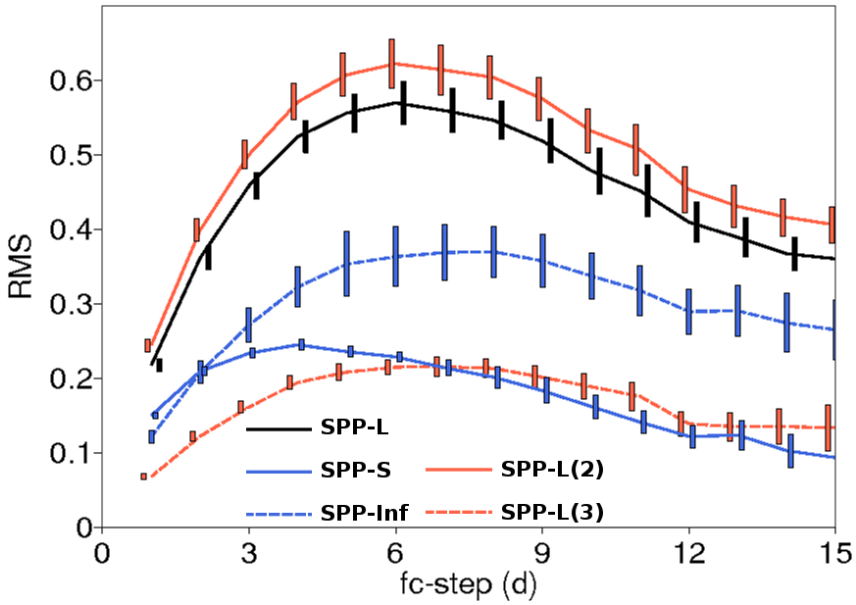

Figure 3: Ensemble spread difference measured against IP_only experiment (zero line). Positive values indicate a larger spread than in IP_only. u-component of wind (in m/s) at 850 hPa in Tropics up to day 15 lead times for different SPP configurations: SPP-L (black), SPP-L(2) (red), SPP-L(3) (dashed red), SPP-S (blue), and SPP-Inf (dashed blue). The vertical bars indicate the 95% confidence inverval. Statistics are computed from 23 start dates between Dec 2013 and Nov 2014.

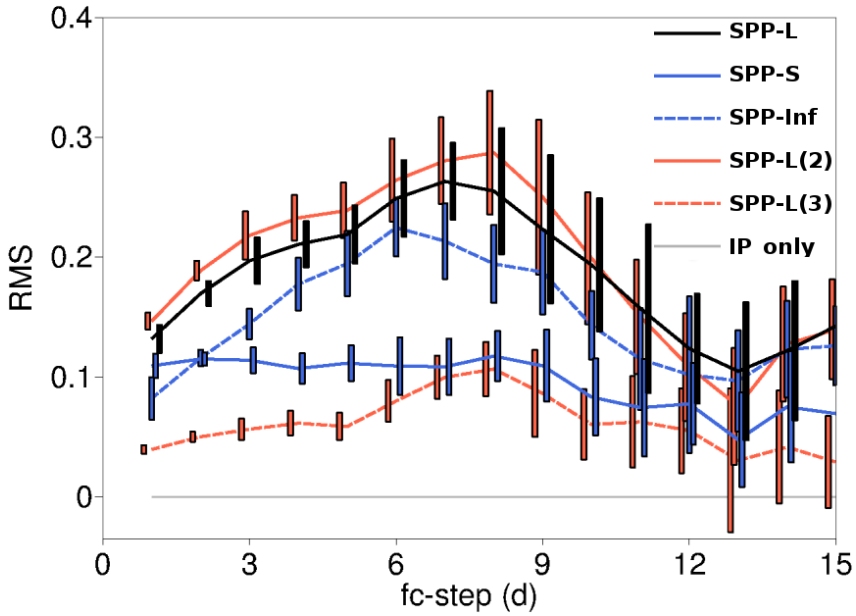

Figure 4: As in Fig 3, but for Northern extra-tropics.

two metrics to a set of surface variables at forecast day 5. In Figure 7 two metrics are shown for the ensemble mean (em) values of 11 surface variables: (i) local *absolute difference* from unperturbed control forecast (cf), calculated as $\sum|em-cf|/\sum cf*100$; and (ii) *relative change* of area means, calculated as $(\sum em/\sum cf-1)*100$. In both equations the summations go over all grid points, and apply areal weights. The subtraction in (ii) is done to centre the fraction around 0 rather than 1. The first metric

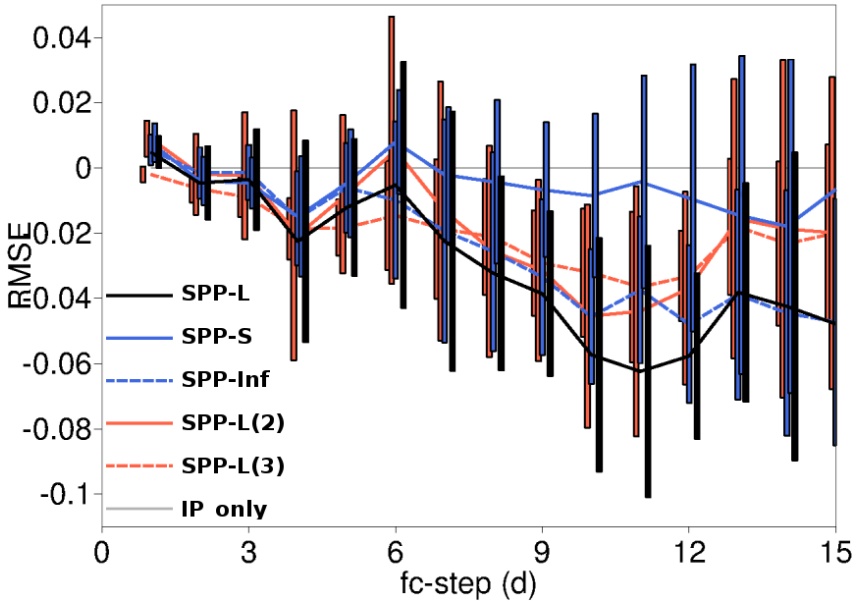

Figure 5: Ensemble mean RMSE difference measured against IP_only experiment (zero line). Negative values indicate an improved RMSE compared to IP_only. u-component of wind (in m/s) at 850 hPa in Tropics up to day 15 lead times for different SPP configurations: SPP-L (black), SPP-L(2) (red), SPP-L(3) (dashed red), SPP-S (blue), and SPP-Inf (dashed blue). The vertical bars indicate the 95% confidence inverval. Statistics are computed from 23 start dates between Dec 2013 and Nov 2014.

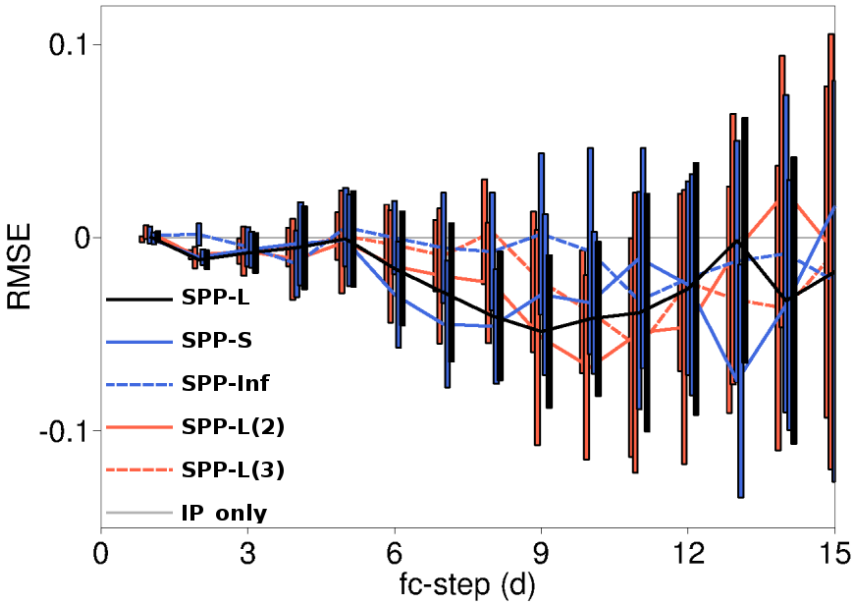

Figure 6: As in Fig 5, but for Northern extra-tropics.

quantifies the amount of change in the ensemble mean w.r.t. cf; the second indicates whether the bulk ensemble mean quantities are increased or decreased w.r.t cf.

The ensemble means of all SPP configurations, as well as that from SPPT, have large *absolute differences* to the cf in the cloud cover and precipition quantities (up to 28% for MCC). However, when comparing the *relative changes* all the experiments indicate a much more modest change in the bulk values. Thus

the large absolute differences originate mostly from relocation of cloudy areas and precipitation. The largest variation in the absolute differences among the SPP configurations (above 2 percentage points) is observed in MCC, HCC, LSP and CP. The absolute differences in the other variables are more tightly distributed. Most notable differences in *relative change* within the SPP configurations are found in MCC, LCC, TTR and TSR, in which the different configurations show a mix of both an increase and a decrease of the bulk quantities. Large differences are also observable in CP and LSP, although for these variables all the configurations change the ensemble mean towards the same direction w.r.t. cf. The SPP-L configuration shows a rather conservative behaviour among all the SPP configurations: it is never the largest in absolute difference or relative change. The SPPT experiment indicates quite a similar ensemble mean impact in relative differences to the SPP configurations. However, SPPT produces the largest relative changes for all variables except LSP, LCC and TSR. Especially large differences between SPPT and the SPP configurations are found in SF, HCC and TCWV. The distinct impact of the two schemes on the forecasts is clearly seen in SF and TCWV, for these variables all the SPP configurations produce an increase and SPPT a decrease in the ensemble mean states.

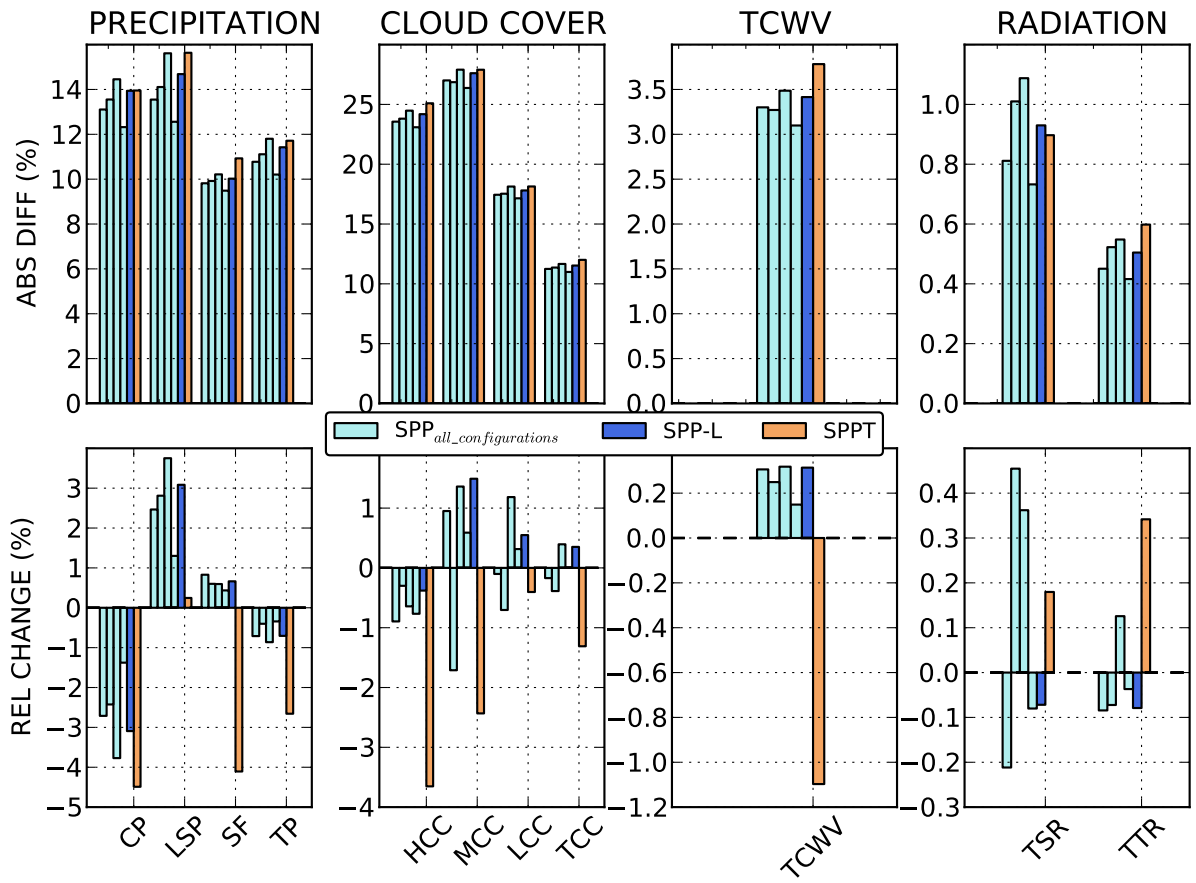

Figure 7: Global integrated values for local absolute difference (upper row, see text for details) and relative change of area means (lower row, see text for details) between ensemble mean and unperturbed forecast. SPPT (orange), SPP-L (blue), and the different SPP configurations (cyan). The variables shown: 120 h accumulated convective (CP), large scale (LSP) and total (TP) precipitation, 120 h accumulated snow fall (SF); high (HCC), medium (MCC), low level (LCC) and total cloud cover (TCC); total column water vapour (TCWV); top-of-the-atmosphere net solar (TSR) and thermal radiation (TTR) fluxes averaged over 120 h. All units in %. Mean of 6 start dates between Dec 2013 and Feb 2014.

5 Perturbation structure

We will next focus on identifying any structural differences in the short range model forecasts produced by the SPP and the SPPT schemes. In order to better understand the model responses we disable the initial state perturbations, i.e. any differences between the SPP-L^{noi} and SPPT^{noi} experiments originate from representing the model uncertainties in different ways.

Figure 8 presents vertical profiles of statistics for temperature tendencies produced by the model physical parametrisations, and accumulated over 6 model time steps, i.e. 3 h of simulations. The top row presents the (Pearson's) correlation of the ensemble mean (em) with control forecast (cf), calculated for a single

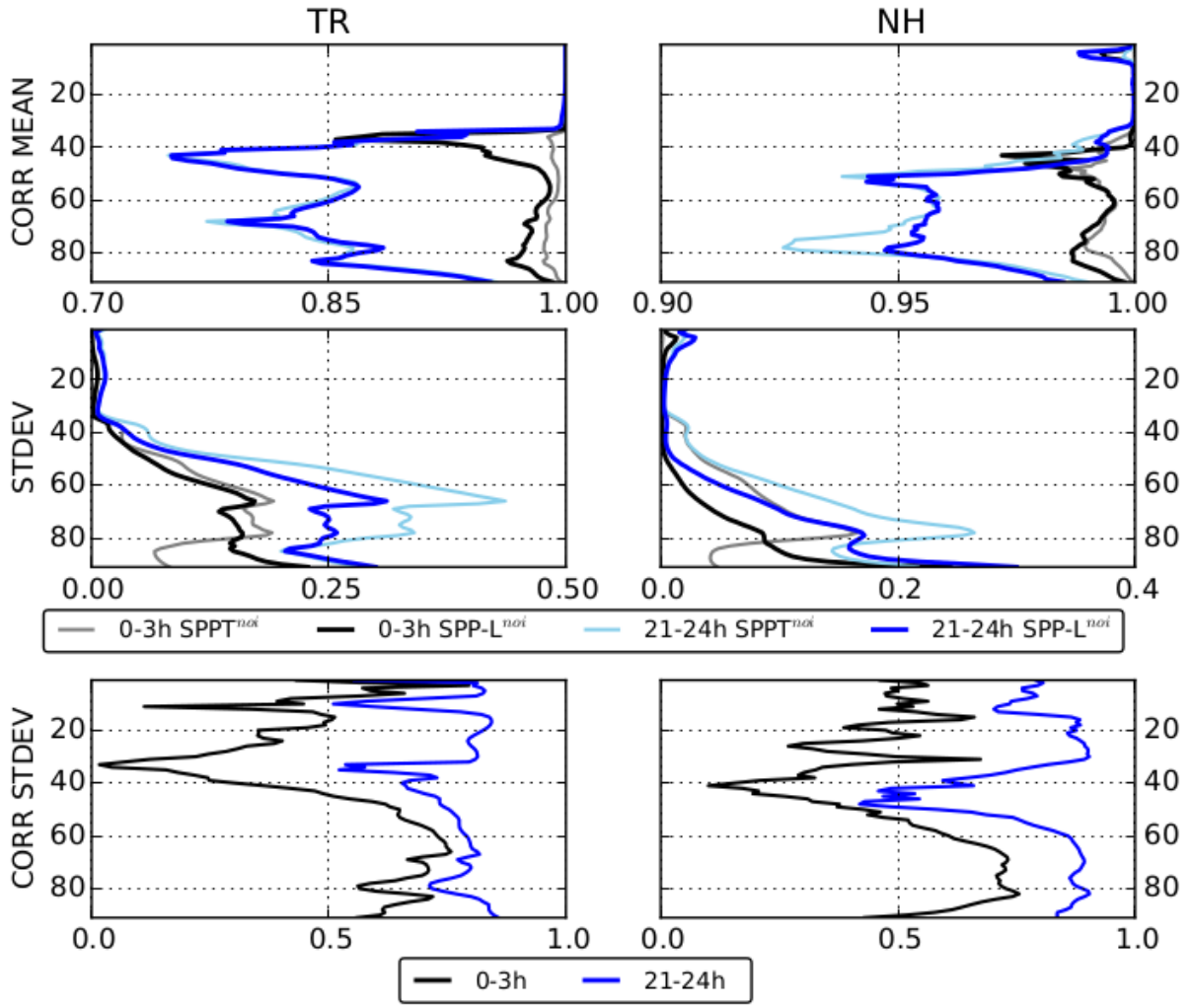

Figure 8: Vertical profiles of temperature tendencies for (top) ensemble mean correlation with control forecast, (middle) ensemble standard deviation, and (bottom) correlation of SPP-L^{noi} ensemble standard deviation with that of SPPT^{noi} (see text for details). The profiles are shown for 3 h cumulated tendencies for forecast lengths of 0–3 h for SPP-L^{noi} (black) and SPPT^{noi} (grey), and of 21–24 h for SPP-L^{noi} (blue) and SPPT^{noi} (cyan). Left (right) column presents tropical (northern extra-tropical) scores. Model levels are presented on y-axis (as a rough guidance to the model levels: lvl 81 = 925 hPa, lvl 64 = 500 hPa, and lvl 30 = 50 hPa). Units for top and bottom are 1 (dimensionsless) and for middle K/(3 h). Statistics from 6 start dates between Dec 2013 and Feb 2014.

model level as

$$\mathrm{CORR\_M} = \frac{\sum_{x,t}(em_{x,t} - \overline{em})(cf_{x,t} - \overline{cf})}{\sqrt{\sum_{x,t}(em_{x,t} - \overline{em})^2}\sqrt{\sum_{x,t}(cf_{x,t} - \overline{cf})^2}} \tag{2}$$

Here $\overline{em}$ ($\overline{cf}$) is the area-weighted mean of em (cf) tendencies over all grid points and time steps (6 dates between Dec 2013 and Feb 2014) on a single model level. Similarly, the sums ($\sum_{x,t}$) go over all the grid points (x), and time steps (t). The cosine-latitude area-weighting is applied consistently in all summations. The correlation indicates how similar to the unperturbed model state the tendency of the ensemble mean is. Lower values of the tendency correlations indicate that the ensemble members differ systematically from the unperturbed state. We want to emphasize that the tendencies shown are the contribution of the physical parametrisations to the total tendencies.

In the first 3 h of the forecast, the correlation to the cf is overall high for both of the experiments. Furthermore, the correlations are higher for NH than TR (note the different x-axes). Some differences between the two experiments emerge already during this time: In TR, the ensemble mean (em) is less correlated with the cf for SPP-L$^{\mathrm{noi}}$ than for SPPT$^{\mathrm{noi}}$, especially in the lower stratosphere (levels 36–46). In NH, the two profiles look more alike except in the boundary layer (level >80), in the lower stratosphere (levels 41–44), and near the top of the model (levels 1–10), where em of SPP-L$^{\mathrm{noi}}$ is less correlated with the cf. By lead times 21–24 h, a general decrease of correlation of em with cf is observed for both experiments (as expected due to perturbation growth). In TR, the two experiments exhibit very similar correlations of ensemble mean to cf by 21–24 h. This similarity in the globally averaged correlations does not equate to similar responses from SPPT$^{\mathrm{noi}}$ and SPP-L$^{\mathrm{noi}}$ across the globe: from plotted maps (not shown) it is observed that the differences between em and cf in each experiment are of a similar magnitude to the differences between em of SPPT$^{\mathrm{noi}}$ and em of SPP-L$^{\mathrm{noi}}$. In NH, SPPT$^{\mathrm{noi}}$ exhibits greater decorrelation from cf in model levels 65–80. Near the top of the model (levels 1–10), em of SPP-L$^{\mathrm{noi}}$ shows greater decorrelation from cf. This is due to the SPP scheme perturbing gravity waves excited by orography, and thus also altering how these gravity waves break in the stratosphere.

The mean ensemble standard deviations (stdev.) are shown in the middle row of Fig 8. Area-weighting is applied to the standard deviations and the values are furthermore averaged over the 6 dates. This is normally interpreted as how much the tendencies of the ensemble members differ from each other. In this case the larger the standard deviation is the greater the difference between the temperature tendencies of the ensemble members.

The stdevs. of the tendencies in the free-troposphere tend to be greater in TR than in NH (note the different x-axes) for both schemes, and for both lead times. In the 0–3 h lead times, SPP-L$^{\mathrm{noi}}$ displays larger stdev. than SPPT$^{\mathrm{noi}}$ within the boundary layer (level >80) for both TR and NH. However, in NH above the boundary layer, SPPT$^{\mathrm{noi}}$ shows larger stdev. At forecast range 21–24 h, the larger spread for SPPT$^{\mathrm{noi}}$ above the boundary layer becomes more visible, and is observed in TR as well. For these lead times the larger stdev. of SPP-L$^{\mathrm{noi}}$ in the boundary layer is still visible in NH. Interestingly, the difference in the boundary layer between SPP-L$^{\mathrm{noi}}$ and SPPT$^{\mathrm{noi}}$ experiments has become very small in TR. Although the SPPT$^{\mathrm{noi}}$ experiment has no perturbations in the boundary layer, the differences in the free-troposphere have by this lead time forced the boundary layer differently, thus generating differences in the lowest model levels also. Both experiments exhibit an increased stdev. in the upmost 10 model levels, indicating that both ensembles generate differences in gravity wave breaking. For SPPT the differences originate from non-orographic gravity wave sources, for SPP from both non-orographic and orographic.

Finally, the bottom row presents the correlations between SPP-L^{noi} and SPPT^{noi} stdevs., calculated from:

$$\mathrm{CORR\_S} = \frac{\sum_{x,t}(s^{\mathrm{SPP}}_{x,t} - \overline{s^{\mathrm{SPP}}})(s^{\mathrm{SPPT}}_{x,t} - \overline{s^{\mathrm{SPPT}}})}{\sqrt{\sum_{x,t}(s^{\mathrm{SPP}}_{x,t} - \overline{s^{\mathrm{SPP}}})^2}\sqrt{\sum_{x,t}(s^{\mathrm{SPPT}}_{x,t} - \overline{s^{\mathrm{SPPT}}})^2}}$$

Here, s^{SPP} is the standard deviation of the SPP-L^{noi} ensemble and s^{SPPT} the standard deviation of the SPPT^{noi} ensemble. This quantifies the differences in the location and timing of ensemble stdev. between SPP-L^{noi} and SPPT^{noi} experiments. The larger the correlation is, the more similar in space and time the stdev. of the two experiments. The stdev. correlation is shown for 0–3 h (black) and for 21–24 h forecast lead times (blue).

The correlations between the stdevs. of the two experiments tend to be small in both TR and NH, indicating that the perturbations from the two schemes yield physical structures that differ spatially and/or temporally. The two schemes appear to be less correlated at lead times 0–3 h than at the later times 21–24 h. This indicates that the two schemes introduce quite different perturbation structures, which yield low correlations at the earliest lead times; but by 21–24 h, regional variations in the flow regime modulate perturbation growth and may (partly) cause the tendency spread to become more similar for the two stochastic schemes.

Next, differences between the SPP and SPPT schemes are studied through spread in 11 surface variables. In Figure 9, the average global spread for the 6 start dates is presented for SPP-L^{noi} (blue) and SPPT^{noi} (orange) experiments at 3 h forecast lead time. Large differences in the spread are present for all but TCWV. SPP-L^{noi} spread in precipitation is almost twice that in SPPT^{noi}. For cloud cover, SPP-L^{noi} spread is also $\sim$1.5 times higher than in SPPT^{noi}. The larger TCWV spread for SPPT^{noi} originates from the extra-tropics (not shown). In TR, SPP-L^{noi} spread is actually larger than in SPPT^{noi}. The absence of radiation spread in SPPT^{noi} can be explained through the IFS time-stepping method: Radiation fluxes are updated every 3 h and are moreover calculated after the model state is written out. Because the SPPT perturbations are applied after the first radiation update (and because no initial state perturbations are present), all the ensemble members in SPPT^{noi} have precisely the same radiation fields when the model state is written out at the 3 h forecast lead time. The SPP perturbations are, however, applied during the first radiation update, resulting in each ensemble member having different radiative fluxes (which are applied throughout the 3 h radiation window) when the model state is written out.

Figure 10 presents spread for the same variables, but at forecast lead time of 24 h. The SPPT^{noi} spread has by this time grown w.r.t. SPP-L^{noi}. The spread differences in precipitation variables are much smaller, although SPP-L^{noi} retains slightly greater spread in CP, LSP and TP than SPPT^{noi}. In the cloud cover variables, SPP-L^{noi} has now more spread only for LCC. The larger spread in SPPT^{noi} originates from the extra-tropics. The differences in TCWV spread have also grown larger, with SPPT^{noi} spread being $\sim$1.5 times greater than SPP-L^{noi} spread. Finally, spread in radiation fluxes is (still) larger for SPP-L^{noi}. An interesting aspect in Fig 10 is that even though SPP-L^{noi} produces less variations than SPPT$^{noi}_{L}$ in the cloud cover, it still generates more spread than SPPT$^{noi}_{L}$ in the net radiative fluxes at the top-of-the-atmosphere.

6 Impact on ensemble forecasts

Next, we include again initial state perturbations in the ensembles and compare the probabilistic skill of the SPP-L experiment with the SPPT and IP_only benchmark experiments. Both SPP-L and SPPT

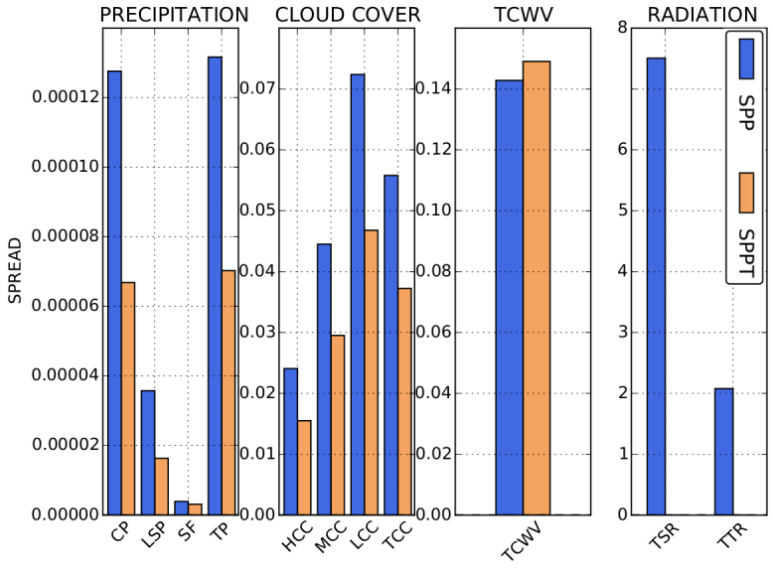

Figure 9: Global integrated values of ensemble stdev. at forecast lead time of 3 h for SPPT^{noi} (orange) and SPP-L^{noi} (blue) experiments. The variables shown: 3 h accumulated convective (CP), large scale (LSP) and total (TP) precipitation (in m), 3 h accumulated snow fall (SF) (in m of water equivalent); high (HCC), medium (MCC), low level (LCC) and total cloud cover (TCC) (in fraction); total column water vapour (TCWV) (in kg/m^{-2}); top-of-the-atmosphere net solar (TSR) and thermal radiation (TTR) fluxes averaged over 3 h (in W/m^{-2}). Mean of 6 start dates between Dec 2013 and Feb 2014.

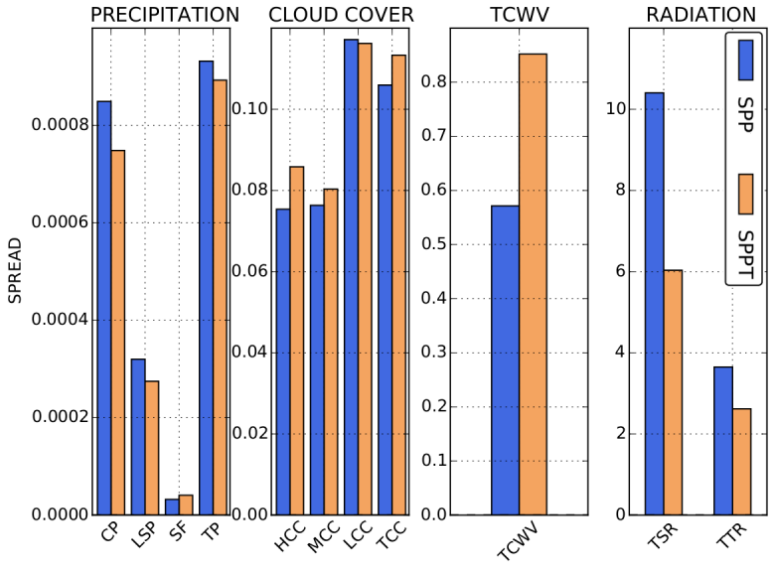

Figure 10: As in Figure 9, but for forecast lead time of 24 h. The precipitation amounts (radiative fluxes) are accumulated (averaged) over 24 h.

experiments exhibit significantly better Continuous Ranked Probability Scores (CRPS; calculated against operational IFS analyses) than IP_only throughout the forecast window (Fig 11). Furthermore, beyond day 1 forecasts, the SPPT scores are better than those of SPP-L. The SPP-L CRPS are nevertheless closer to the values of SPPT than to IP_only; the CRPS improvement over IP_only at forecast day 5 is 9% for SPPT and 6% for SPP-L (the evolution of the CRPS itself is shown in supplementary Fig S1).

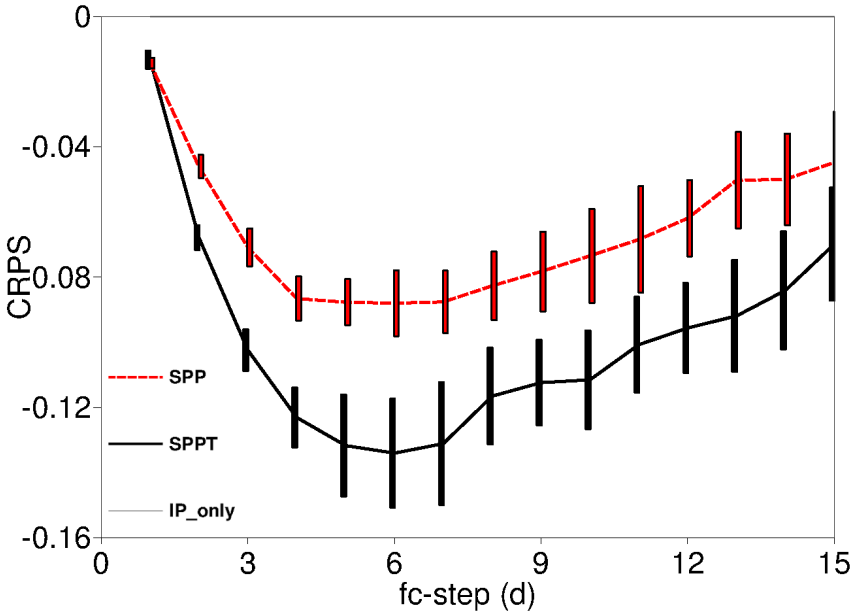

Figure 11: Differences in Continuous Ranked Probability Scores (CRPS) for u-component of wind (in m/s) at 850 hPa in Tropics for forecast lead times up to day 15. Negative values indicate an improved skill compared to IP_only. SPP-L (red dashed), SPPT (black), IP_only (zero line). The vertical bars indicate the 95% confidence inverval. Statistics are computed from 46 start dates between Dec 2013 and Nov 2014.

Both SPP-L and SPPT also improve the CRPS w.r.t. IP_only for the u-wind component in the Nothern extra-tropics (Fig 12). This holds at 95% significance level up to forecast day 9 for SPP-L, and up to forecast day 11 for SPPT. SPPT scores are significantly better than those of SPP-L for shorter than 5 day lead times. Beyond day 5, the two experiments overlap in the 95% confidence interval. The CRPS at forecast day 5 is improved by approximately 2% for both experiments when compared against IP_only benchmark (supplementary Fig S2).

Figure 13 extends the comparison of probabilistic scores to a set of model variables used typically to assess ensemble skill. At forecast day 1, compared to IP_only, SPP-L scores are significantly better for all the variables, whereas changes due to SPPT for Z500 are neutral or worse. In general, the SPP-L and SPPT experiments have quite similar CRPS, with the largest differences appearing in Z500 in SH, U850 in NH and T850 in TR. At forecast day 3, the improvements in scores due to both schemes compared to IP_only have increased. Changes due to SPP-L are still significantly better for all the variables, and from SPPT, changes for Z500 are now neutral or better. In relative terms, SPPT is now significantly better than SPP-L for Z500 in NH, U850 in NH and TR, and T850 in NH. By forecast day 5, SPPT performs best for all variables and regions (not shown).

SPP-L and SPPT experiments are next verified against surface observations. The Continuous Ranked Probability Skill Score (CRPSS) for 2 m temperature in the extra-tropics from verification against SYNOP is presented in Fig 14. Both SPP-L and SPPT exhibit greater skill than IP_only throughout the forecast range. Furthermore, SPP-L shows the greatest skill in the first couple of forecast days, after which the SPP-L and SPPT experiments show a similar skill.

Figure 15 presents CRPSS values for 24 h accumulated precipitation verified against SYNOP observations. The two experiments again show more skill than IP_only for the whole forecast length. SPP-L and SPPT exhibit very similar skill at day 1, with SPPT showing slightly greater skill thereafter.

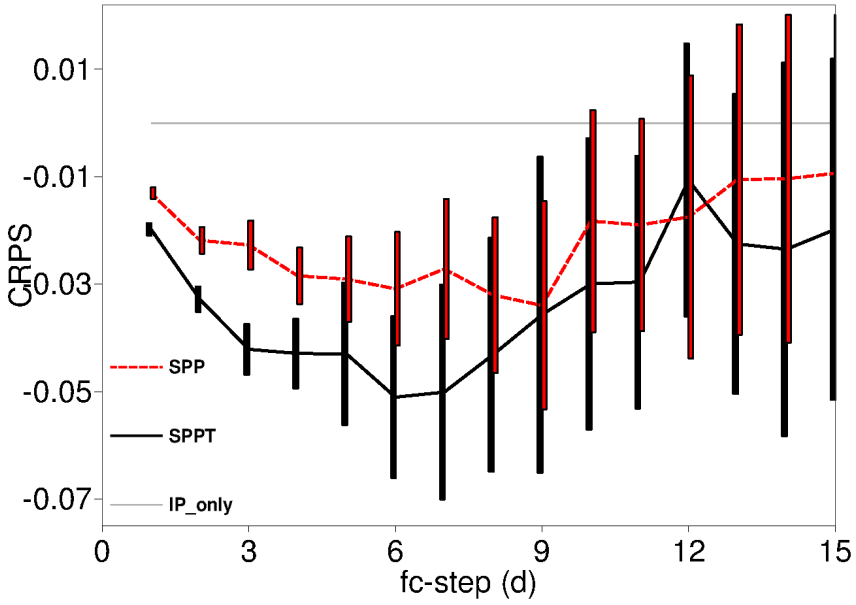

Figure 12: As in Fig 11, but for Northern extra-tropics.

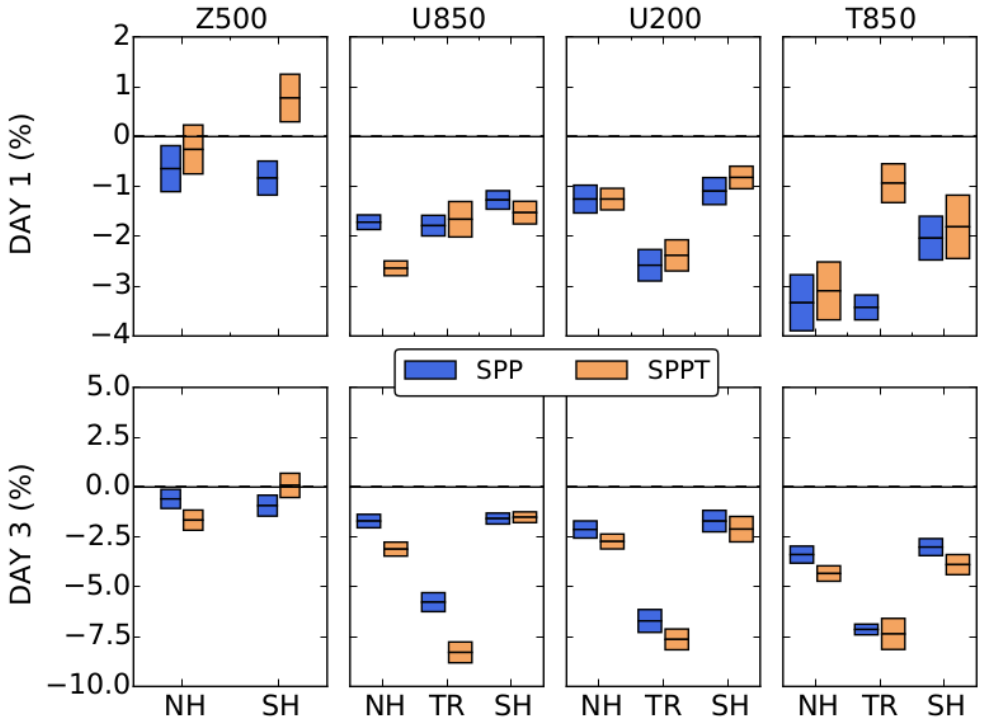

Figure 13: CRPS difference (%) from IP_only for SPP-L and SPPT experiments at forecast lead times of 1 (top) and 3 days (bottom). Scores are shown for geopotential height at 500 hPa (Z500), u-component of wind at 850 and 200 hPa (U850, U200), and temperature at 850 hPa (T850). Areas given: Tropics (TR), Northern extra-tropics (NH), and southern extra-tropics (SH). Negative (positive) values indicate improved (degrated) skill. The vertical bars represent the 95% confidence interval of the mean. Statistics are computed from 46 start dates between Dec 2013 and Nov 2014.

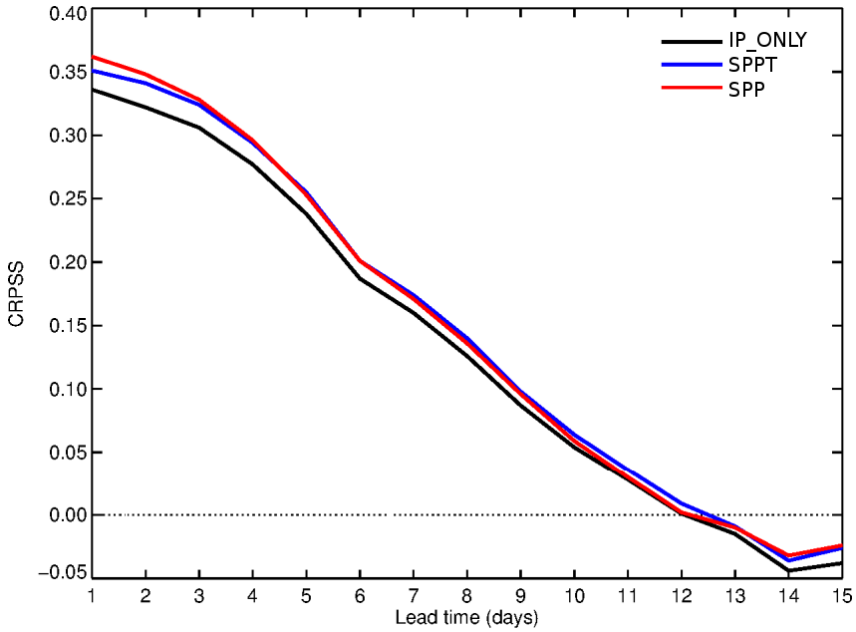

Figure 14: Continuous Ranked Probability Skill Scores (CRPSS) for extra-tropical 2 m temperatures up to forecast lead time of 15 days. PPL (red), SPPT (blue) and IP_only (black). Mean skill of 46 start dates between Dec 2013 and Nov 2014.

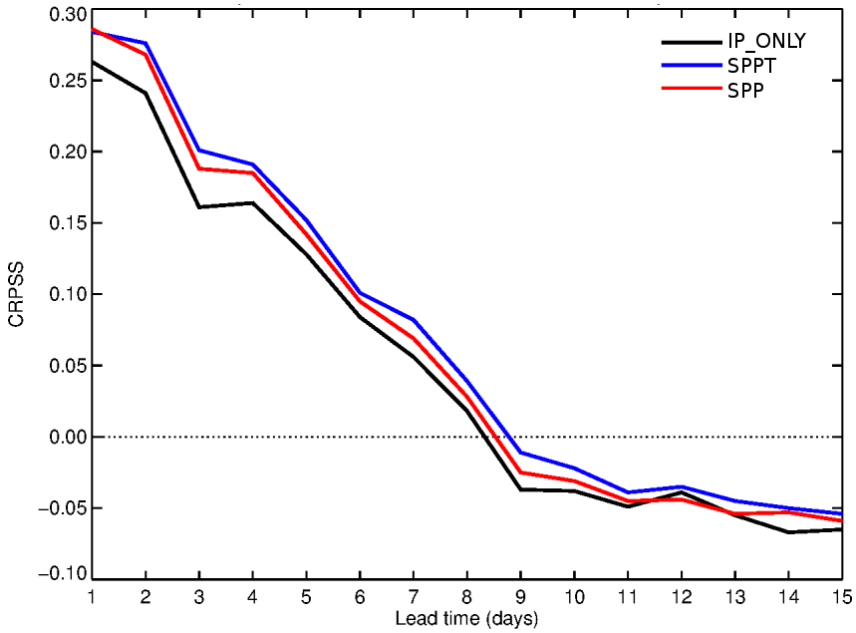

Figure 15: As in Fig 14, but for accumulated precipitation over 24h.

In order to better understand the differences in precipitation amounts, frequency bias for 24 h accumulated precipitation[3] at forecast day 5 is presented in Fig 16. Interestingly, SPP-L produces a better representation of rainfall amounts above the 5 mm threshold, i.e. for the heavier precipitation. However,

[3]Frequency bias is calculated as the ratio between forecast and observed rainfall amounts above a given threshold (with 1 being the perfect score). This provides direct validation of how well the model represents the amounts of light to heavy precipitation observed.

since the SPPT experiment is able to represent the more frequent light rain events better, it scores better than SPP-L in CRPSS. Dominance of the light rain events in forming CRPSS is confirmed when looking into the day 1 frequency bias (not shown): for the short lead times SPP-L produces a better representation of light (to heavy) rainfall events, which then results in very similar (or slightly better) CRPSS for SPP-L.

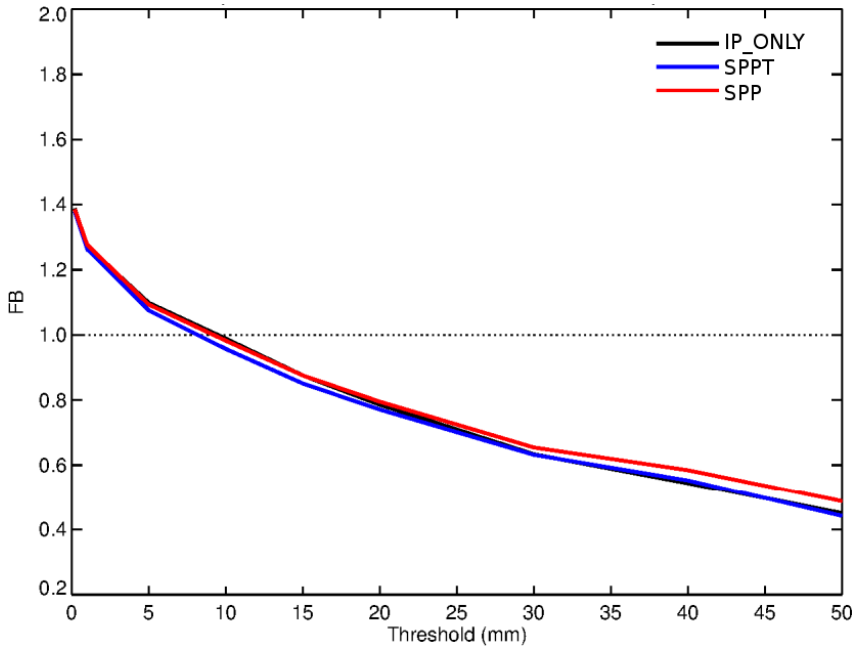

Figure 16: Frequency bias of accumulated precipitation over 24h in extra-tropics for forecast lead time of 5 days. Precipitation thresholds are set to 1, 5, 10, 15, ..., 50 mm. PPL (red), SPPT (blue) and IP_only (black). Means computed from 46 start dates between Dec 2013 and Nov 2014.

7 Impact on model climate

Finally, we evaluate the impact of the SPP scheme on the model climate. These model climate runs are set up to complement the earlier SPP-L, SPPT and IP_only experiments. Each "ensemble" is here constructed from a total of four 13-month model integrations at T_L255 resolution (∼80 km), initialised from four start times separated by 30 hours, starting from 00UTC 1st August 2000. Under this configuration, we limit ourselves to assessing the mean of these 4 integrations, with the first month omitted. Differences due to initial conditions ("IP_only" in the previous experiments) are now represented by the deterministic model integrating from four different initial states. In the perturbed experiments (SPP-L and SPPT), each of the four model integrations also includes the effect of the perturbation method. The verification is done against satellite observations and reanalysis data from the year covered by the model integrations. ERA-Interim (Dee et al., 2011) reanalysis data was used for tropical wind vector verification at 200, 700 and 925 hPa levels (W200, W700 and W925). The observation datasets used are Special Sensor Microwave Imager (SSM/I) (Hilburn and Wentz, 2008) for precipitation (TP(1)) and total column water vapour (TCWV(1)); Global Precipitation Climatology Project (GPCP) (Adler et al., 2003) for precipitation (TP(2)) and total column water vapour (TCWV(2)); International Satellite Cloud Climatology Project (ISCCP) for total cloud cover (TCC); Clouds and the Earth's Radiant Energy System (CERES) – Energy Balanced and Filled (EBAF) (Kato et al., 2013) for top-of-the-atmosphere net solar (TSR) and

thermal radiation (TTR).

Figure 17 presents the change of RMSE values of annual means w.r.t to IP_only for the SPP-L and SPPT experiments. The SPP-L experiment is consistently better than IP_only for all variables. Furthermore, the SPP-L experiment has the lowest RMSE values for all variables except W925, TSR and TCC. The SPPT experiment has the best RMSE in W925, TSR and TCC, but is worse than IP_only in W700, TTR and TCWV. The TCWV degradation in SPPT is especially large. The improved wind climatology in SPP-L originates largely from more accurate representation of the Indian summer monsoon (not shown). Consequently, the positive precipitation bias in the region during the summer months is reduced. Another, though less significant, cause for the improved wind climatology is better circulation in the spring-time Atlantic ocean. The reduced RMSE for TTR in the SPP-L experiment stems from a smaller positive bias over latitudes north of 50°N (not shown). A secondary cause is a reduced negative bias in the tropical regions. The TSR RMSE in SPP-L and IP_only are quite close to each other. It would have been feasible to tune the radiation parameter distributions to produce an overall improved radiation budget. (This became evident during the screening of the parameter distributions). However, we focused on finding a SPP configuration with good medium-range forecast skill and the improved climatology was a side product. Reduction in the TCWV RMSE in the SPP-L experiment is due to removal of (a part of) a negative tropical bias. The large RMSE values in SPPT are a consequence of these negative biases getting larger. Finally, the TCC RMSE for the SPP and SPPT experiments are close to each other. Interestingly, the causes for the improvements are quite different: whereas SPPT improves the climatology mostly through reducing a positive bias in the tropics, SPP reduces a negative bias in the mid-latitudes.

8 Discussion

The choice of the "large"-scale correlation patterns (2000 km and 72 h) as the default SPP configuration was done based on the ensemble skill in the upper air variables. It is, however, a counter-intuitive choice when linking the perturbations to the real atmosphere: synoptic weather systems move at faster speeds and encompass smaller areas than described by these large correlation scales. Nonetheless, the degraded ensemble skill due to the smaller-scale correlation patterns (SPP-S) indicates that the smaller scale patterns might be too noisy (for our purposes). The better skill of the larger perturbation patterns presumably stems from the perturbations having more time to affect the flow downstream and actively influence areas of developing critical weather phenomena. However, the globally frozen parameter values resulted in the worst ensemble mean RMSE scores among the SPP configurations. This indicates that global parameter perturbations are suboptimal as well. An explanation might lie in the close physical connections between the chosen parameters: as the scale of the correlation patterns increases, the covariances between parameters become more important (due to interactions between the perturbed processes). Persistently perturbing parameters independently can introduce systematic errors into the forecasting system due to misbalanced interaction between processes, reducing the overall skill.

Although parameter covariances for IFS are available from research experiments (Ollinaho et al., 2013a), how to apply them effectively still remains an open question (as was indicated by the experiments of Christensen et al. (2015)). We also note that multiple space and time correlation scales that differ from parameter to parameter are technically possible (as per the multiple scales in SPPT, Palmer et al., 2009). At this stage of the SPP development, it is not clear how to adequately constrain the degrees of freedom in a more complex scheme.

The set of standard deviations $\sigma(1)$, which was selected as the default, and the slightly altered set $\sigma(2)$ produced similar amounts of spread in all model variables and levels. Even the ensemble mean RMSE

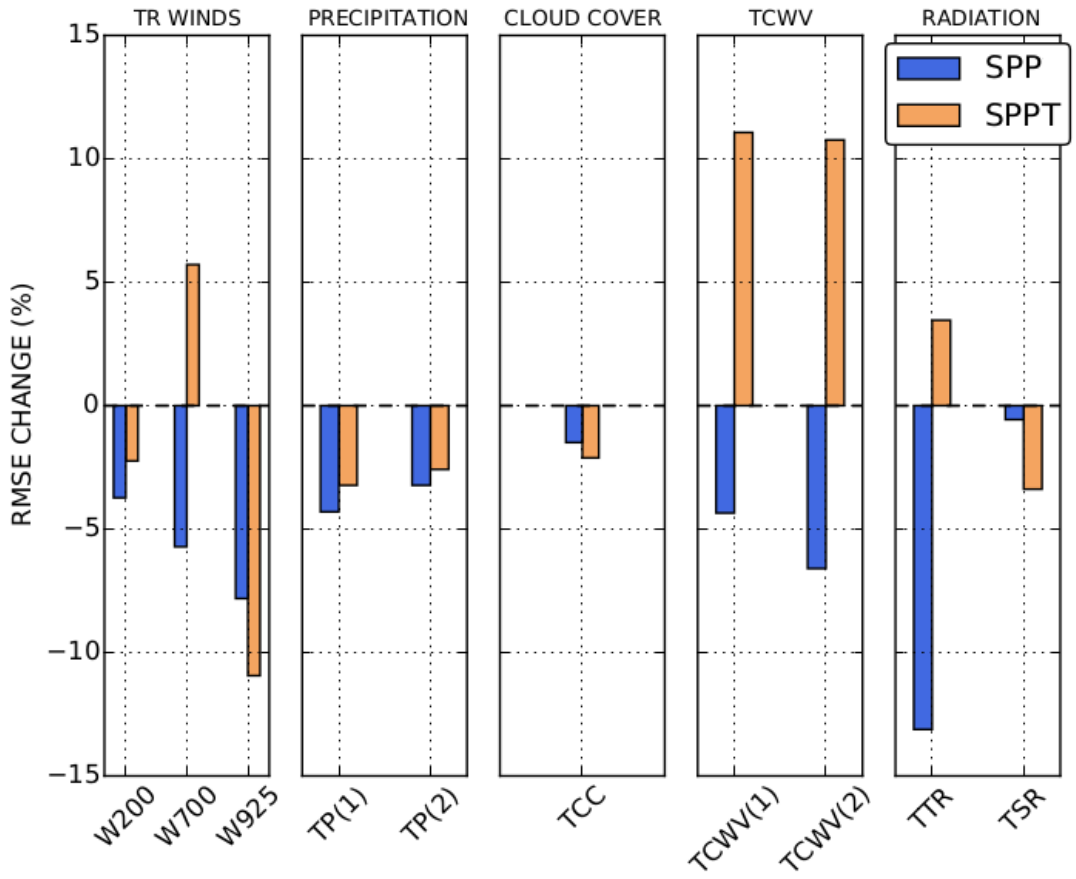

Figure 17: The change (%) of mean RMSE value for 1-year long model integrations verified against satellite observations and reanalysis data (see text for details) w.r.t. IP_only for SPPT and SPP-L experiments. Tropical winds at 200, 700 and 925 hPa verified against reanalysis (W200, W700, and W925); global precipitation scores verified against 2 sources (TP(1) and TP(2)); total cloud cover (TCC); total column water vapour verified against 2 sources (TCWV(1) and TCWV(2)); top-of-the-atmosphere net solar and thermal radiation (TTR and TSR).

values of the experiments using these distributions were relatively close to each other. This indicates that the SPP scheme is quite robust, and is not subject to major tuning procedures in order to become effective. Nonetheless, fine tuning the perturbation distributions does lead to improvements in the probabilistic scores. As the number of perturbed parameters increases, so does the challenge of finding the optimal perturbation distribution for each. A possibility in reducing the (person) resources needed for this task would be to utilize existing parameter estimation algorithms to automate the tuning process (like that introduced in Järvinen et al., 2012; Laine et al., 2012). However, exploring this possibility remains a challenge for the future.

Due to the nature of the log-normal distribution the tails of the distribution tend to extend to very high and low values when the perturbation standard deviations exceeds 0.6. Even though perturbations as large as an order of magnitude bigger (smaller) than the default parameter values can be argued to be reasonable for some parameters (e.g. aerosol optical thicknesses), tails reaching even larger (smaller) values are hard to justify. As discussed in Section 3, the largest standard deviations for any perturbation distributions used in this study (a value of 0.8) remain rather conservative. It would be interesting to experiment with some broader perturbation distributions with enforced clipping to avoid unphysically extreme values.

The very distinct uncertainty growth mechanism for the SPP and SPPT schemes is noticeable in the upper air and surface validation. As indicated by Figures 11–12 the scores of the two schemes are close to each other at forecast lead time of 24 h (and before), but also at the longest forecast lead times outside the Tropics. A confirmation for this is seen in Figure 13, where the SPP experiment produces similar or slightly better CRPS than SPPT at short lead times, but the SPPT experiment starts to gain lead at longer lead times. Similarly the surface verification (Figs 14–15) indicates that shortest forecast lead times are indeed more skilful in the SPP experiment. A thorough analysis of the precipitation frequency bias shows that the SPP scheme improves the represention of medium to heavy rainfall events throughout the forecast range, and light to medium events in the shortest forecast lead times (shorter than 24 h). Thus, the SPP scheme introduces fast-growing perturbations into the forecast from the beginning. However, these perturbations do not continue to grow as strongly after the initial period.

The advantages of the SPP-style model uncertainty representation are evident when the impact on model tendencies is compared to that of SPPT: the physical consistency of the model can be maintained in the boundary layer, where the SPPT style tendency perturbations may produce instabilities. As seen in Fig 8, directly perturbing the boundary layer via SPP yields larger ensemble spread than that due to the propagation of perturbations above the boundary layer via SPPT. The consistent application of perturbations throughout the model column due to SPP also ensures that individual ensemble members conserve energy: top-of-the-atmosphere and surface fluxes adjust to the tendency changes that are applied within the column. The strength of parameter perturbations is also obvious when comparing the spread in the cloud cover and radiation for the two schemes: even though SPP produces less variation in the cloud cover (for lead times greater than 24 h), it produces more variation in radiation since the cloud optical properties are perturbed directly. The observed climatological improvements to a range of variables imply benefits due to the more physically consistent model uncertainty representation from SPP than from SPPT and demonstrate that the SPP scheme does not increase model biases. The latter has been confirmed through a visual comparison of 2D-maps of biases. A multi-year climate experiment to verify the impact (as well as the impact on seasonal predictions) is left for future work.

Nonetheless, the SPPT scheme produces more skilful ensembles for lead times beyond day 1. This is evident from the upper air probabilistic scores. The greater spread observed in the free-troposphere in the SPPT experiment is already visible in the first 3 h of the forecasts (Fig 8), and is most pronounced in the extra-tropics. We note that the initial state perturbations used here have been tuned to produce the most skilful operational ensemble, which includes SPPT. Future testing will explore whether the initial state perturbations should be retuned for use with SPP.

The SPP stdev. of the tendencies in NH is close to zero near the tropopause (i.e. where convection stops), whereas the differences in the SPPT ensemble extend up to the stratosphere. The inactivity of the SPP scheme can be explained (to some extent) by the perturbed radiation parameters not affecting the critical processes at these levels (strongly enough). Parameters and variables associated with vertical diffusion above the boundary layer are also currently untouched by SPP. Perturbing such quantities could increase the spread and skill due to SPP at higher levels. Figure 8 also shows that the two ensembles have quite similar total spread in the beginning of the forecast in a deep layer (in the Tropics), which coincides with large decorrelations in the spread, implying differences in the spatial and temporal structures generated by the two schemes.

Because the SPP and SPPT schemes seem to provide quite different mechanisms for model uncertainty representations, they offer an interesting possibility for a comprehensive study of model uncertainties. Furthermore, a hybrid system with both of the schemes active will form the subject of our future investigations. Since it is likely that neither the SPPT nor the SPP scheme can alone fully represent all model uncertainty stemming from the physical parametrisations, such a hybrid system seems a natural

evolution towards a more comprehensive representation of the true distribution of model errors. This has already been tested by activating SPP perturbations for the four turbulent diffusion and subgrid orography parameters alongside SPPT. In this simple setup, the two uncertainty representations overlap as little as possible since SPPT is inactive in the boundary layer. The initial results from this system have shown strong potential. We envisage a hybrid system could be active in a future operational upgrade of ENS.

In order for a model uncertainty representation to be useful in an operational ensemble prediction system it should not dominate the overall cost of integrating the model. Accurate estimates of the computational cost of the scheme have been obtained by looking at the distribution of elapsed time for each of the 720 integration time steps of the model for all 20 perturbed forecasts and 45 start dates. Thus, the estimates are based on the run times of 6.5×10^5 model time steps. Outliers in the distribution of elapsed times that are mainly due to model I/O have been removed prior to computing the median of the distribution. The SPP scheme increases the median runtime of non-I/O time steps by 5.7%. This compares with an increase of 1.8% for SPPT. These moderate contributions to computational cost are deemed affordable for operational usage of either scheme.

We also note that the presented SPP scheme is an initial implementation and there is scope for improvement by optimising the choice of parameters, as well as the magnitude, time and space scales of the perturbations. It is important that the chosen parameters encompass the uncertainty across the whole range of meteorological regimes and phenomena in order to represent the true uncertainty in the model forecast and this may mean adding additional parameters. This modularity is a major strength of the SPP approach, i.e. that the uncertainty of individual processes can be represented with increasing fidelity. Moreover, the ensemble system can be adjusted to take proper account of new physical parameters and representations. The limit is to represent uncertainty in *every* process in the model. In practice we have to consider resource costs (both computational as well as personnel) and limit the parameter set to a manageable number.

9 Conclusions

In this paper, we have introduced a process-level stochastic representation of model uncertainties by perturbing a set of 20 parameters and variables in the ECMWF IFS model. The perturbed quantities are chosen from the IFS parametrisations of (a) turbulent diffusion and subgrid orography, (b) convection, (c) clouds and large-scale precipitation, and (d) radiation. The scheme, Stochastically Perturbed Parametrisations (SPP), applies perturbations, which vary in time and space, independently to each of the parameters and variables. The perturbations are drawn from log-normal and normal distributions, with a prescribed mean and standard deviation for each parameter and variable. The perturbation patterns, containing spatial and temporal correlations of 2000 km and 72h respectively, are evolved according to a first order auto-regressive (AR(1)) process.

The scheme is tested in the ECMWF ensemble (ENS) with a resolution of T_L399 up to forecast day 15. The experiments assess the performance of the ENS when SPP is applied alongside initial state perturbations. SPP exhibits little sensitivity to small changes to the standard deviations of the perturbation patterns. Halving the standard deviations, however, leads to a large drop in the ensemble spread. The choice of correlation scales is also shown to impact the ensemble skill. Correlation scales of 2000 km and 72 h are chosen, since both smaller and larger scales have been observed to reduce skill.

The probabilistic skill of the scheme is benchmarked against the SPPT scheme, which is one of the two currently operational model uncertainty representations in the ENS. A comparison of upper air scores shows that the SPPT scheme produces more skilful ensembles for medium range lead times. However,

for short lead times (up to 24 h) the two schemes display a similar skill in forecasting the upper air variables, with SPP producing more consistent improvements when compared to initial state perturbations only. A verification against surface observations of 2 m temperature shows SPP having more skill in the first couple of forecast days. A comparison of precipitation frequency biases between the two schemes indicates that SPP improves the representation of light to heavy rain in the short range. Examining the behaviour of the model tendencies reveals that the SPP scheme introduces greater variability into the ensemble in the boundary layer. We show that by directly perturbing the processes associated with radiative effects due to clouds a greater impact is gained on radiative fluxes than by indirectly perturbing the fluxes via SPPT. Long model integrations reveal that SPP generates a better model climate for a range of variables, advocating the advantages of a physically consistent model uncertainty representation.

This initial implementation of the SPP scheme has already shown the potential of a process-level representation of model uncertainties. The differences in the perturbed tendencies are an indication that the SPP scheme represents different sources of uncertainty than the SPPT scheme. The differences between the two schemes could be exploited to construct a more comprehensive model uncertainty representation based on a combination of both schemes.

Acknowledgements

We thank Dr. Simon Lang for computing the changes in the run times. We are grateful to Dr. Nils Wedi, Dr. Roberto Buizza and Prof. Erland Källen for their valuable comments on the manuscript. We also want to thank the two anonymous reviewers for their feedback that helped to improve and clarify the manuscript.

References

Adler, R., et al., 2003: The version 2 global precipitation climatology project (GPCP) monthly precipitation analysis (1979-present). *J. Hydrometeor.*, **4**, 1147–1167.

Baker, L., A. Rudd, S. Migliorini, and R. Bannister, 2014: Representation of model error in a convective-scale ensemble prediction system. *Nonlinear Processes in Geophysics*, **21**, 19–39.

Bechtold, P., N. Semane, P. Lopez, J.-P. Chaboureau, A. Beljaars, and N. Bormann, 2014: Representing equilibrium and nonequilibrium convection in large-scale models. *J. Atmos. Sci.*, **71**, 734–753.

Beljaars, A., A. R. Brown, and N. Wood, 2004: A new parametrization of turbulent orographic form drag. *Quarterly Journal of the Royal Meteorological Society*, **130 (599)**, 1327–1347.

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.

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 (632)**, 703–722, doi:10.1002/qj.234, URL http://dx.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.

Buizza, R., M. Miller, and T. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF Ensemble Prediction System. *Quart. J. Roy. Meteor. Soc.*, **134**, 2041–2066.

Charron, M., G. Pellerin, L. Spacek, P. L. Houtekamer, N. Gagnon, H. L. Mitchell, and L. Michelin, 2010: Toward random sampling of model error in the Canadian ensemble prediction system. *Mon. Wea. Rev.*, **138**, 1877–1901, URL http://dx.doi.org/10.1175/2009MWR3187.1.

Christensen, H. M., I. M. Moroz, and T. N. Palmer, 2015: Stochastic and perturbed parameter representations of model uncertainty in convection parameterization. *J. Atmos. Sci.*, **72**, 25252544, doi: http://dx.doi.org/10.1175/JAS-D-14-0250.1.

Dee, D. P., et al., 2011: The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, **137 (656)**, 553–597, doi: 10.1002/qj.828.

ECMWF, 2015: Forecasts. http://www.ecmwf.int/en/forecasts.

Grubišić, V. and M. W. Moncrieff, 2000: Parameterization of convective momentum transport in highly baroclinic conditions. *J. Atmos. Sci.*, **57**, 3035–3049.

Hilburn, K. A. and F. J. Wentz, 2008: Intercalibrated passive microwave rain products from the unified microwave ocean retrieval algorithm (UMORA). *J. Appl. Meteor. Clim.*, **47**, 778–794.

Iversen, T., A. Deckmyn, C. Santos, K. Sattler, J. B. Bremnes, H. Feddersen, and I.-L. Frogner, 2011: Evaluation of 'GLAMEPS' — a proposed multimodel eps for short range forecasting. *Tellus A*, **63 (3)**, 513–530, doi:10.1111/j.1600-0870.2010.00507.x, URL http://dx.doi.org/10.1111/j.1600-0870.2010.00507.x.

Järvinen, H., M. Laine, A. Solonen, and H. Haario, 2012: Ensemble prediction and parameter estimation system: the concept. *Q.J.R. Meteorol. Soc.*, **138**, 281–288, doi:10.1002/qj.923.

Kato, S., N. G. Loeb, F. G. Rose, D. R. Doelling, D. A. Rutan, T. E. Caldwell, L. Yu, and R. A. Weller, 2013: Surface irradiances consistent with CERES-derived top-of-atmosphere shortwave and longwave irradiances. *J. Climate*, **26**, 2719–2740, doi:doi:10.1175/JCLI-D-12-00436.1.

Laine, M., A. Solonen, H. Haario, and H. Järvinen, 2012: Ensemble prediction and parameter estimation system: the method. *Q.J.R. Meteorol. Soc.*, **138**, 289–297, doi:10.1002/qj.922.

Leutbecher, M. and S. T. K. Lang, 2014: On the reliability of ensemble variance in subspaces defined by singular vectors. *Quart. J. Roy. Meteor. Soc.*, 1453–1466.

Leutbecher, M. and T. Palmer, 2008: Ensemble forecasting. *J. Comp. Phys.*, **227 (7)**, 3515–3539, doi: 10.1016/j.jcp.2007.02.014.

Lott, F. and M. J. Miller, 1997: A new subgrid-scale orographic drag parametrization: Its formulation and testing. *Quarterly Journal of the Royal Meteorological Society*, **123 (537)**, 101–127.

Martin, G. M., D. W. Johnson, and A. Spice, 1994: The measurement and parameterization of effective radius of droplets in warm stratocumulus clouds. *J. Atmos. Sci.*, **51**, 1823, doi:http://dx.doi.org/10.1175/1520-0469(1994)051⟨1823:TMAPOE⟩2.0.CO;2.

Morcrette, J. J., H. W. Barker, J. N. S. Cole, M. J. Iacono, and R. Pincus, 2008: Impact of a new radiation package, McRad, in the ECMWF integrated forecasting system. *Mon. Wea. Rev.*, **136**, 4773–, doi: http://dx.doi.org/10.1175/2008MWR2363.1.

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 in Geophysics*, **20 (6)**, 1001–1010, doi:10.5194/npg-20-1001-2013.

Ollinaho, P., 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 (675)**, 1520–1532, doi:10.1002/qj.2044.

Palmer, T., R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G. Shutts, M. Steinheimer, and A. Weisheimer, 2009: Stochastic parameterization and model uncertainty. Tech. Mem. 598, ECMWF, 42 pp. [Available online at: http://www.ecmwf.int/publications/library/ecpublications/_pdf/tm/501-600/tm598.pdf].

Pincus, R., H. Barker, and J.-J. Morcrette, 2003: A fast, flexible, approximate technique for computing radiative transfer in inhomogeneous cloud fields. *J. Geophys. Res.*, **108 (D13)**, doi:10.1029/2002JD003322, URL http://dx.doi.org/10.1029/2002JD003322.

Sandu, I., P. Bechtold, A. Beljaars, A. Bozzo, F. Pithan, T. G. Shepherdx, and A. Zadra, 2015: Impacts of parameterized orographic drag on the northern hemisphere winter circulation. *Journal of Advances in Modeling Earth Systems*.

Shutts, G., M. Leutbecher, A. Weisheimer, T. Stockdale, L. Isaksen, and M. Bonavita, 2011: Representing model uncertainty: stochastic parametrizations at ECMWF. *ECMWF Newsletter*, **129**, 19–24, [Available online at: http://www.ecmwf.int/publications/newsletters/pdf/129.pdf].

Tegen, I., P. Hollrig, M. Chin, I. Fung, D. Jacob, and J. Penner, 1997: Contribution of different aerosol species to the global aerosol extinction optical thickness: Estimates from model results. *J. Geophys. Res.*, **102**, 23 895–23 915, doi:doi:10.1029/97JD01864.

Weaver, A. and P. Courtier, 2001: Correlation modelling on the sphere using a generalized diffusion equation. *Quart. J. Roy. Meteor. Soc.*, **127**, 1815–1846.

Weisheimer, A., S. Corti, T. Palmer, and F. Vitart, 2014: Addressing model error through atmospheric stochastic physical parametrizations: impact on the coupled ECMWF seasonal forecasting system. *Phil. Trans. R. Soc. A*, **372 (2018)**, doi:10.1098/rsta.2013.0290.

Supplemental Materials

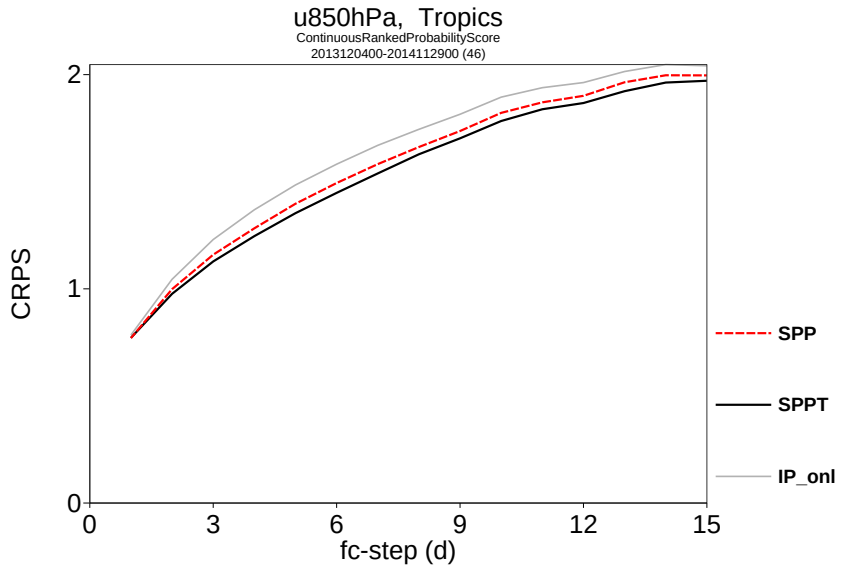

Figure S1: Continuous Ranked Probability Score for u-component of wind (in m/s) at 850 hPa in Tropics for 15 day lead times. IP_only (grey), SPPT (black) and SPP-L (red dashed) experiments.

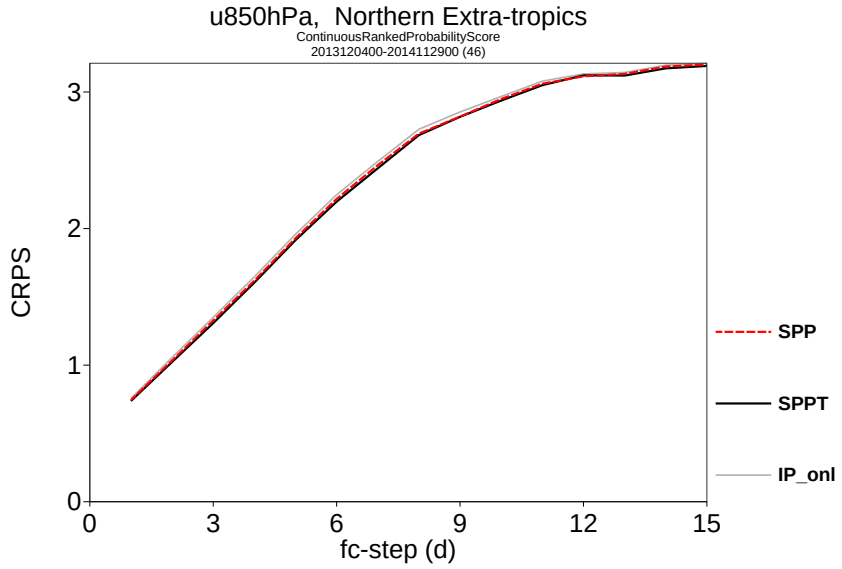

Figure S2: Continuous Ranked Probability Score for u-component of wind at 850 hPa in Northern extra-tropics for 15 day lead times. IP_only (grey), SPPT (black) and SPP-L (red dashed) experiments.