

Model Uncertainty Representation in a Convection-Permitting Ensemble—SPP and SPPT in HarmonEPS

INGER-LISE FROGNER,^a ULF ANDRAE,^b PIRKKA OLLINAHO,^c ALAN HALLY,^d KAROLIINA HÄMÄLÄINEN,^c
JANNE KAUFANEN,^c KARL-IVAR IVARSSON,^b AND DANIEL YAZGI^b

^a Norwegian Meteorological Institute (Met Norway), Oslo, Norway

^b Swedish Meteorological and Hydrological Institute, Norrköping, Sweden

^c Finnish Meteorological Institute, Helsinki, Finland

^d Irish Meteorological Service (Met Éireann), Dublin, Ireland

(Manuscript received 30 April 2021, in final form 15 December 2021)

ABSTRACT: The stochastically perturbed parameterizations scheme (SPP) is here implemented and tested in HarmonEPS—the convection-permitting limited area ensemble prediction system by the international research program High Resolution Limited Area Model (HIRLAM) group. SPP introduces stochastic perturbations to values of chosen closure parameters representing efficiencies or rates of change in parameterized atmospheric (sub)processes. The impact of SPP is compared to that of the stochastically perturbed parameterization tendencies scheme (SPPT). SPP in this first version in HarmonEPS perturbs 11 parameters, active in different atmospheric processes and under various weather conditions. The main motivation for this study is the lack of variability seen in cloud products in HarmonEPS, as reported by duty forecasters. SPP in this first version is able to increase variability in a range of weather variables, including the cloud products. However, for some weather variables the root-mean-squared error of the ensemble mean is increased and the mean bias is impacted, especially in winter. This indicates that (some) parameter perturbation distributions are not optimal in the current configuration, and a further sensitivity analysis is required. SPPT resulted in a smaller increase in variability in the ensemble, but the impact was almost completely masked out when combined with perturbations of the initial state, lateral boundaries, and surface properties. An in-depth investigation into this lack of impact from SPPT is here presented through examining, among other things, accumulated tendencies from the model physics.

SIGNIFICANCE STATEMENT: Small inaccuracies, simplifications, or errors in any part of a complex and nonlinear system like a weather model can amplify and in a short time become significant. We wanted to introduce a physically consistent way of representing these uncertainties in a model that is used in several European countries. To do this we introduce variations in a few parameters that are used in the model description, and that we know are uncertain. By doing this we were able to increase the variability of the cloud products as desired. We see this as a promising approach for capturing the possibilities of fog occurring or not in this model. Further refinements are needed before it can be used in operational weather forecasts.


KEYWORDS: Fog; Clouds; Ensembles; Mesoscale models; Model errors

1. Introduction

Ensemble prediction systems (EPSs) are the commonly used framework in numerical weather prediction (NWP) to provide information on the possible future states of the atmosphere, taking into account uncertainties that exist in different parts of the forecasting system. The main sources of uncertainty in NWP models originate from (i) incomplete reconstruction of the current atmospheric state (due to lack of observations, limitations in data assimilation, etc.), and (ii) errors in model construction (arising from the need to approximate and discretize the atmospheric governing equations, which then results in parameterization of unresolved processes). These are referred to as initial state uncertainty and model uncertainty, respectively. A third source of uncertainty

arises from how interactions are handled between the atmosphere and other Earth system components (oceans, glaciers, etc.). In limited area modeling (LAM) an additional uncertainty source comes from how lateral boundary conditions from the host model are handled (see e.g., Frogner et al. 2019).

HarmonEPS (Frogner et al. 2019) is a convection-permitting LAM EPS developed by the international research program High Resolution Limited Area Model (HIRLAM) group. The EPS configuration used here includes initial state, surface and lateral boundary uncertainty representations. This study is motivated by feedback from duty forecasters related to insufficient spread characteristics in HarmonEPS cloud products. From a forecaster's perspective, the uncertainty regarding clouds, and especially low clouds, is of special interest. One reason is that it also affects other prognostic

 Denotes content that is immediately available upon publication as open access.

Corresponding author: Inger-Lise Frogner, i.l.frogner@met.no



This article is licensed under a [Creative Commons Attribution 4.0 license](http://creativecommons.org/licenses/by/4.0/) (<http://creativecommons.org/licenses/by/4.0/>).

DOI: 10.1175/MWR-D-21-0099.1

© 2022 American Meteorological Society.

parameters such as temperature near the ground. Currently, the spread of the ensemble is considerably lower than the forecast error [e.g., root-mean-square error (RMSE) of the ensemble mean]. It is often seen that all members have the same misplacement or have the same over or underprediction of the clouds. Thus, the issue of too confident ensembles in difficult-to-predict weather situations (related especially to cloud products) is often brought up by duty forecasters. On top of this, forecasting clouds in the correct place is important for generating products for atmospheric icing for wind energy, power lines, and aviation (see e.g., Bernabò et al. 2015; Kraj and Bibeau 2010; Nygaard et al. 2016).

Some operational HarmonEPS configurations have opted to use multiphysics or multimodel approaches, but there is currently no universal approach to handling model uncertainties in HarmonEPS. Stochastically perturbed parameterization tendencies (SPPT) (see e.g., Bouttier et al. 2012) is available as an option, but is not used operationally in any suites due to the small effect seen in previous studies (see Frogner et al. 2019). Finding an efficient model uncertainty representation is therefore the most obvious pathway that could result in a more realistic variability in cloud products.

Representing model uncertainties in weather forecasts is challenging, and it continues to be an active area of research (see e.g., Leutbecher et al. 2017). Model errors are complex and arise from a multitude of sources, therefore several different approaches for representing them in EPS settings have been developed over the years (see e.g., Ollinaho et al. 2017). In this paper, the focus is on model uncertainty representations accounting for errors in (i) total tendency contributions from physical parameterizations of the model, and in (ii) chosen values of closure parameters controlling the efficiencies or rates of change of parameterized (sub)processes. For item i, SPPT is used. For item ii, the recently developed stochastically perturbed parameterizations (SPP) methodology is applied following the implementation from the European Centre for Medium-Range Weather Forecasts (ECMWF) (Ollinaho et al. 2017; Lang et al. 2021). SPP has been applied in several LAM EPSs, e.g., (i) Wastl et al. (2019a) applied SPP in a hybrid setup in Convection-Permitting Limited Area Ensemble Forecasting (C-LAEF), (ii) Jankov et al. (2019) use SPP in High-Resolution Rapid Refresh (HRRR) ensemble, and (iii) Thompson et al. (2021) have implemented SPP in Weather Research and Forecasting (WRF) Model. The latter two implementations use a slightly different SPP configuration to that used by ECMWF and in this study (described in section 5). In this first implementation of SPP in HarmonEPS, 12 key parameters have been chosen and tested with an emphasis on parameters related to clouds and microphysics (7 of the 12) with the aim of trying to overcome the lack of variability in clouds in particular. Based on initial testing, 11 of the 12 parameters were accepted for further studying. Details about the excluded parameter are, however, provided in section 5.

The ensemble system is described in section 2, the experimental setup in section 3 and the verification methodology used in section 4. Section 5 describes the SPPT and SPP implementations and how the two model uncertainty schemes perform in HarmonEPS. Section 6 is devoted to an in-depth

investigation on the different perturbations in HarmonEPS and how they interact. Finally a general discussion and conclusions are presented in section 7.

2. HarmonEPS—The HARMONIE-AROME ensemble prediction system

The convection-permitting HARMONIE-AROME model (Bengtsson et al. 2017) is developed within the ALADIN-HIRLAM NWP system (Termonia et al. 2018) and the system serves as the operational forecasting tool in a number of countries in Europe. The forecast model is run with a 2.5-km horizontal grid spacing with 65 levels in the vertical. The upper air data assimilation system is based on three-dimensional variational data assimilation (3DVAR) (Brousseau et al. 2011) with 3-hourly cycling. At the surface 2-m temperature (T2m) and relative humidity (RH2m) as well as snow cover are assimilated using optimal interpolation (Giard and Bazile 2000). The ensemble prediction part of the system, HarmonEPS (Frogner et al. 2019), is used in this study, and it supports a wide range of perturbation methodologies dealing with initial, model and boundary uncertainty.

The following perturbations described in Frogner et al. (2019) are used in this study: (i) the initial condition (IC) perturbations created from applying the difference between the ECMWF operational EPS (ECMWF ENS) member and control to the HarmonEPS control member analysis. (ii) The lateral boundary perturbations (LBC). It must be noted that the LBCs are not actual perturbations, but rather balanced states from the corresponding ensemble member from the ECMWF ENS (Sleigh et al. 2019). (iii) At the surface, perturbations are applied to each members' surface analysis following Bouttier et al. (2016) with perturbations added to model fields kept constant during the forecasts (such as sea surface temperature, vegetation and leaf area index) as well as to fields evolving during the forecasts (temperature and moisture in the soil). For more details on IC, LBC, and surface perturbations in HarmonEPS the reader is referred to Frogner et al. (2019).

3. Experimental setup

HarmonEPS has been run in three different setups in this study, all with the operational horizontal and vertical resolution of HARMONIE-AROME, and with 3-hourly cycling with 3DVAR for the control: (i) each member is using the control member's upper air analysis with initial perturbations (IC) added as described above. All members run their own surface analysis using optimal interpolation, and surface perturbations are applied. The handling of the lateral boundaries is as described above (LBC). Since we are initializing the model surface fields from an ECMWF ENS model state, a two-week spinup period is run prior to the start of the experiment periods to allow the slow soil variables to adapt to the HarmonEPS model climate. This is the reference setup in HarmonEPS (experiments called REF). Experiments where SPPT or SPP is added to REF also belong to this experiment type. The forecasts are run for +48-h lead time. (ii) In section 6 how each of the perturbation types affect the ensemble in

TABLE 1. Characteristics of the different experiment types used in this study.

	Type i	Type ii	Type iii
Upper air assimilation	Control only	Control only	Control only
Surface assimilation	All members	Control only	Control only
Activated perturbations	IC, LBC, and surface. Adding either SPPT or SPP as stated in the text.	IC, LBC, surface, SPPT, SPP. One at a time.	SPPT or SPP with individual parameters, members start from control initial conditions every cycle
Spinup	Yes	Yes	No
Forecast length	+48 h	+48 h	+36 h
Forecast period and number of start dates	February 2019—28 start dates (SPP and SPPT in full setup). June 2019—30 start dates (SPP in full setup).	1 Feb 2019–5 Feb 2019 (5 start dates)	30 May 2016–5 Jun 2016 (7 start dates) (SPP and SPPT sensitivity). 20 Feb 2019–26 Feb 2019 (7 start dates) (SPP sensitivity).

isolation is studied. Here, the same setup as in setup i is used, except that each perturbation type is tested separately, including SPPT and SPP, and that the ensemble members use the surface analysis of the control in order to clearly distinguish the effect of the different perturbations without the ensemble members getting different surface histories. (iii) The third setup is used for the sensitivity experiments with SPPT and SPP. This setup is designed to only perturb the model physics, keeping everything else equal for the ensemble members. Here, the ensemble members start from the control member initial conditions every cycle. No prior spinup period was deemed necessary for these tests. The forecast lead time used is +36 h. For all three experiment types we run the full forecast length from 0000 UTC, while the intermediate times are only short forecasts used for cycling. A summary of the experiment details is presented in Table 1.

All the experiments are run with six ensemble members plus a control member. Initial tests were conducted with more ensemble members, but in order to make testing computationally affordable there was a need to reduce the ensemble member count. Six members was found to be sufficient to maintain the signal on how the ensemble skill was changing between the different experiments. This is also in line with the results of Clark et al. (2011) who showed that their relatively small ensemble (3–9 members) had statistically indistinguishable average ROC areas relative to their full 17-member ensemble, and Keil et al. (2019) who tested the effect of the different ensemble sizes by applying a resampling method with replacement and got qualitatively similar results.

4. Verification methodology

The verification of the different experiments is done against point observations and against satellite-observed cloud masks. For the point observations near surface parameters are verified against SYNOP stations, while upper air parameters are verified against radiosondes. Bilinear interpolation is used to interpolate the forecast values to the observation points. For

2-m temperature a correction is applied to account for the different elevation in model and observations, using the standard adiabatic lapse rate of 6.5 K km^{-1} . A gross error check is performed, where unrealistic values are removed. Then a further check is performed, where observations that are more than six standard deviations away from the forecast values are removed.

For the point verification the following validation metrics are used to show the relative performance of the different experiments. All ensemble members, including the control, are used in the calculations:

- RMSE: The root-mean-square error of the ensemble mean of the forecast compared with observations.
- The ensemble spread, or variability: the standard deviation of the ensemble members around the ensemble mean:

$$\text{spread} = \sqrt{\frac{1}{N-1} \sum_{n=1}^N (x_n - \mu)^2},$$

where μ is the ensemble mean. The ensemble spread should be equal to the RMSE for a well calibrated ensemble.

- Mean bias: The ensemble mean – the observation, averaged over all cases.
- fCRPS: The fair continuous ranked probability score (Leutbecher 2019). CRPS is a negatively oriented score that measures the distance of a continuous distribution function to the observed value. CRPS reduces to the mean absolute error of the forecast in the case of a one-member ensemble. The fair version of CRPS corresponds to the expected CRPS of an infinite sized ensemble. Leutbecher (2019) conclude that using the fair CRPS with an ensemble size of four to eight ensemble members is sufficient for most research experiments.

For the spatial verification the following validation metric is used:

- FSS: Fractions skill score (Roberts and Lean 2008; Roberts 2008). FSS is a measure of model forecast skill as a function of spatial scale.

FSS is used here to evaluate the model forecast skill for clouds, where the forecast total cloud cover C_f is assessed against a satellite-observed cloud mask. Crocker and Mittermaier (2013) employed a cloud mask to assess spatial model bias using contingency table metrics and object-based methods. This work concluded that using a cloud mask can give a quick assessment of the forecast model tendency to contaminate clear sky with low cloud fractions when low thresholds are used.

To undertake the model evaluation, a forecast cloud mask M_f is extracted from the predicted total cloud cover by defining a threshold q in the following way:

$$M_x = \begin{cases} 1 & \text{if } C_x \leq q \\ 0 & \text{otherwise} \end{cases}, \quad (1)$$

where M_x is the masked field, C_x is the cloud fraction and subscript x is o for observations and f for forecast fields.

More than 50% of the model's domain is covered by clouds in all dates covered by the satellite data used in this study. Thus, the model performance is assessed by forecasting clear areas instead of clouds (Crocker and Mittermaier 2013). For this reason, M_x is set to 1 where C_x is less than a given threshold. In this case an event of being cloud free is defined at each model grid cell.

The satellite-observed cloud masks used in this study are a product of the Polar Platform System (PPS) of EUMETSAT Satellite Application Facilities for Nowcasting and very short range forecasting (SAFNWC) (Thoss 2014a,b). Since the resolution of satellite-observed cloud masks [1 km for AVHRR and 750 m for VIIRS, Thoss (2014b)] is higher than the model resolution (2.5 km), the fraction of cloudy pixels is computed for each model grid cell. This process results in observed cloud fractions of values between 0 and 1. The resulting field is converted to binary (M_o) by Eq. (1) after defining a threshold.

In this study, the threshold q is chosen to be 0.2. A low threshold means more clouds and less cloud-free grid cells. Using a lower threshold mimics the cloud mask generation algorithm which describes a cell as being cloudy even when only thin cirrus clouds are present. Only dates when the satellite data covers more than 80% of the model domain are considered. The maximum satellite time deviation from forecast valid time is chosen to be 5 min.

To compute the FSS, first, the fraction of event occurrences in the grid cell neighborhood is calculated. A grid cell neighborhood of scale s is defined as the square centered on that grid cell and covers $(2s + 1)^2$ grid cells where the scale s is 0, 1, 2, ... etc. Since the model grid spacing is 2.5 km, the scale number s corresponds to the spatial-scale length of $2.5(2s + 1)$ km.

As the experiments in this study have one control member and six perturbed members, seven values are defined at each grid cell for a given lead time when the control member is included and six values when it is excluded.

FSS is defined by the formula below:

$$\text{FSS} = 1 - \frac{\sum (\text{PF}_{ij} - \text{OF}_{ij})^2}{\sum (\text{PF}_{ij}^2 + \text{OF}_{ij}^2)}, \quad (2)$$

where PF_{ij} and OF_{ij} are the forecast and observed event fractions at grid cell ij , respectively, and following (Schwartz et al. 2010)

$$\text{PF}_{ij} = \frac{1}{N} \sum_{k=1}^N \text{PF}_{ijk}, \quad (3)$$

where PF_{ijk} is the forecast event fraction (here, being cloud free) for grid cell ij and member k .

As this study was motivated by the insufficient spread characteristics in HarmonEPS cloud products, there is naturally a focus on the spread when evaluating the experiments. Also included are the FSS for clouds, RMSE of the ensemble mean, the mean bias and the fCRPS for a range of weather variables to see how the perturbations introduced may or may not alter the mean behavior of the ensemble. Other metrics were also looked at, but were found not to add any extra insight and are therefore not included.

The statistical significance of the differences between two experiments are calculated using a bootstrap approach with 1000 replicates, computed independently at each lead time from the observation/forecast data pooled for each forecast start date. The test is insensitive to spatial autocorrelations since all stations are represented in each pool. The sample size varies between approximately 350 for the cloud variables, to 850 for the near surface weather parameters. Statistical significance is calculated for the levels 99.7%, 95%, and 68%, meaning that the differences are considered to be significant or not at these confidence levels. Examples of how this is used and presented in scorecards are shown in Figs. 7 and 8. Observations uncertainty are not taken into account, this is motivated by the findings of Frogner et al. (2019) who argue that it is of less importance when comparing relative performance of different model configurations. This is also in line with Jankov et al. (2017) and Lang et al. (2021).

5. Model uncertainty schemes—SPPT and SPP

a. SPPT

The stochastically perturbed parameterization tendencies scheme introduces stochastic perturbations to the tendencies of horizontal wind components, specific humidity and temperature produced by the physical parameterizations of the model. The perturbations are applied at each model time step and after the contribution of each physics parameterization has been calculated. SPPT has been described in detail by many other authors including Palmer et al. (2009), Bouttier et al. (2012) and Frogner et al. (2019). The formulation used here is based on the LAM implementation of SPPT described in Bouttier et al. (2012). SPPT is tapered in the stratosphere and below approximately 1200 m in the boundary layer to avoid instabilities. Other systems such as the Austrian C-LAEF use a partially perturbed parameterization tendency technique or pSPPT, based on the work of Wastl et al. (2019a). In this approach, the partial tendencies of the physics parameterization schemes are perturbed separately, which is in contrast to the traditional SPPT approach implemented in HarmonEPS. This approach allows the boundary layer tapering to be switched off and thus tendency style perturbations can play an enhanced role (Wastl et al. 2019a).



FIG. 1. The integration area.

In the HarmonEPS implementation of SPPT the stochastic pattern generator (SPG; [Tsyrlunikov and Gayfulin 2017](#)) is employed for the generation of the random perturbation fields. This pattern generator has the advantage of accounting for “proportionality of scales,” meaning it takes into account the fact that longer spatial scales live longer than shorter spatial scales, which die out quicker, a widespread feature in geophysics. In SPG, the perturbations vary spatially and temporally, and are correlated through a third order in time stochastic differential equation with a pseudodifferential spatial operator defined on a limited area. The implementation in HarmonEPS interfaces the code provided by [Tsyrlunikov and Gayfulin \(2017\)](#) and is solely defined by the spatial (XLCOR) and temporal (TAU) correlation length scales, and the standard deviation, SDEV. Furthermore, SPG provides an initialization to ensure stationary statistics from the start of the integration.

1) SPPT SENSITIVITY TESTS

A number of sensitivity tests were carried out to investigate the optimum settings for XLCOR and TAU for the SPPT perturbations. Sensitivity tests were also designed to look into the influence of the clipping ratio (XCLIP) and size of the perturbations (controlled by the standard deviation of the perturbations, SDEV). Various ranges were used to test each one of these SPPT control parameters; for XLCOR, lengths of 100–2000 km; for TAU, time scales of 6–24 h; for SDEV and XCLIP values of 0.1–1.0 and 10.0–1.0 were used, respectively. The ranges used for SDEV and the clipping ratio XCLIP ensured the perturbation coefficients were clipped at -1 and 1 . All sensitivity tests were carried out over the domain shown in [Fig. 1](#), over the 7 days from 30 May 2016 to

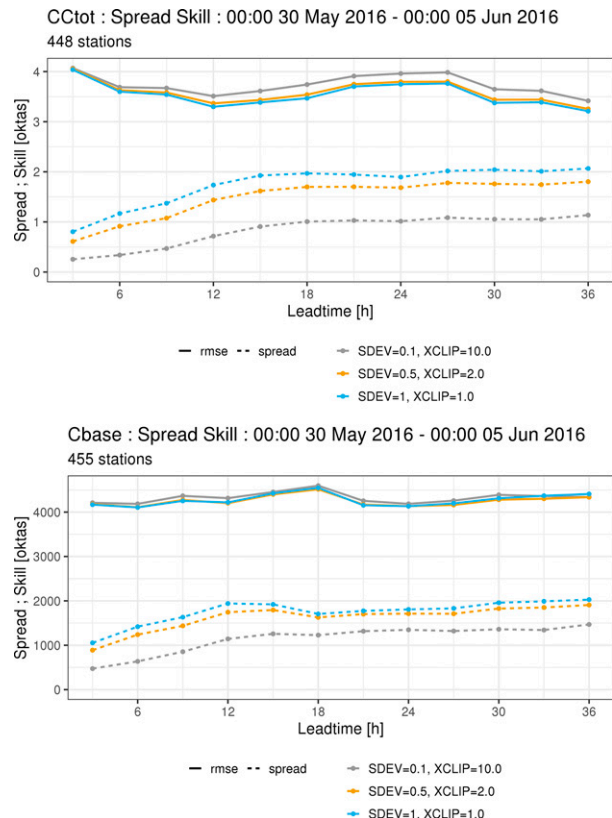


FIG. 2. Spread and skill scores from sensitivity tests for SPPT for (top) total cloud cover (CCtot) and (bottom) cloud base (Cbase) for experiments with varying SDEV (standard deviation of perturbation sizes) and XCLIP (perturbation clipping ratios); SDEV = 0.1, XCLIP = 10.0 (gray), SDEV = 0.5, XCLIP = 2.0 (orange), SDEV = 1.0, and XCLIP = 1.0 (blue). The score differences are calculated with respect to the experiment with SDEV = 1.0. For CCtot the score differences for both the spread and the RMSE are statistically significant at the 99.7% level. For Cbase the score differences for the spread are also at the 99.7% level, while for RMSE the results are mixed.

5 June 2016, and with 6 perturbed members plus one unperturbed member (type iii experiments as defined in [section 3](#)).

Sensitivity tests with varying TAU were first compared in order to find a suitable setting for the time correlation length scale. Settings for XLCOR and SDEV were compared using a similar methodology.

Sensitivity tests for SPPT control parameters XLCOR and TAU did not demonstrate statistically significant results. Despite the range of values used to test both parameters, the difference in spread/skill scores, and indeed all verification scores, was negligible (not shown). [Figure 2](#) displays the impact of SPPT for a sample of tests from the sensitivity experiments undertaken for the SDEV and XCLIP control parameters. The three tests shown represent the low, middle, and high end of the tuning ranges. These sensitivity experiments give a much larger response than those for XLCOR and TAU. Increased standard deviation sizes and reduced clipping ratios lead to statistically significant improvements in

TABLE 2. Summary of SPP parameters (PAR.). Det. is the deterministic value of the parameter, STD#1 is the original standard deviation the sensitivity process was started with, STD#2 is the standard deviation after the sensitivity analysis, and 5 perc. and 95 perc. are the 5th and 95th percentiles, respectively, of the resulting pdf for STD#2, scaled by the deterministic value. Type stands for liquid microphysics (LM), ice microphysics (IM), radiation (RAD), convection (CONV), and turbulence (TURB).

No.	Description	PAR.	Det.	STD#1	STD#2	5 perc.	95 perc.	Type
1	Threshold for cloud thickness used in shallow/deep convection decision	CLDDPTHDP	4000	0.1	0.4	0.07	3.50	CONV
2	Cloud ice content impact on cloud thickness	ICE_CLD_WGT	1	0.1	0.4	0.07	3.50	IM
3	Ice nuclei concentration	ICENU	1	0.35	0.7	0.03	31.6	IM
4	Saturation limit sensitivity for condensation	VSIGQSAT	0.03	0.1	0.4	0.07	3.50	LM
5	Kogan autoconversion speed	KGN_ACON	10	0.25	0.5	0.03	3.81	LM
6	Kogan subgrid-scale (cloud fraction) sensitivity	KGN_SBGR	0.5	0.1	0.2	0.31	2.24	LM
7	Graupel impact on radiation	RADGR	0.5	0.15	0.3	0.15	2.93	RAD
8	Snow impact on radiation	RADSN	0.5	0.15	0.3	0.15	2.93	RAD
9	Top entrainment efficiency	RFAC_TWO_COEF	2	0.1	0.4	0.07	3.50	TURB
10	Stable conditions length scale	RZC_H	0.15	0.1	0.4	0.07	3.50	TURB
11	Asymptotic free atmospheric length scale	RZL_INF	100	0.15	0.6	0.01	3.84	TURB

the spread skill relationships for almost all variables at all lead times (not shown). However, it was discovered that values of SDEV above 0.3 result in the undesirable effect of having a non-Gaussian distribution of perturbation values. This non-Gaussian characteristic arises when clipping is performed, as all values outside the clipping range are added to the tails of the interval. In all further experiments presented here TAU is set to 8 h, XLCOR is set to 200 km, and SDEV and XCLIP are set to 0.3 and 3.33, respectively.

2) SPPT IN FULL EPS SETUP

Although SPPT demonstrated a clear impact when tested separately (see Fig. 2), experiments where SPPT was combined with initial, surface, and lateral boundary perturbations (type i experiments as defined in section 3) only resulted in a small amount of additional variability in the ensemble. This result was seen for the same dates used in the sensitivity experiments (not shown) as well as in a month long experiment covering February 2019 (also seen in Fig. 15 which is discussed later, compare dark blue and gray lines). This conclusion holds for all investigated parameters, also upper air parameters where SPPT is not tapered. This minor impact of SPPT on ensemble spread and skill in HarmonEPS is in contrast to what has been reported at other centers using SPPT (Bouttier et al. 2012; Palmer et al. 2009), and despite the effort to optimize it in HarmonEPS and the promising results when it was tested separately. These results are dissected and discussed further in section 6.

It is clear, however, that the SPPT results motivate the implementation of another method to account for model error in HarmonEPS. In the next section, the implementation and first results of the stochastically perturbed parameterizations (SPP; Ollinaho et al. 2017; Lang et al. 2021) scheme in HarmonEPS are described.

b. SPP

The SPP scheme introduces stochastic perturbations to selected closure parameters in physical parameterizations of a model. The control variables in SPP influencing the magnitude

of the perturbations are (i) the spatial and (ii) temporal correlation lengths of the perturbation patterns, and (iii) the shape of the distribution from which the perturbations are drawn. In HarmonEPS, variables i and ii are controlled using SPG, as for SPPT. The ECMWF implementation of SPP is followed here where the parameter perturbations sample a lognormal distribution (Ollinaho et al. 2017; Lang et al. 2021). The other practical choice would be to sample a normal distribution (Jankov et al. 2017, 2019; Thompson et al. 2021). Following the ECMWF implementation, variable iii is regulated through choosing the width of the distribution (standard deviation) accompanied by a shape choice determining whether the mean or the median of the distribution is equal to the unperturbed closure parameter value.

The current implementation of SPP in HarmonEPS perturbs 12 different parameters, 11 of them are activated and listed in Table 2. The selected parameters are included after advice from physics experts and are not only uncertain in their nature but also active in different processes and under various atmospheric conditions. Parameter 1 works on convection,

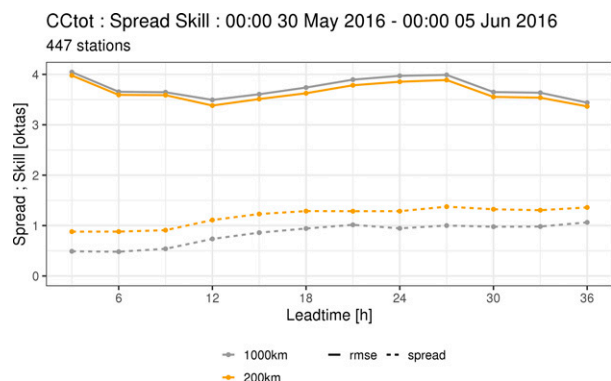


FIG. 3. Spread and skill scores for total cloud cover (CCtot) for one week in spring 2016 showing the sensitivity of changing the spatial scale for the SPP perturbations, in gray for 1000 km and in orange for 200 km. The score differences are statistically significant at the 99.7% level.

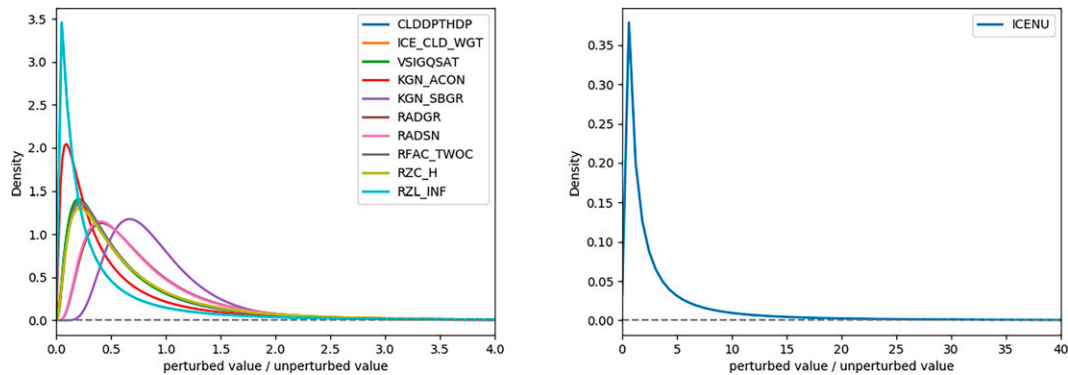


FIG. 4. The pdfs for the parameters perturbed in SPP. Parameters are described in Table 2.

2–6 on condensation, 7 and 8 on cloud affected radiation and the last three on turbulence. SPP perturbations were also introduced for a parameter controlling the threshold for cloud thickness for stratocumulus/cumulus transition (not included in the table). This parameter is involved in choosing the stratocumulus regime in the convection scheme (only moist updrafts and no dry updraft calculations); however, the chance of it becoming active is very small as it depends on a very strong inversion strength threshold in the atmosphere. Since the parameter is only seldom active, it was decided to switch off perturbations to it.

1) SPATIAL AND TEMPORAL SCALES

All parameters are perturbed using the same spatial and temporal scales, but using a unique random seed for each parameter. The spatial and temporal length scales were tested in an early implementation of SPP which included 9 out of the 11 parameters listed in Table 2. Two different spatial length scales were tested, 200 and 1000 km. An example of the effect of changing the spatial scale is shown in Fig. 3. In another test (not shown) a spatial scale of 100 km was tested. This did not give any significant differences from using 200 km. In the following, 200 km is used for all SPP results shown. A range of temporal scales were tested, from 1 h to infinity, with very little sensitivity seen (not shown). In the following a temporal length scale of 12 h is used.

2) PARAMETER DISTRIBUTIONS

Experts were consulted on the different parameterizations and in particular about the range of values the (originally deterministic) parameters could take. The STD#1 value in Table 2 results in approximately this range. All parameters implemented so far follow a lognormal distribution with the mean of the distribution equal to the unperturbed parameter value, as in the latest ECMWF setup (Lang et al. 2021), except for ice nuclei concentration (ICENU) where the median is used. The number concentration of the cloud ice crystals may vary with several orders of magnitude depending on the occurrence of splintering processes. Those are active only under special conditions, e.g., temperatures near -5°C together with high concentrations of cloud liquid and of

graupel. Therefore, a very long tail for ICENU has a physical relevance, and the median was chosen as it resulted in a longer tail than using the mean.

The sensitivity to the width of the distribution (the standard deviation) was examined for each parameter separately. These tests included a summer and a winter period (30 May 2016–5 June 2016 and 20 February 2019–26 February 2019), the summer period being the same as the one used for the SPPT tests. The experiments were run with the same number of ensemble members and the same cycling as for the SPPT experiments, i.e., experiment type iii as defined in section 3. The final parameter distributions were decided based on how the perturbations affected the ensemble spread and also minimized any degradations to the ensemble mean RMSE. For parameters KGN_SBGR, RADGR, and RADSN (see Table 2 for a description of the different parameters), a clipping function was introduced to ensure the parameters were kept within physical bounds. The resulting parameter

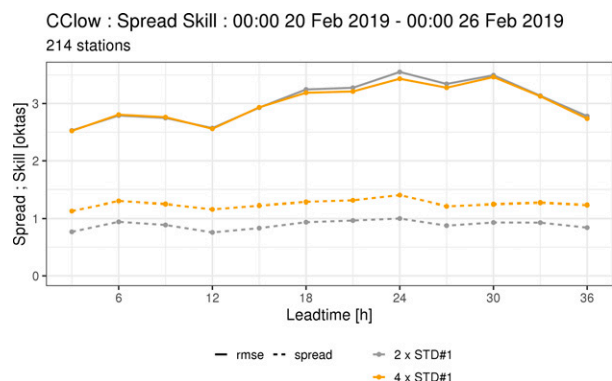


FIG. 5. Spread and skill scores for fraction of low cloud cover (CClow) for one week in February 2019 showing the sensitivity of increasing the standard deviation of the pdf for the SPP parameter saturation limit sensitivity (VSIGQSAT); in gray at 2 times the SDEV and in orange at 4 times the SDEV. The score differences for the spread are statistically significant at the 99.7% level. For the RMSE there is no significant difference for +15 and +33 h, it is significantly worse to increase the SDEV at the 68% level or higher for the first lead times up to +9 h and significantly better to increase the SDEV at the 68% level or higher from +12 h.

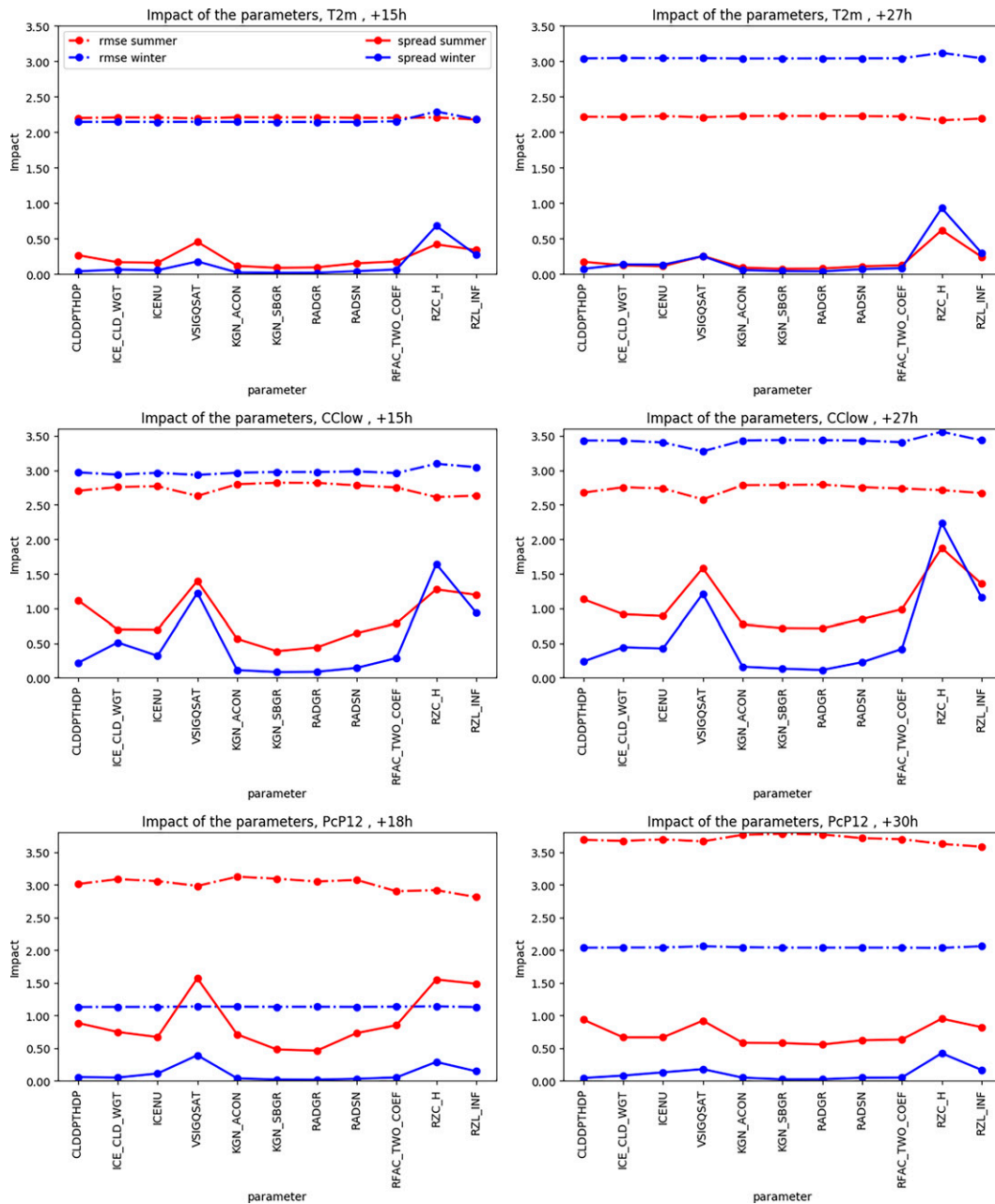


FIG. 6. Spread (solid line) and RMSE of the ensemble mean (dashed line), for summer (red) and winter (blue). The parameters on the x axis correspond to those in Table 2. (top) 2-m temperature (T2m), (middle) fraction of low cloud cover (CClow), and (bottom) 12-h accumulated precipitation (PcP12h). Forecast length is +15 and +27 h for T2m and CClow, and +18 and +30 h for PcP12h.

densities are shown in Fig. 4. As mentioned above, ICENU is the only parameter that uses the median. Its density distribution is quite different from the other parameters, hence it is also plotted separately in the right panel of Fig. 4. The resulting 5th and 95th percentiles of the distributions are shown in Table 2 together with the deterministic values (Det.).

In Fig. 5 spread and skill for low clouds are shown as an example of the sensitivity to the width of the distribution for the saturation limit sensitivity (VSIGOSAT). As seen in Fig. 5 there was a clear positive impact of increasing the width of the probability density function (pdf) for this parameter. For other parameters, only a limited effect was seen. A rather conservative approach of only increasing the standard

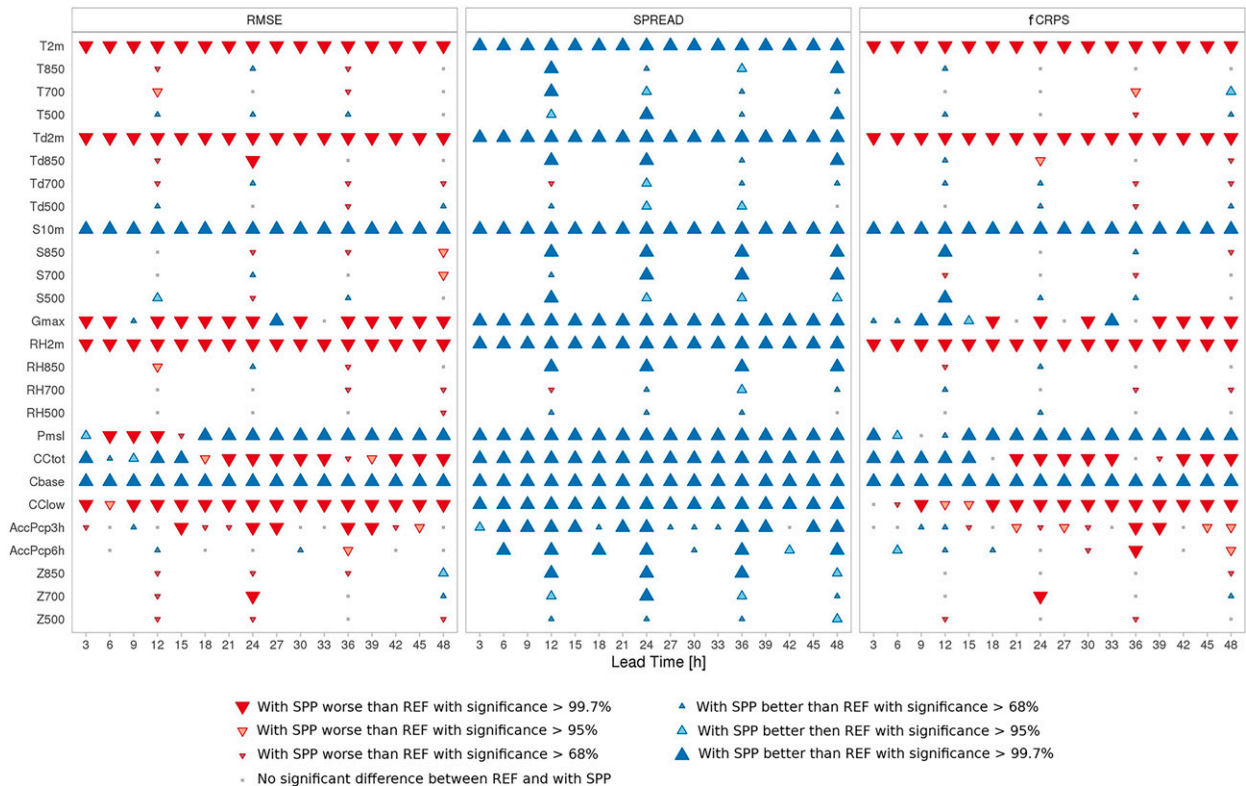


FIG. 7. Scorecards for February 2019, including the statistical significance for the score differences between an experiment with SPP and a reference experiment without SPP (REF). (left) RMSE of the ensemble mean, (center) ensemble spread, and (right) fCRPS. Temperature (T), dewpoint temperature (T_d), wind speed (S), relative humidity (RH), and geopotential (Z). The prevailing numbers indicate the pressure level (850, 700, and 500 hPa) or the height above surface (2 and 10 m). Also shown are, maximum wind gust (Gmax), mean sea level pressure (Pmsl), fraction of total cloud cover (CCtot), cloud base height (Cbase), fraction of low clouds (CClow), and 3- and 6-hourly accumulated precipitation (AccPcp3h and AccPcp6h).

deviation when it had a clear positive impact on the system was chosen.

3) IMPACT OF INDIVIDUALLY PERTURBED PARAMETERS

As expected, perturbing some parameters has a bigger impact than perturbing others. Due to the nature of the perturbed parameters, there is a clear difference in impact for the summer and winter testing periods for some parameters. In Fig. 6 the spread and RMSE from individually perturbing the parameters in Table 2 for two forecast lead times is shown for summer and winter testing periods. The +15- and +27-h forecast lead times represent the maximum and minimum responses seen in the verification for some of the parameters and are connected to the diurnal cycle (all forecasts start at 0000 UTC). These two forecast lead times are shown for T2m and fraction of low cloud cover (CClow) in Fig. 6. For 12-h accumulated precipitation, +18- and +30-h forecast lead times are used, as the 12-h accumulation is not available at +15 and +27 h. The spread and RMSE shown here is with the final parameter distributions (STD#2). The stable condition length scale (RZC_H), the saturation limit sensitivity for condensation (VSIGQSAT, especially in summer), the threshold cloud thickness used in shallow/deep convection

decision (CLDDPTHDP) and the asymptotic free atmospheric length scale (RZL_INF) are clearly the most effective parameters for increasing the spread. It is also quite evident that in winter KGN_ACON, KGN_SBGR, RADGR and RADSN do so to a lesser extent. We can also observe that the spread depends much more on the parameter perturbed than the ensemble mean RMSE does.

4) SPP IN FULL EPS SETUP

After the individual adjustment of the parameter pdf's, SPP was added to the reference setup of HarmonEPS (Frogner et al. 2019) and compared to the reference experiment (REF) (type i experiments as defined in section 3). Two months were tested, one in winter (February 2019) and one in early summer (June 2019). In Figs. 7 and 8 scorecards show the effect of adding SPP in HarmonEPS for a selection of weather parameters in February and June, respectively. There is a clear overall increase in spread when SPP is included, for all parameters and most lead times, for both the summer and winter periods. For RMSE the results are more mixed. There is a significant increase in RMSE for near surface weather parameters like T2m and RH2m from applying SPP. The increased RMSE is more evident in winter than in summer. However, RMSE for

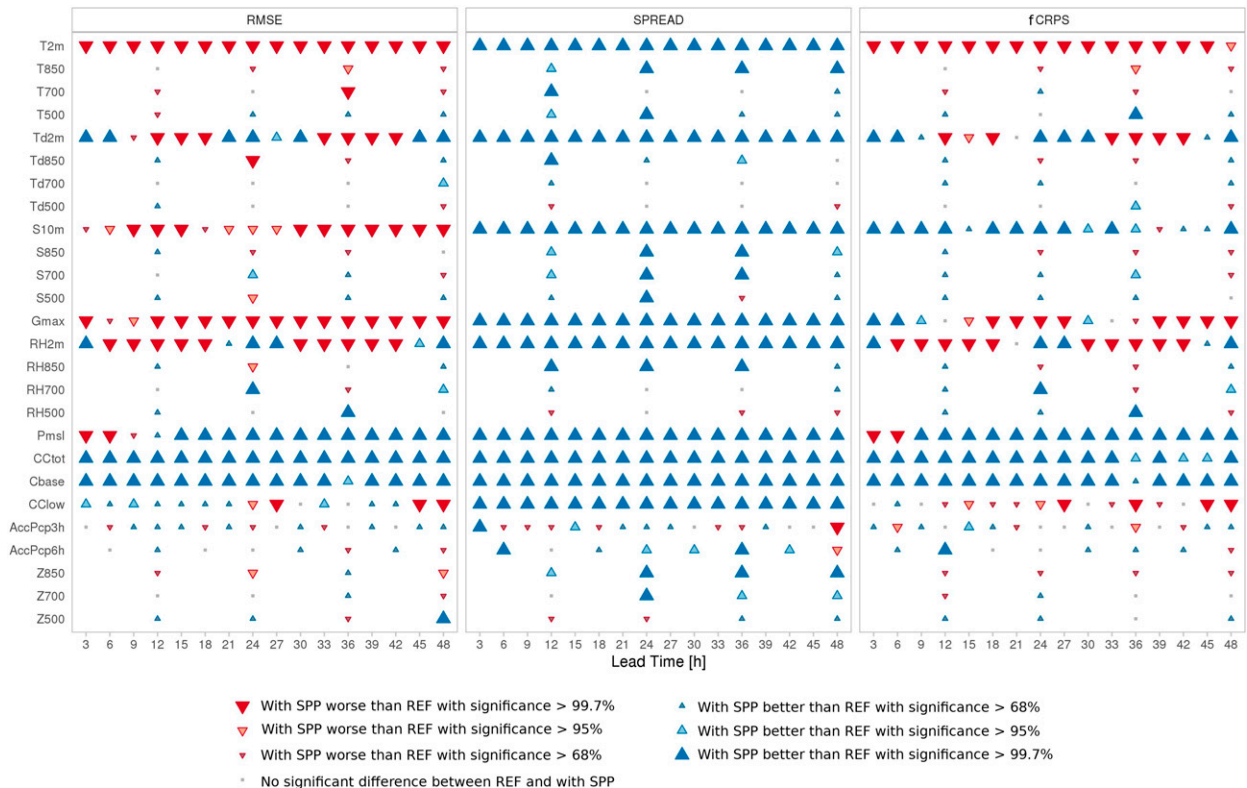


FIG. 8. As in Fig. 7, but for June 2019.

the cloud variables (fraction of total cloud cover (CCtot), cloud base height (Cbase) and fraction of low clouds (CClow)) are mainly improved by adding SPP in summer. For winter the impact is more mixed. fCRPS also shows a mixed effect from including SPP, but there is a significant improvement from SPP for CCtot and Cbase for the summer period.

Recalling what was seen in Fig. 6, parameter RZC_H was the most influential among this set of parameters, and as expected reducing the width of the pdf for RZC_H has a large impact on the ensemble skill (not shown). The parameter has also a clear impact on the mean bias as seen in Fig. 9. It is naturally undesirable that the perturbations change the mean bias of the system (here making the ensemble colder). Reducing the STD#2 value of RZC_H helps, to a large extent, to alleviate this effect. Interestingly, it is also seen that for CCtot the behavior is exactly opposite to that of T2m, with the mean bias becoming closer to the reference with increasing STD#2. Note that other parameter perturbations are active in this test, so this bias change for CCtot might be due to the interaction with other perturbations. Another possible explanation is compensating errors in the forecast model. This will be looked into in a future study in connection with revised pdfs for the parameters.

A case with poorly predicted fog has been selected to illustrate the low cloud/fog-related forecast response of the SPP perturbations. Figure 10 shows a satellite image from 16 February 2019 where widespread areas of fog cover e.g., southern Sweden and Denmark, and some areas of southwestern

Norway and northeastern Finland are covered with scattered fog. In the reference setup (REF), all the perturbed members (Fig. 11) represent the scattered fog quite well, but the larger fog covered areas in Sweden and Denmark are not present in the forecasts at all. In the SPP experiment (Fig. 12), a larger variability between the ensemble members and a tendency for more fog can be seen. The fog predicted in REF is still present, but in addition larger areas of fog in better agreement with the satellite imagery can be found. The larger variability seen in this case is in line with what is seen in the average scores for the cloud parameters in Fig. 7. SPP also increases the average cloud cover for the period (not shown).

For the convective summer cases investigated (not shown), the ensemble sensitivity to the SPP perturbations is less pronounced. In these cases, precipitation areas are redistributed, but without any significant changes in ensemble skill.

A pairwise FSS comparison of total clouds between REF and SPP experiments for February and June 2019 is presented in Fig. 13. SPP performs better than REF in June (right panel) with a relatively larger statistical confidence (FSS is a positively oriented skill score). For February (left panel) there is no statistically significant difference between the two experiments. The differences between the first and third quartiles are larger in February compared to June, meaning the differences between the two ensembles are greater in February. By looking at the median and the first and third quartiles, it can be seen that the distributions of the differences for all scales are left skewed in February (and oriented toward negative

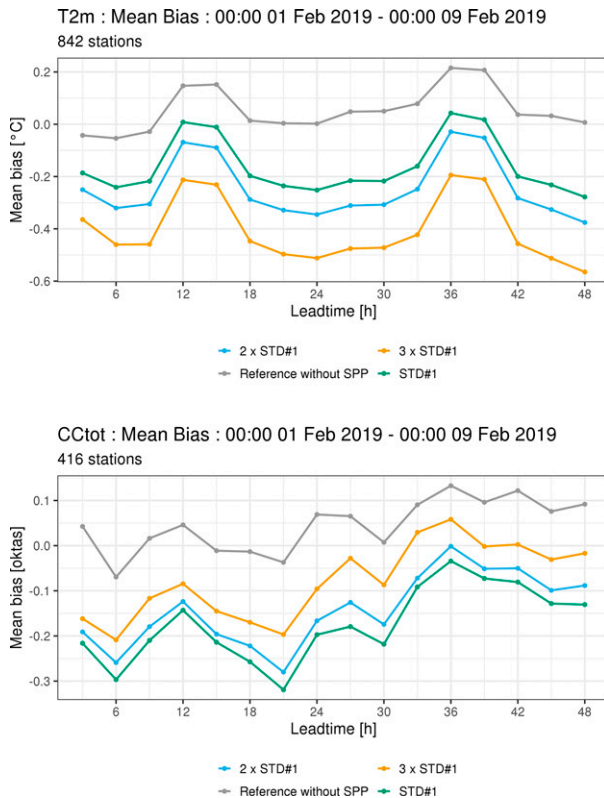


FIG. 9. Mean bias for (top) 2-m temperature (T2m) and (bottom) total cloud cover (CCtot) for a reference HarmonEPS run without SPP (gray), with SPP using STD#1 for RZC_H (green), with SPP and $2 \times$ STD#1 for RZC_H (blue), and with SPP and $3 \times$ STD#1 for RZC_H (orange).

values), indicating that the differences are larger to the left of the median. However, it is difficult to determine which ensemble performs better in February. It can also be observed in Fig. 13 that the difference between SPP and REF does not change significantly across the spatial scales used in this verification.

According to Roberts and Lean (2008) the forecast is considered skillful for scales at which FSS exceeds $FSS_{\text{uniform}} = 0.5 + f_0$, where f_0 is the fraction of cloud-free grid cells. Figure 14 illustrates the difference between FSS and FSS_{uniform} , where f_0 values are calculated from satellite observations for each observation time. A forecast is considered skillful for scales with the difference $\text{diff}_{\text{uniform}} = FSS - FSS_{\text{uniform}}$ larger than 0. Moreover, larger values of the difference correspond to a higher skill. Finding the minimum scale, $\text{scale}_{\text{min}}$, at which FSS exceeds FSS_{uniform} is useful in estimating the forecast skill. The median of $\text{diff}_{\text{uniform}}$ crosses the zero line at spatial scale 32.5 km for both SPP and REF in February (with relatively lower statistical confidence), at spatial scale 12.5 km for SPP and spatial scale 17.5 km for REF in June (with relatively higher statistical confidence). Here, the values are rounded to the nearest scale. Thus, both models achieve FSS_{uniform} at smaller scales in June than in February. By considering the median and the first quartile, SPP achieves FSS_{uniform} at

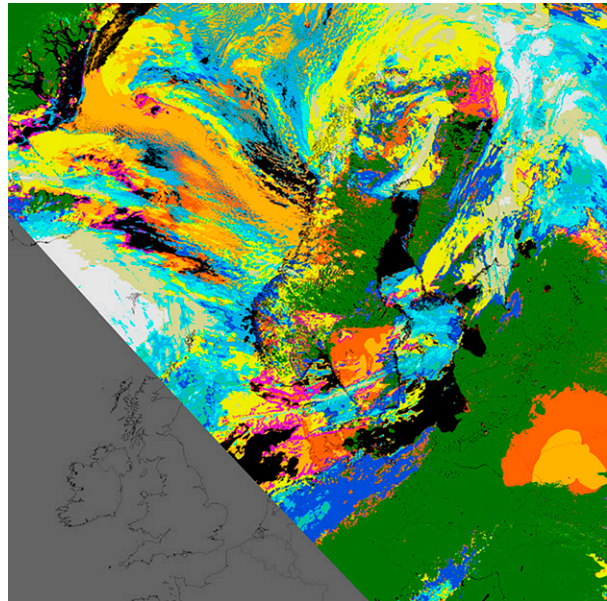


FIG. 10. Satellite picture from NOAA-18 at 1728 UTC 16 Feb 2019. Areas in dark orange can be characterized as fog.

smaller scales more frequently than REF. In February, the two models could be considered the same given the lower statistical confidence.

The FSS skill is also affected by the chosen masking threshold. For example, a threshold of 0.7 slightly reduces the model forecast skill (not shown). This is expected because a higher threshold corresponds to more cloud-free grid cells, while the satellite products used in this study tend to underestimate the number of cloud-free grid cells, as mentioned in section 4. Thus, by increasing the threshold, the forecast cloud mask is expected to be less similar to the satellite-observed cloud mask and therefore FSS is expected to be smaller. In this case, FSS medians for SPP become slightly lower than REF and lower in February than in June. This means that in general SPP produces more cloudy areas, especially in winter. Changing the threshold to 0.7 does not change the statistical confidence of the results considerably when compared to the threshold of 0.2. These changes are also reflected in the $\text{diff}_{\text{uniform}}$ values; however, the changes are small and do not change the overall assessment of the forecast skill.

6. Interactions between perturbation types

As discussed in section 5, SPPT did not produce any significant impact on the ensemble when it was combined with the other perturbation types in HarmonEPS, whereas SPP had a clear positive impact on the ensemble spread (Fig. 15, top panel). A series of experiments was conducted to understand the cause of this lack of impact from SPPT, using experiment setup ii as described in section 3. Due to computational affordability, a subset of the full one-month-long testing period was used. This allowed eight additional experiments to be run in order to get qualitative answers for the lack of

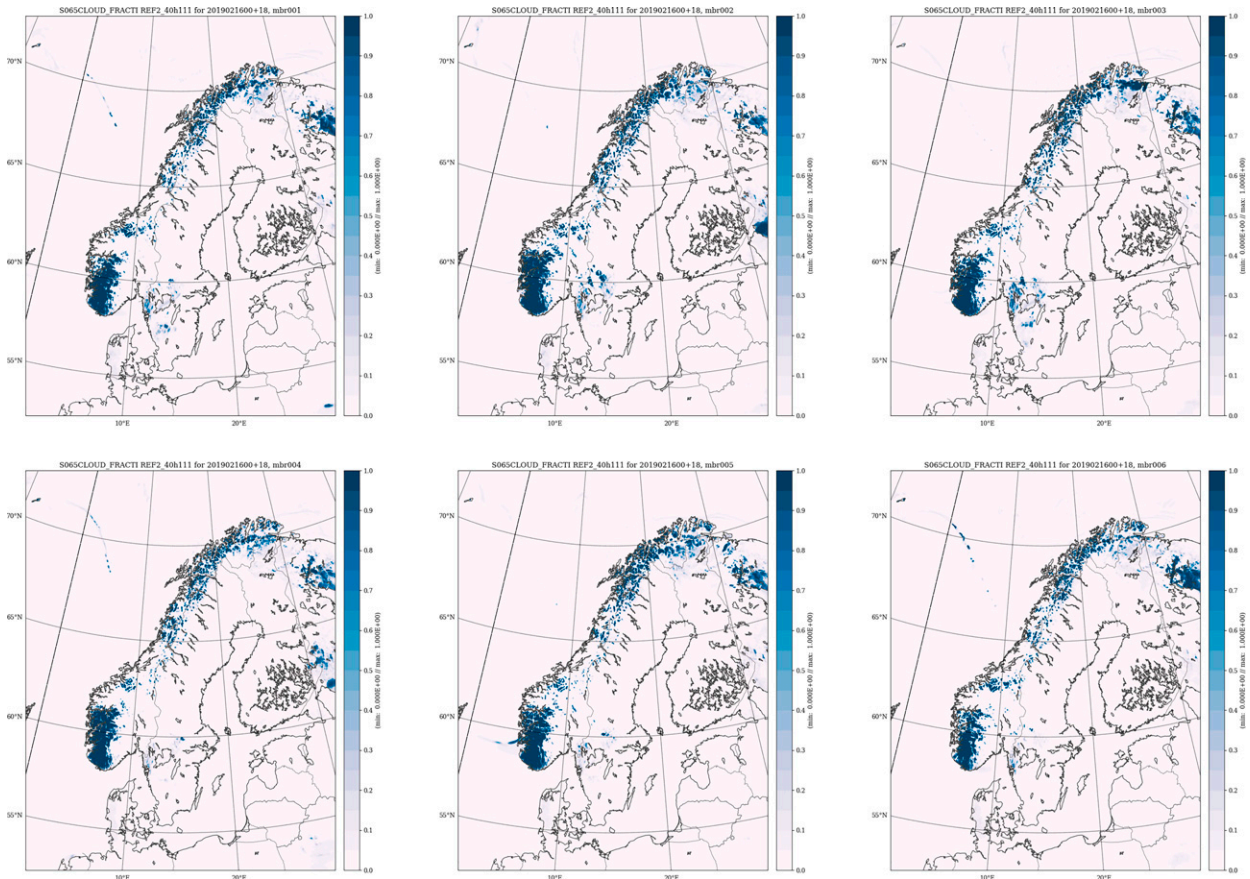


FIG. 11. Fog in reference forecast, members 1–6.

impact from SPPT. A one week period in February 2019 was chosen for this purpose. Obviously, the perturbations in a nonlinear system like an NWP model are not additive, but with the distinct nature of the perturbations that act and focus on different aspects and time ranges of the forecast, it is reasonable to assume they will all contribute to some extent to the variability of the ensemble. Although, the geographical location of the effect will likely be closely linked to the places in the modeled atmosphere where there is sensitivity/instability present. Looking at the effect of all the perturbations individually (Fig. 15, bottom panel) we can see they are all able to create spread. At +18 h the initial and lateral boundary perturbations result in the same amount of spread, while SPPT and surface perturbations result in less and SPP more spread. The experiment with SPP has comparable or higher spread than the other experiments until about +33 h, when the experiment with lateral boundary perturbations begins to dominate. There is a steady increase in the spread emanating from the lateral boundary perturbations, and after about 33 h they are creating the largest spread among the individually active perturbations. This is reasonable, as the perturbations can only be within the large-scale solution inherited from the global model. As noted before, the lateral boundary conditions are not perturbations per se, but rather balanced states

from ECMWF ENS (members from ECMWF ENS). The combination of initial, lateral boundary and surface perturbations (REF, note that this is mainly hidden under the REF + SPPT curve) clearly has higher spread than the three perturbations have individually, as one would expect. Adding SPPT on top has very little effect (curve is almost on top of the experiment REF), while adding SPP has a clearly noticeable effect on the spread. The ensemble mean RMSE (skill) differs less than the spread for the different experiments, which indicates that the different perturbations increase the spread and have less of an effect on the mean. However, it is worth noting that the experiment where SPP is added to REF has the lowest RMSE.

In contrast to the initial and surface perturbations that are applied at only the analysis time over the grid and the lateral boundary perturbations that are applied in a narrow lateral boundary zone, the model uncertainty representations add perturbations at every time step over the grid, and act on the physics parameterizations through perturbing the total physical tendencies (SPPT) or through stochastic perturbations to selected closure parameters in physical parameterizations (SPP). SPPT and SPP are therefore quite different in nature compared to the other perturbations. It is natural to expect that both would thus add variability on top of the other perturbations. This is, however, not the case for SPPT. In section 5

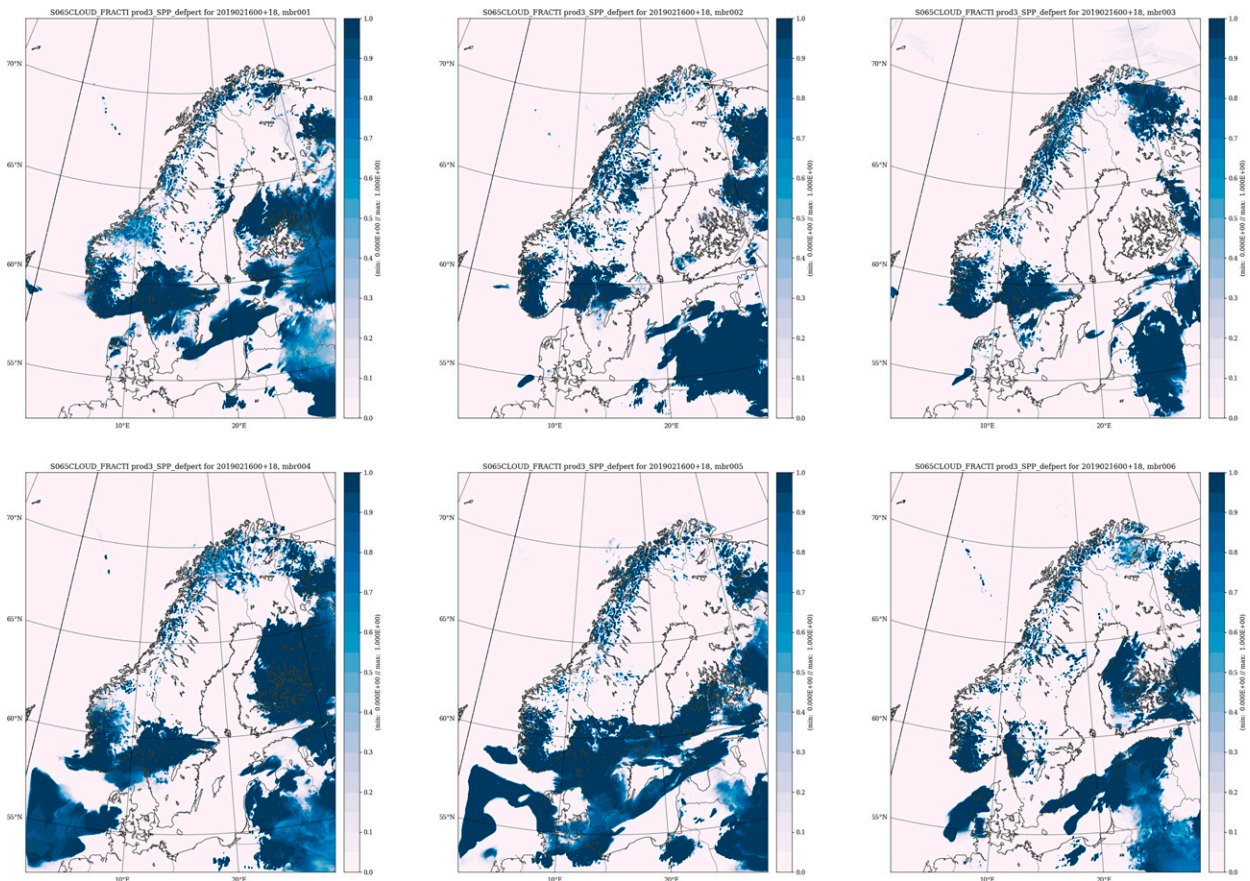


FIG. 12. Fog in SPP forecast, members 1–6.

a clear impact on the spread was seen from SPPT when it was the only perturbation method applied, and the ensemble was sensitive to an increase in the standard deviation of the SPPT perturbations (see Fig. 2). This is confirmed in Fig. 16, where the spread and skill is shown for SPPT experiments with two different values for the standard deviation. While a clear impact on the spread can be seen from increasing the standard deviation when only SPPT is used (top panel), the effect from the same increase is minor when SPPT is combined with initial, lateral boundary and surface perturbations (bottom panel). A standard deviation of 0.3 for SPPT is about the highest one can have in order to maintain the correct histogram shape, as explained in section 5. A standard deviation of 0.9 for SPPT is therefore not recommended, but it is used here to maximize the effect seen on the spread.

The following endeavors to understand why the SPPT perturbations have so little effect in experiments where it is combined with the other perturbations. The 3-h accumulated humidity tendencies from the model physics are investigated for a range of experiments with different combinations of perturbations for several dates and forecast lengths, looking at several model levels and cross sections. Only a single forecast is presented in detail here (0000 UTC 1 February 2019 + 24 h), but the conclusions holds for all cases investigated. In Fig. 17 the

weather situation on 0000 UTC 2 February 2019 is shown together with a cross section used later in the analysis.

In Fig. 18 the difference in ensemble standard deviation of the 3-h accumulated humidity tendencies for the experiments where different standard deviations of 0.9 and 0.3 are used for SPPT is shown. The left panel shows the effect of the increased SPPT standard deviation when only SPPT is active. We again clearly see the effect of the increased SPPT standard deviation, with higher spread especially over the middle part of Norway. The ensemble mean is mainly unchanged by increasing the SPPT perturbations (not shown). The change in ensemble standard deviation over the ocean to the left in the figure has both positive and negative values, hence, the change in SPPT standard deviation is only making a small shift in the ensemble standard deviation in that area. The right panel in Fig. 18 is the same as the left panel, except here SPPT perturbations are introduced on top of the initial, lateral boundary and surface perturbations. In contrast to what was seen for the SPPT only experiments (left panel), the effect seen from increasing the size of the SPPT perturbations is much smaller over the main active area in the middle part of Norway. Again, this is in line with the results in Fig. 16.

The cause for this behavior seems to be that the spread added by SPPT is located in areas where the other

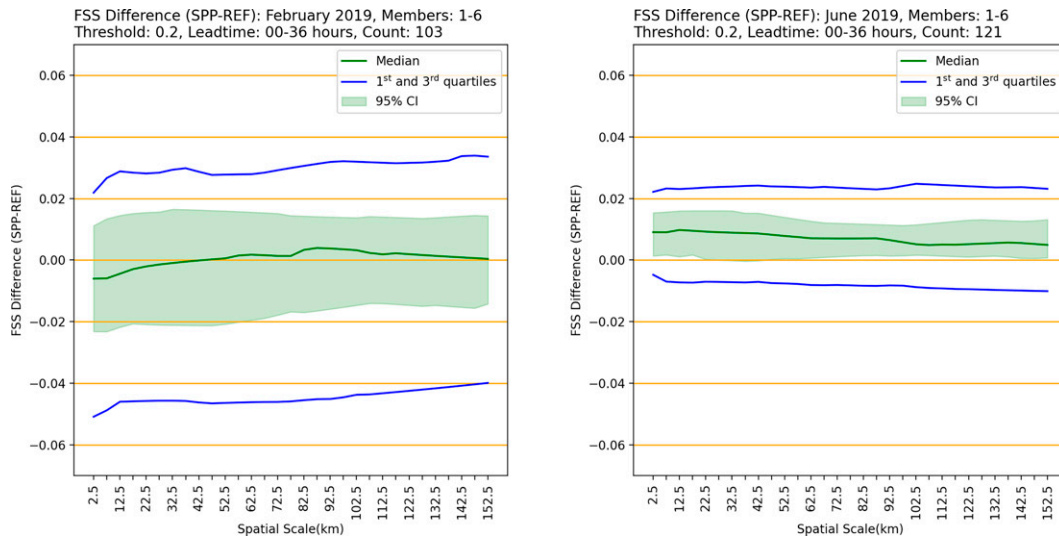


FIG. 13. FSS difference of total clouds between REF and SPP experiments as a function of spatial scale for (left) February and (right) June 2019. Median, first, and third quartiles and the confidence interval of the median (CI) are shown. Ensemble members 1–6, threshold 0.2 and lead times 0–36 h are considered. Count is the number of pairs (forecast valid date, satellite observation date) that match provided that the model domain is covered with more than 80% satellite valid data.

perturbations have already accounted for the variability. It is therefore interesting to see which of the other perturbations masks the effect of SPPT. In the following, the effect that SPPT has on top of the initial, lateral boundary and surface perturbations is looked at separately. Figure 19 shows the spread and skill for these perturbation type combinations. The curves with/without SPPT for the different perturbations are almost on top of each other, except for the surface perturbations, i.e., both the initial and lateral boundary perturbations mask the effect of SPPT. To study this in more depth we again look at the standard deviation of 3-h accumulated humidity tendencies, Fig. 20, now for the experiment with only SPPT (left), only initial perturbations (center) and for the experiment where SPPT is included in addition to the initial perturbations (right). The areas where SPPT creates variability coincides with the areas where the initial conditions create variability, and in the case where both initial and SPPT perturbations are active, hardly any difference is seen from the case with only initial perturbations. The geographical areas where SPPT tries to add variability are thus in the same locations as the variability generated by the initial condition perturbations, and very little extra is introduced by SPPT.

Similarly, the effect of SPPT on top of the lateral boundary perturbations was investigated. A cross section along the line in Fig. 17 is shown in Fig. 21. As for the initial perturbations, SPPT perturbations are also clearly masked by the lateral boundary perturbations throughout the atmospheric column. The geographical areas where SPPT tries to add variability are the same areas as the variability generated by the lateral boundary perturbations, explaining why SPPT adds very little to the spread of the ensemble as seen in Fig. 15, bottom panel.

There is some effect of SPPT on top of the surface perturbations, as it adds variability in the middle part of Norway

where surface perturbations contribute very little (not shown). However, the increased variability from SPPT on top of the surface perturbations is in the places where initial perturbations are already active (not shown).

In contrast to SPPT, SPP adds variability in the ensemble when acting alone, but also when added to the other perturbations in the ensemble (see Fig. 15). It is interesting to see if this is due to perturbations being introduced in other geographical areas (or in other weather situations), if it is an amplification of the spread already created by the other perturbations, or a combination. In Fig. 22 the standard deviations of the 3-h accumulated humidity tendencies are shown applying a mask that identifies in what geographical areas the perturbations are active. The mask is set individually for each perturbation type (each panel) by dividing the maximum value in the plot by 20, hence the size of the perturbations are irrelevant in this figure, favoring the perturbations which are small (SPPT mainly). While mostly the same areas show up for all the perturbation types, SPP does show active areas where the others do not, e.g., along the southwest coast of Norway, along the southwestern part of Finland, and in areas of southern Sweden. SPP therefore seems to be capable of adding variability that is not captured by the other perturbations.

7. Discussion and conclusions

SPP in this first HarmonEPS configuration is a promising scheme for including a model uncertainty representation in HarmonEPS. The main motivation for finding an effective model uncertainty representation for HarmonEPS was the current lack of variability in cloud products. It has been demonstrated in this study, through the impact on general skill scores and a case study, that SPP is able to accomplish this.

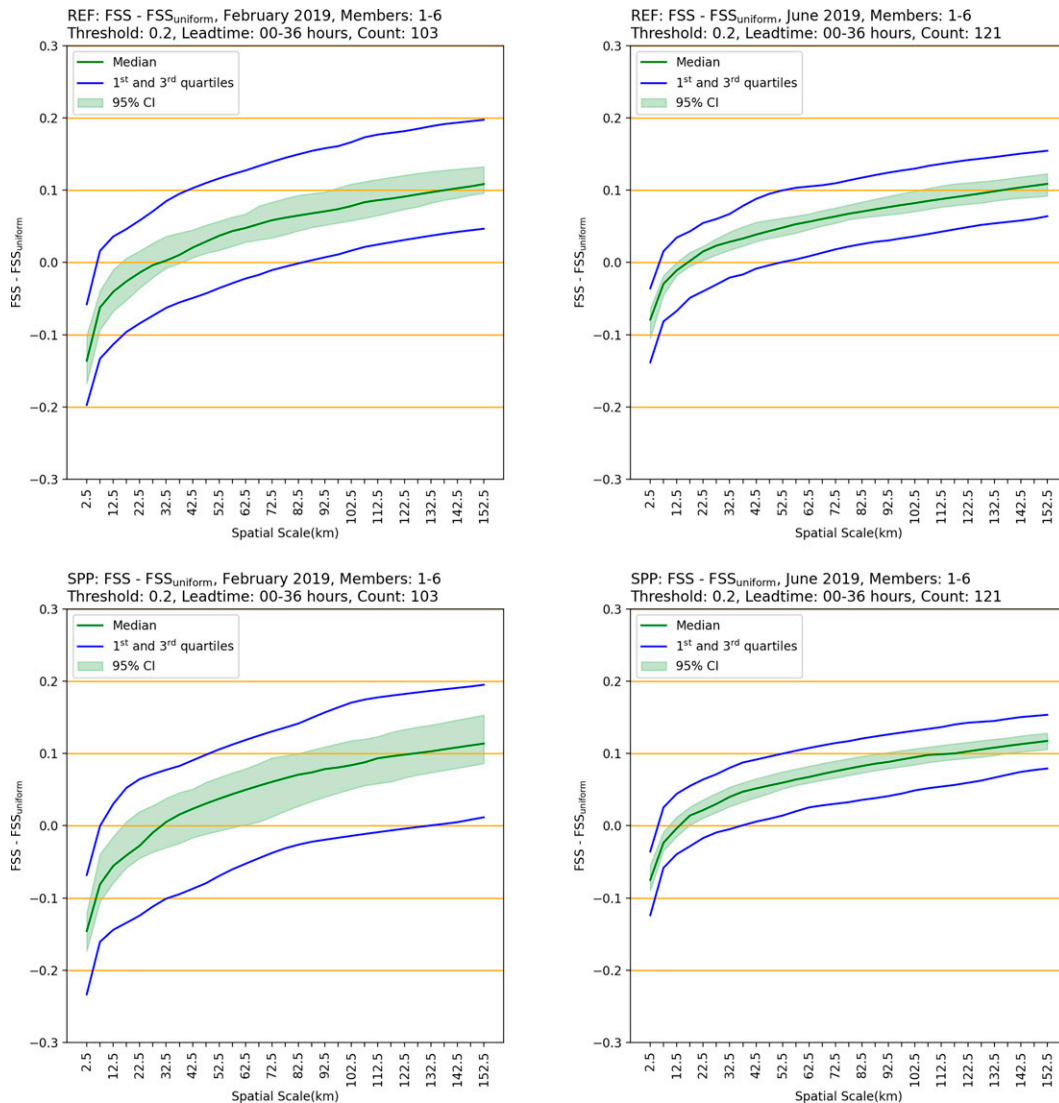


FIG. 14. Difference between FSS and FSS_{uniform} for (top) REF and (bottom) SPP as a function of spatial scale for (left) February and (right) June 2019. All other figure information is as in Fig. 13.

Interestingly, it has been demonstrated that SPP is able to add variability in geographical areas where the other perturbations are not active. The case study also illustrated SPP's ability to capture a fog event that was missed by a forecast without SPP. McCabe et al. (2016) also demonstrated that perturbing parameters with the random parameter (RP) scheme in the Met Office's convection-permitting EPS for the United Kingdom (MOGREPS-U.K.) enabled the EPS to capture observed fog events that were otherwise missed. Moreover, SPP is able to add to HarmonEPS' variability much more than SPPT. SPP is also in line with the objective of physically consistent perturbations as it does not violate local conservation properties of energy and moisture. Although SPP was able to improve the variability of the ensemble, in some cases a degradation of the ensemble mean RMSE was observed. Moreover, perturbations of some parameters affected the mean bias

of the model, especially during the cold season. In particular, three parameters were found to be more active in winter (VSIGQSAT, RZC_H, and RZL_INF). While these parameters are also active in summer, the relative impact of all parameters is more even in summer (see Fig. 6). Reducing the standard deviation for RZC_H perturbations was seen to help in reducing the bias change with respect to T2m, but had the opposite effect on the cloud variables. Some of the perturbed parameters are involved in the same processes and are thus likely influenced by each other. So far this has not been taken into account, as each perturbed parameter has its own realization of the random perturbation pattern. In theory, this could possibly result in perturbations of one parameter working against perturbations of another parameter. This in turn could lead to too large perturbations of the individual parameters to get the desired effect on the variability, as well as perturbations simply

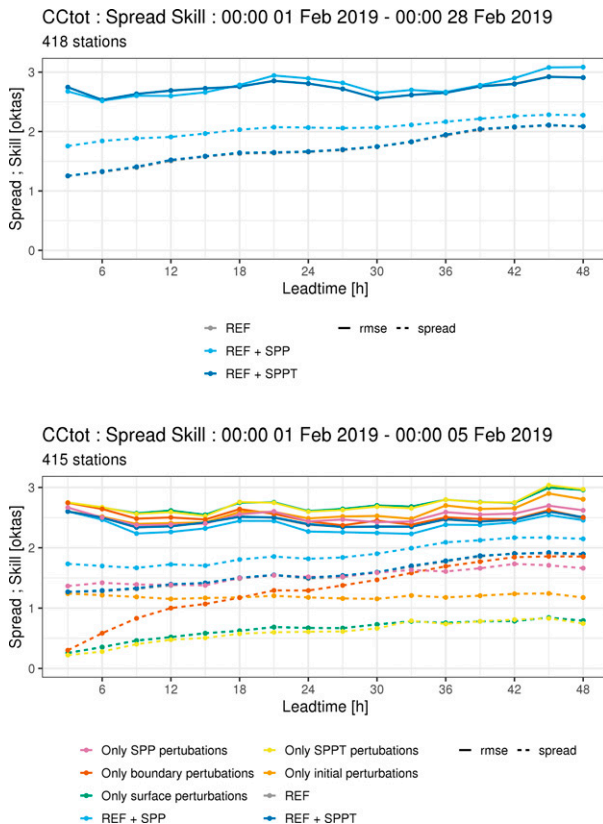


FIG. 15. Spread and skill scores for (top) total cloud cover (CCtot) for February 2019 and (bottom) the one week testing period in February 2019. SPP perturbations (pink), SPPT perturbations (yellow), lateral boundary influence (brown), initial perturbations (orange), surface perturbations (green), boundary, surface and initial perturbations combined (called REF, gray), REF + SPP (light blue), and REF + SPPT (dark blue). Note that REF is partly hidden under REF + SPPT. Statistical significance is calculated for each experiment with respect to REF. The score differences for the spread are all statistically significant at the 99.7% level, except for experiment REF + SPPT, which is statistically significant at the 68% level or higher for the top figure, and statistically significant at the 95% level up to +9 h and from +33 to +48 h, and not statistically significant from +12 to +30 h. For the RMSE the results of the significance tests are mixed.

cancelling each other out. The impact of the correlation of the perturbation patterns is currently being studied, starting with RZC_H and RZL_INF. If this proves successful, this could further improve the SPP scheme and might also help in reducing the above mentioned ensemble mean RMSE/bias change. Although taking possible parameter correlations into account further adds to the maintenance of the scheme, it might be possible to utilize some algorithmic parameter estimation methods (see e.g., Ollinaho et al. 2013, 2014) to inform about these correlations.

SPPT in HarmonEPS is not performing as well as reported in other EPSs, including other convection-permitting EPSs (e.g., Bouttier et al. 2012). Increasing the size of the SPPT perturbations only resulted in a minor change in ensemble

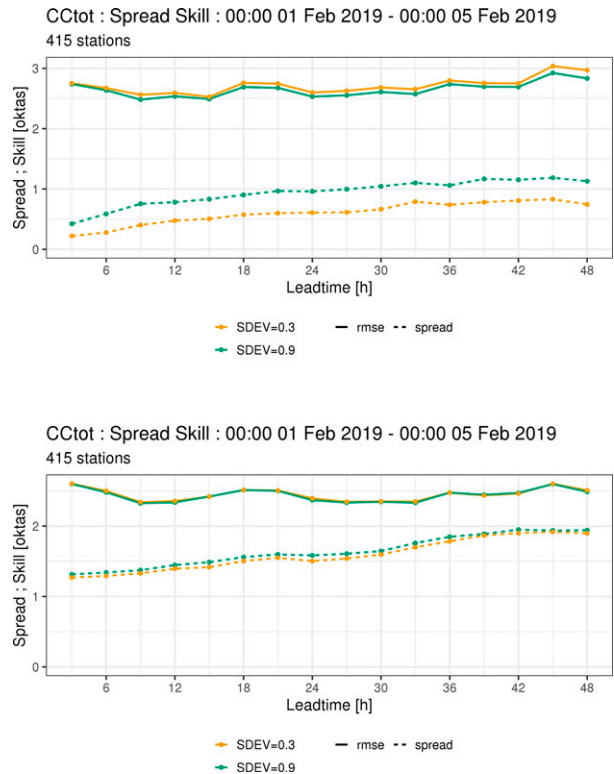


FIG. 16. Spread and skill scores for total cloud cover (CCtot) for the one week testing period in February 2019. SPPT with standard deviation for the perturbation of 0.3 (orange) and with standard deviation for the perturbation of 0.9 (green). (top) Only SPPT perturbations, and (bottom) all other perturbations in addition to SPPT. For the top panel the score differences are statistically significant at the 99.7% level both for the ensemble spread and for RMSE, except for RMSE for +3 h where it is significant at the 68% level. For the bottom panel the score differences are statistically significant at the 68% level or higher for the ensemble spread. For RMSE there is no significant difference from increasing the standard deviation for +27, +30 and +48 h, it is statistically worse at the 68% level or higher for +15, +36, +39, and +42 h and better at the 68% level or higher for the remaining forecast lead times.

variability when SPPT was combined with the other perturbations used in this study. Similar negligible effects on the ensemble variability have been seen when SPPT was tested on a domain over Ireland (not shown). One possible explanation for this reduced impact from SPPT compared to other studies could be the relative magnitude of SPPT perturbations compared to the initial and lateral boundary perturbations (e.g., see Fig. 15, bottom panel). HarmonEPS is known to have a larger initial spread compared to e.g., ECMWF ENS, as reported in Frogner et al. (2019). Looking at the tendencies, it was seen that SPPT was only able to create variability in the same geographical areas as the other perturbations, at least for the cases which have been investigated in this study. Disentangling the different perturbations in HarmonEPS showed that all of the other perturbations are taking part in

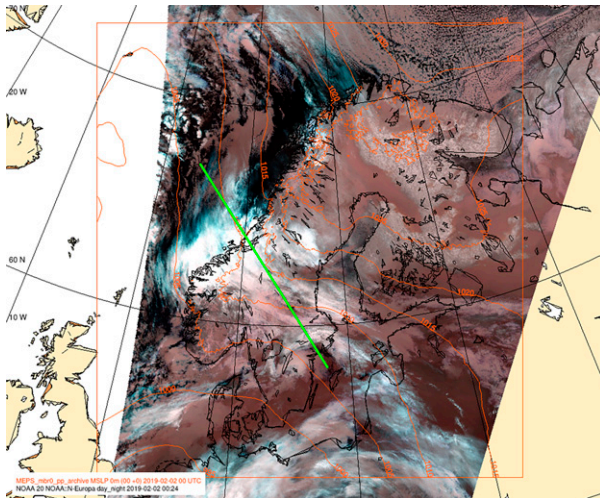


FIG. 17. Weather situation at 0000 UTC 2 Feb 2019. Satellite picture from NOAA. In orange is the MSLP analysis from operational model. The green line shows the position of the cross section used (67°N , 4°E), (58°N , 18°E).

masking the effect of SPPT; although, the surface perturbations do so to a lesser extent than the initial and boundary perturbations.

The time periods used in this investigation may also have an impact on the response from SPPT. Previous studies have reported that the impact of physics perturbations is quite case dependent with a particular dependence on synoptic forcing [greater response to physics perturbations in weakly forced cases e.g., Hally et al. (2014), Keil et al. (2014)]. More recently, Keil et al. (2019) demonstrated that hourly

precipitation rates illustrate a clear weather regime dependence, with spatial variability increased during weak forcing. While such investigations are beyond the scope of the current study, the minimal effect of SPPT illustrated in this study could be related to strong synoptic forcing over the Scandinavian region during February 2019 (images not shown illustrate some features consistent with strong upper-level forcing). Despite the potential imbalances in strength between the initial, lateral boundary and SPPT perturbations, SPP is able to create variability in areas where the other perturbations do not, and also produce more useful forecasts seen from e.g., the resolution component of the Brier score and the area under the relative operating characteristic (ROC) curve (not shown). Unlike SPPT, where the perturbations are zero in cases where the total tendency is zero, SPP acts on individual processes through perturbing the closure parameters in the physical parameterizations and can trigger new states even in such situations.

One option for trying to improve SPPT in HarmonEPS is pSPPT (Wastl et al. 2019b), where the partial tendencies of the physics parameterization schemes are sequentially perturbed. This was also shown to improve the numerical stability of SPPT, making it possible to switch off the tapering in the boundary layer for SPPT for all parameterizations, except for the turbulence scheme. Applying pSPPT has the potential to create more variability near the surface than seen in the standard SPPT. pSPPT also results in a more physically consistent scheme, as the interaction between the uncertainties of the different physics parameterization schemes is sustained.

Model error is complex and originates from many sources, e.g., unresolved processes at subgrid scale, simplified process description, incomplete knowledge of processes and uncertain closure parameters in the parameterizations. Debate exists

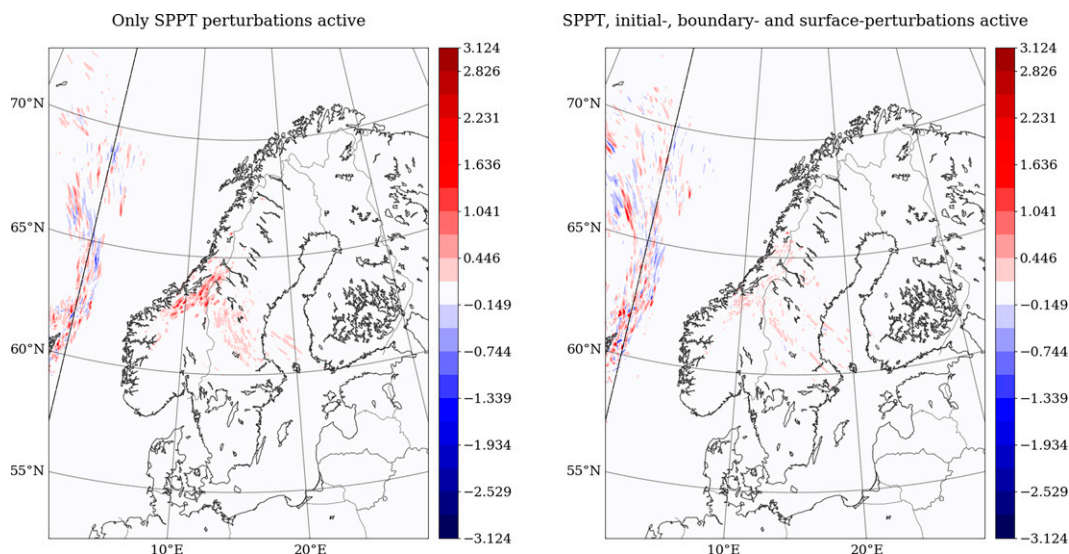


FIG. 18. Difference in ensemble standard deviation for two experiments where the SPPT standard deviation is 0.9 and 0.3. For 3-h accumulated specific humidity tendencies for 24-h forecast from 0000 UTC 1 Feb 2019 for level 28 (600 hPa). (left) For the experiments with only SPPT and (right) for the experiments with all other perturbations on in addition to SPPT. The values are scaled by 1.00×10^5 .

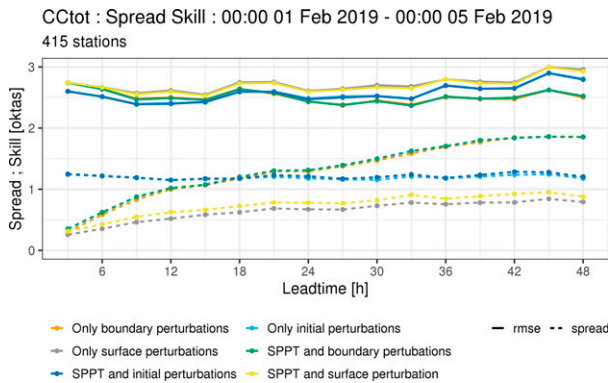


FIG. 19. Spread and skill scores for total cloud cover (CCtot) for the one week testing period in February 2019. Experiment with only initial perturbations (light blue), experiment with initial perturbations and SPPT (dark blue), experiment with only lateral boundary perturbations (orange), experiment with lateral boundary perturbations and SPPT (green), experiment with only surface perturbations (gray), and experiment with surface perturbations and SPPT (yellow). Even though the experiments without SPPT are almost invisible under the SPPT ones, the spread is significantly increased by applying SPPT at the 68% level or higher for all experiments, except for a few forecast lengths where no significant difference is found. For the RMSE the results of the significance tests are mixed.

over whether or not a single model uncertainty scheme is sufficient and that perhaps a combination of schemes is needed to account for the full model uncertainty present. For example, Jankov et al. (2019) argue that SPP on its own in their 3-km ensemble is not sufficient, and that combining it with SPPT is necessary. This is in contrast to the results presented in this paper where SPP is seen to contribute considerably to the variability of the ensemble, while activating SPPT adds very little. Jankov et al. (2019) applies SPP only to the planetary boundary layer (PBL) scheme and Thompson et al. (2021) only for a few microphysics parameters, and this could

be part of the explanation why a reduced effect from SPP is seen. Interestingly, Lang et al. (2021) show a revised version of SPP in ECMWF ENS that is as skillful as SPPT, while the first version was not (Ollinaho et al. 2017). Clearly, the additional probabilistic skill provided by SPP is, among other things, tied to the quantity and quality of the parameters perturbed.

The impact of SPP reported in this study, in Jankov et al. (2019), and in the two versions of ECMWF SPP (Ollinaho et al. 2017; Lang et al. 2021), highlights that the actual setup of the SPP scheme is important. This includes targeting influential parameters important in different weather situations, the shape of the parameter pdfs, the influence of the spatial and temporal scales used, and possible correlations between the parameters. An obvious improvement to the current SPP implementation in HarmonEPS is to increase the number of perturbed parameters, and also to extend perturbations to better cover the different parts of the model physics. Perturbations to the semi-Lagrangian horizontal diffusion will be added and investigated in a future study. Also the choice of spatial and temporal length scales will be revisited in a future study, possibly with different scales for different parameters. SPG, as used here, has the possibility to be extended to three dimensions, which might be important for some processes. For this to be computationally affordable in an operational setting, further work is needed to decrease the cost of the generation of the random fields and in optimizing how often they are applied. Currently, perturbing all 11 parameters used in this study at the same time gives rise to a 5% increase in computer resources compared to a run without SPP when the pattern is updated every time step (in 2D), and 0.1% when updated every hour. Carefully choosing parameters that are proven to be influential will also be important for the affordability of SPP. As seen in Fig. 6 some of the parameters have little impact in both seasons, and further studies will illustrate if these parameters can be excluded or if they prove to be important in certain situations.

One drawback of SPP compared to SPPT is the issue of maintenance. While SPPT requires very little maintenance,

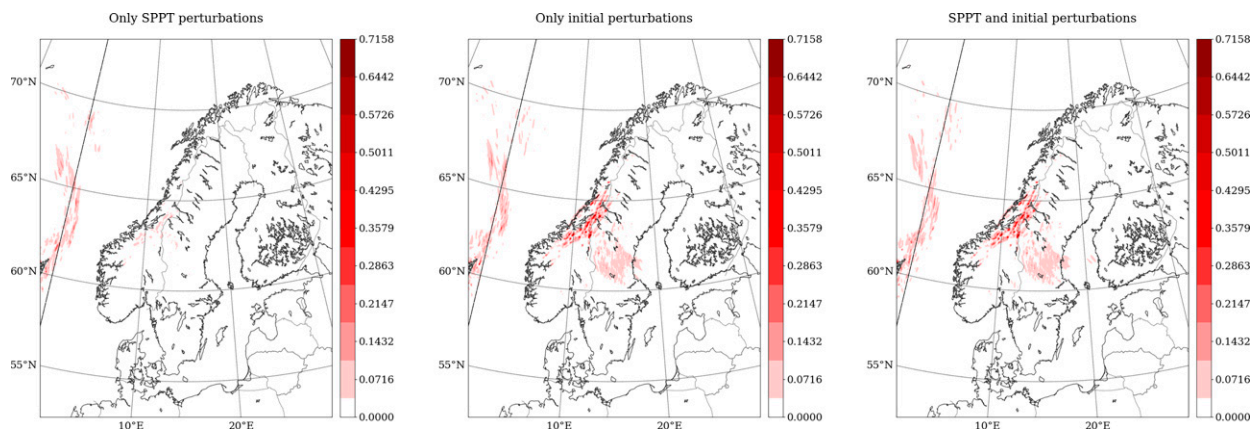


FIG. 20. Standard deviation for 3-h accumulated specific humidity tendencies for 24-h forecast from 0000 UTC 1 Feb 2019 for level 28 (600 hPa) for the experiments with (left) only SPPT, (center) only initial perturbation, and (right) initial and SPPT perturbations. The values are scaled by 1.00×10^4 .

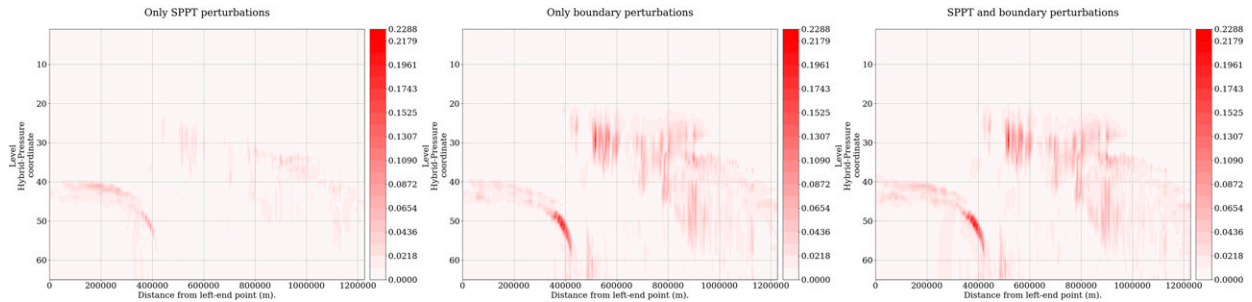


FIG. 21. Standard deviation for 3-h accumulated specific humidity tendencies for 24-h forecast from 0000 UTC 1 Feb 2019 for the cross section seen in Fig. 17 for the experiments with (left) only SPPT perturbations, (center) only lateral boundary perturbations, and (right) lateral boundary and SPPT perturbations. The values are scaled by 1.00×10^4 .

SPP needs reassessment and adjustments when new physics are developed. However, Lang et al. (2021) argue that the conservation properties of SPP make it nonetheless an attractive option over SPPT. Currently physics developments are based mainly on deterministic experimentation. It will be important for optimal use of resources, as well as optimal

uncertainty representation, that in the future stochasticity is taken into account at an early stage of physics development.

SPP works well for the convection-permitting ensemble tested here, even when perturbing so few parameters involved in a rather limited set of physical parameterizations and processes within them. An ensemble size of 6 + 1 members as

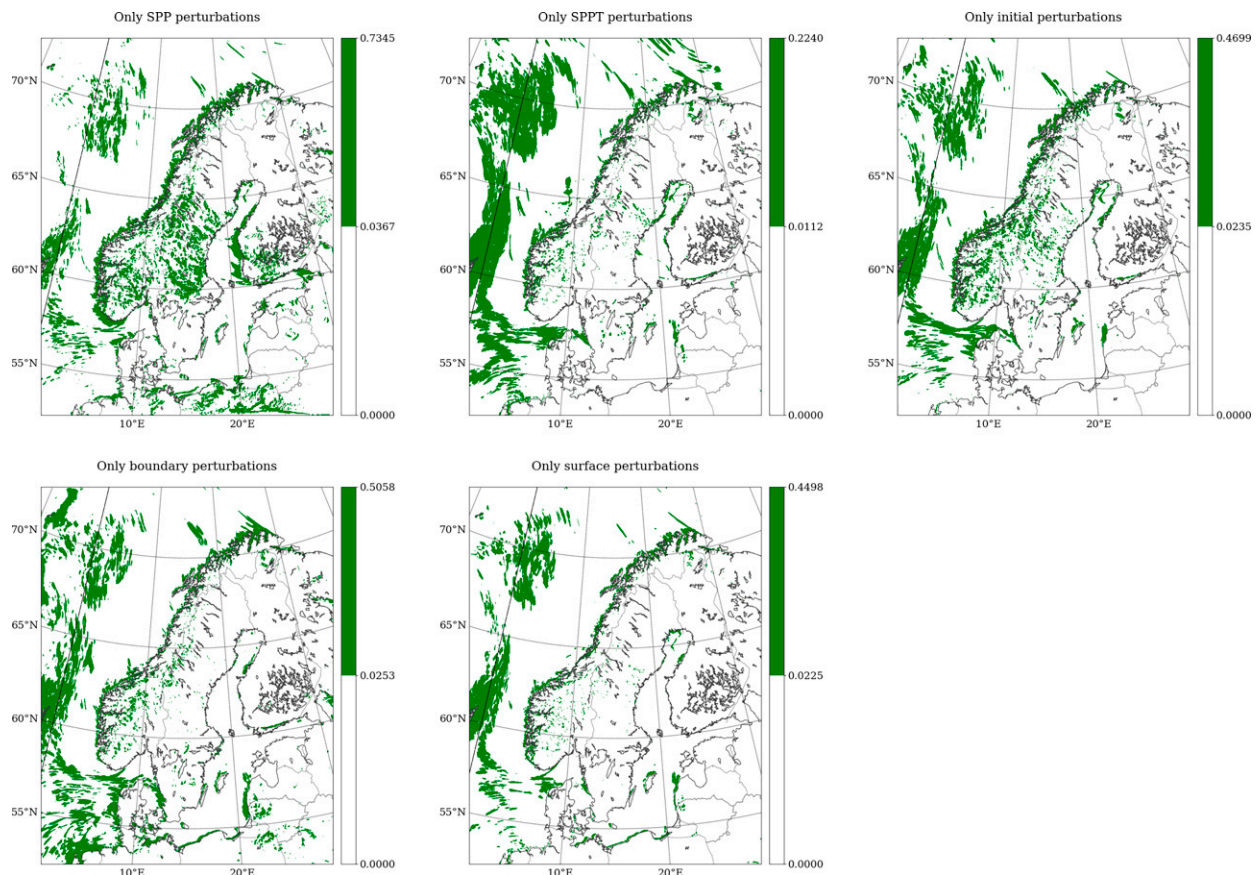


FIG. 22. Standard deviation for 3-h accumulated specific humidity tendencies for 24-h forecast from 0000 UTC 1 Feb 2019 for model level 61. The scaling in the plots is constructed to highlight the areas where the different perturbations add variability, with the transition from white to green equaling the maximum value in each plot divided by 20. (top left) Only SPP perturbations, (top center) only SPPT perturbations, (top right) only initial perturbations, (bottom left) only lateral boundary perturbations, and (bottom center) only surface perturbations. The values are scaled by 1.00×10^4 .

used here, and experiment periods of two weeks for the sensitivity of SPP parameters and two months for testing SPP with respect to REF, are obviously not enough to adequately sample the full probability distribution of atmospheric states. Longer experiment periods that include a wider variety of differently forced atmospheric situations and large ensemble sizes would be desirable to confirm the results presented in this paper. However, we are confident, based on tests comparing our 6 + 1 member ensemble with a 20 + 1 member ensemble and the use of fair CRPS, that our results are a good foundation for further development of SPP in HarmonEPS. Further work will first focus on getting SPP ready for operational implementation in HarmonEPS suites (Frogner et al. 2019) by adjusting the pdfs for the already implemented parameters, while also taking into account the correlations of some of the parameters, and paying special attention to possible undesirable bias changes.

Acknowledgments. This study was conducted as part of the HIRLAM program (<http://hirlam.org/>). Some of the experiments referred to in this paper were run with computer resources provided by Special Projects from the ECMWF. This work was also supported by Academy of Finland (Grant 316939). Special thanks go to Andrew Singleton for his help in adapting the verification system to our needs and to three anonymous reviewers for valuable comments and suggestions.

Data availability statement. All model data produced during this study have been archived locally and are available upon request to the corresponding author.

REFERENCES

- Bengtsson, L., and Coauthors, 2017: The HARMONIE-AROME model configuration in the ALADIN-HIRLAM NWP system. *Mon. Wea. Rev.*, **145**, 1919–1935, <https://doi.org/10.1175/MWR-D-16-0417.1>.
- Bernabò, P., F. Cuccoli, and L. Baldini, 2015: Icing hazard for civil aviation. *2015 IEEE Metrology for Aerospace (MetroAeroSpace)*, Benevento, Italy, IEEE, 295–300, <https://doi.org/10.1109/MetroAeroSpace.2015.7180671>.
- Bouttier, F., B. Vié, O. Nuissier, and L. Raynaud, 2012: Impact of stochastic physics in a convection-permitting ensemble. *Mon. Wea. Rev.*, **140**, 3706–3721, <https://doi.org/10.1175/MWR-D-12-00031.1>.
- , L. Raynaud, O. Nuissier, and B. Ménétrier, 2016: Sensitivity of the AROME ensemble to initial and surface perturbations during HyMeX. *Quart. J. Roy. Meteor. Soc.*, **142**, 390–403, <https://doi.org/10.1002/qj.2622>.
- Brousseau, P., L. Berre, F. Bouttier, and G. Desroziers, 2011: Background-error covariances for a convective-scale data-assimilation system: AROME–France 3D-Var. *Quart. J. Roy. Meteor. Soc.*, **137**, 409–422, <https://doi.org/10.1002/qj.750>.
- Clark, A. J., and Coauthors, 2011: Probabilistic precipitation forecast skill as a function of ensemble size and spatial scale in a convection-allowing ensemble. *Mon. Wea. Rev.*, **139**, 1410–1418, <https://doi.org/10.1175/2010MWR3624.1>.
- Crocker, R., and M. Mittermaier, 2013: Exploratory use of a satellite cloud mask to verify NWP models. *Meteor. Appl.*, **20**, 197–205, <https://doi.org/10.1002/met.1384>.
- Frogner, I.-L., and Coauthors, 2019: HarmonEPS—The HARMONIE ensemble prediction system. *Wea. Forecasting*, **34**, 1909–1937, <https://doi.org/10.1175/WAF-D-19-0030.1>.
- Giard, D., and E. Bazile, 2000: Implementation of a new assimilation scheme for soil and surface variables in a global NWP model. *Mon. Wea. Rev.*, **128**, 997–1015, [https://doi.org/10.1175/1520-0493\(2000\)128<0997:IOANAS>2.0.CO;2](https://doi.org/10.1175/1520-0493(2000)128<0997:IOANAS>2.0.CO;2).
- Hally, A., E. Richard, and V. Ducrocq, 2014: An ensemble study of HyMeX IOP6 and IOP7a: Sensitivity to physical and initial and boundary condition uncertainties. *Nat. Hazards Earth Syst. Sci.*, **14**, 1071–1084, <https://doi.org/10.5194/nhess-14-1071-2014>.
- Jankov, I., and Coauthors, 2017: A performance comparison between multiphysics and stochastic approaches within a North American RAP ensemble. *Mon. Wea. Rev.*, **145**, 1161–1179, <https://doi.org/10.1175/MWR-D-16-0160.1>.
- , J. Beck, J. Wolff, M. Harrold, J. B. Olson, T. Smirnova, C. Alexander, and J. Berner, 2019: Stochastically perturbed parameterizations in an HRRR-based ensemble. *Mon. Wea. Rev.*, **147**, 153–173, <https://doi.org/10.1175/MWR-D-18-0092.1>.
- Keil, C., F. Heinlein, and G. C. Craig, 2014: The convective adjustment time-scale as indicator of predictability of convective precipitation. *Quart. J. Roy. Meteor. Soc.*, **140**, 480–490, <https://doi.org/10.1002/qj.2143>.
- , F. Baur, K. Bachmann, S. Rasp, L. Schneider, and C. Barthlott, 2019: Relative contribution of soil moisture, boundary-layer and microphysical perturbations on convective predictability in different weather regimes. *Quart. J. Roy. Meteor. Soc.*, **145**, 3102–3115, <https://doi.org/10.1002/qj.3607>.
- Kraj, A. G., and E. L. Bibeau, 2010: Phases of icing on wind turbine blades characterized by ice accumulation. *Renewable Energy*, **35**, 966–972, <https://doi.org/10.1016/j.renene.2009.09.013>.
- Lang, S. T. K., S.-J. Lock, M. Leutbecher, P. Bechtold, and R. M. Forbes, 2021: Revision of the stochastically perturbed parametrisations model uncertainty scheme in the integrated forecasting system. *Quart. J. Roy. Meteor. Soc.*, **147**, 1364–1381, <https://doi.org/10.1002/qj.3978>.
- Leutbecher, M., 2019: Ensemble size: How suboptimal is less than infinity? *Quart. J. Roy. Meteor. Soc.*, **145**, 107–128, <https://doi.org/10.1002/qj.3387>.
- , and Coauthors, 2017: Stochastic representations of model uncertainties at ECMWF: State of the art and future vision. *Quart. J. Roy. Meteor. Soc.*, **143**, 2315–2339, <https://doi.org/10.1002/qj.3094>.
- McCabe, A., R. Swinbank, W. Tennant, and A. Lock, 2016: Representing model uncertainty in the Met Office convection-permitting ensemble prediction system and its impact on fog forecasting. *Quart. J. Roy. Meteor. Soc.*, **142**, 2897–2910, <https://doi.org/10.1002/qj.2876>.
- Nygaard, B., L. Moen, Ø. Welgaard, F. Nyhammer, R. Bredesen, and O. Byrkjedal, 2016: Monitoring and forecasting ice loads on a 420 kV transmission line in extreme climatic conditions. *Proc. 16th Int. Workshop on Atmospheric Icing of Structures*, Uppsala, Sweden, Swedish Energy Agency, Abstract 39, https://windren.se/TWAS_p/TWAS2015/00_00_00_Proceedings_TWAS2015_32MB.pdf.
- Ollinaho, P., P. Bechtold, M. Leutbecher, M. Laine, A. Solonen, H. Haario, and H. Järvinen, 2013: Parameter variations in prediction skill optimization at ECMWF. *Nonlinear Processes*

- Geophys.*, **20**, 1001–1010, <https://doi.org/10.5194/npg-20-1001-2013>.
- , H. Järvinen, P. Bauer, M. Laine, P. Bechtold, J. Susiluoto, and H. Haario, 2014: Optimization of NWP model closure parameters using total energy norm of forecast error as a target. *Geosci. Model Dev.*, **7**, 1889–1900, <https://doi.org/10.5194/gmd-7-1889-2014>.
- , and Coauthors, 2017: Towards process-level representation of model uncertainties: Stochastically perturbed parametrizations in the ECMWF ensemble. *Quart. J. Roy. Meteor. Soc.*, **143**, 408–422, <https://doi.org/10.1002/qj.2931>.
- Palmer, T. N., R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G. Shutts, M. Steinheimer, and A. Weisheimer, 2009: Stochastic parametrization and model uncertainty. ECMWF Tech. Memo. 598, 44 pp., <http://www.ecmwf.int/sites/default/files/elibrary/2009/11577-stochastic-parametrization-and-model-uncertainty.pdf>.
- Roberts, N. M., 2008: Assessing the spatial and temporal variation in the skill of precipitation forecasts from an NWP model. *Meteor. Appl.*, **15**, 163–169, <https://doi.org/10.1002/met.57>.
- , and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, <https://doi.org/10.1175/2007MWR2123.1>.
- Schwartz, C. S., and Coauthors, 2010: Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting*, **25**, 263–280, <https://doi.org/10.1175/2009WAF2222267.1>.
- Sleigh, M., P. Browne, M. Diamantakis, T. Haiden, and D. Richardson, 2019: IFS upgrade greatly improves forecasts. ECMWF Newsletter, No. 160, ECMWF, Reading, United Kingdom, 18–22, <https://www.ecmwf.int/node/19164>.
- Termonia, P., and Coauthors, 2018: The ALADIN system and its canonical model configurations AROME cy41t1 and ALARO cy40t1. *Geosci. Model Dev.*, **11**, 257–281, <https://doi.org/10.5194/gmd-11-257-2018>.
- Thompson, G., J. Berner, M. Frediani, J. A. Otkin, and S. M. Griffin, 2021: A stochastic parameter perturbation method to represent uncertainty in a microphysics scheme. *Mon. Wea. Rev.*, **149**, 1481–1497, <https://doi.org/10.1175/MWR-D-20-0077.1>.
- Thoss, A., 2014a: Algorithm theoretical basis document for the cloud mask of the NWC/PPS. Tech. Rep., NWC SAF, 64 pp., [https://www.nwcsaf.org/AemetWebContents/ScientificDocumentation/Documentation/Documentation/PPS/v2014/NWC-CDOP2-PPS-SMHI-SCI-ATBD-1_v1_0.pdf](https://www.nwcsaf.org/AemetWebContents/ScientificDocumentation/Documentation/PPS/v2014/NWC-CDOP2-PPS-SMHI-SCI-ATBD-1_v1_0.pdf).
- , 2014b: User manual for the NWC/PPS application: Science part. Tech. Rep., NWC SAF, 53 pp., https://www.nwcsaf.org/AemetWebContents/ScientificDocumentation/Documentation/PPS/v2014/NWC-CDOP2-PPS-SMHI-SCI-UM-1_v1_0.pdf.
- Tsyrlunikov, M., and D. Gayfulin, 2017: A limited-area spatio-temporal stochastic pattern generator for simulation of uncertainties in ensemble applications. *Meteor. Z.*, **26**, 549–566, <https://doi.org/10.1127/metz/2017/0815>.
- Wastl, C., Y. Wang, A. Atencia, and C. Wittmann, 2019a: A hybrid stochastically perturbed parametrization scheme in a convection-permitting ensemble. *Mon. Wea. Rev.*, **147**, 2217–2230, <https://doi.org/10.1175/MWR-D-18-0415.1>.
- , —, —, and —, 2019b: Independent perturbations for physics parametrization tendencies in a convection-permitting ensemble (pSPPT). *Geosci. Model Dev.*, **12**, 261–273, <https://doi.org/10.5194/gmd-12-261-2019>.