Physically based stochastic perturbations improve a high-resolution forecast of convection

Matjaž Puh[1] | **Christian Keil[1]** | **Christoph Gebhardt[2]** | **Chiara Marsigli[2]** | **Mirjam Hirt[1]** | **Fabian Jakub[1]** | **George C. Craig[1]**

[1]Meteorological Institute,
Ludwig-Maximilians-Universität,
Munich, Germany

[2]Deutscher Wetterdienst, Offenbach,
Germany

Correspondence
Matjaž Puh, Meteorological Institute,
Ludwig-Maximilians-Universität, 80333
Munich, Germany.
Email: M.Puh@physik.uni-muenchen.de

Funding information
Deutsche Forschungsgemeinschaft,
Grant/Award Number: SFB/TR165

Abstract

A physically based stochastic perturbation (PSP) scheme has been implemented in the convection-permitting ICON-D2 ensemble prediction system at Deutscher Wetterdienst (DWD) and run for a three-month trial experiment in summer 2021. The scheme mimics the impact of boundary-layer turbulence on the smallest resolved scales and impacts convective precipitation in particular. A weather-regime-dependent systematic evaluation shows that PSP efficiently increases ensemble spread of precipitation in weak synoptic forcing, while producing realistic convective structures. During strong forcing, the effect of the scheme is negligible, as expected by design. A probabilistic verification shows improvements in the forecast skill of other variables as well, especially the spread-to-skill ratio, but identifies starting points for further improvements of the method.

KEYWORDS

convection, ensemble, model uncertainty, stochastic perturbations

1 | INTRODUCTION

Ensemble prediction systems (EPSs) have been implemented to account for the chaotic nature of the atmosphere, imperfect numerical weather prediction (NWP) models and uncertain initial conditions. They provide multiple scenarios as an alternative to a single forecast started from the best estimate of the initial state. This is particularly beneficial on the kilometre scale that is now resolved by high-resolution NWP models, where errors grow rapidly due to nonlinear processes, limiting the predictability significantly (Hohenegger and Schär, 2007).

One of the main advantages of using an ensemble instead of a single deterministic forecast is its ability to represent different sources of uncertainty. A major source

of error in limited-area models is the uncertainty in the initial and boundary conditions (Lorenz, 1965), which is commonly introduced by constructing the ensemble using perturbed input fields of global ensembles. Another important source of uncertainty is the model itself, as a consequence of our limited knowledge about atmospheric phenomena and the finite grid size. The former can be addressed by perturbing specific components of the model formulation (e.g., parameters), while the latter requires a careful parameterization of subgrid, unresolved processes. Since convection is mostly resolved at the kilometre scale, the quality of its forecast is limited by our understanding of the key physical processes in convection, like boundary-layer turbulence, cloud microphysics (e.g., Thompson *et al.*, 2021; Matsunobu *et al.*, 2022),

and cold-pool dynamics (e.g., Hirt and Craig, 2021). There is still much fundamental research required in this field, including detailed observations of these processes (Clark *et al.*, 2016). On the other hand, convective initiation is often driven by unresolved boundary-layer processes, especially when synoptic forcing is weak and local mechanisms are the main factor in overcoming convection inhibition. In such circumstances, high-resolution models have shown insufficient convective initiation (see e.g., Clark *et al.*, 2016).

In the context of a convection-permitting EPS, the insufficient representation of physical processes in the model is likely a cause for the underdispersion of the ensemble, especially for near-surface variables. This may be mitigated by stochastic schemes: Bouttier *et al.* (2012) find that using the Stochastically Perturbed Parameterization Tendencies (SPPT) scheme in the underdispersive application of research to operations at mesoscale (AROME) ensemble is an effective technique for enhancing spread. Keil *et al.* (2019) study the relative contributions of soil moisture heterogeneities, a stochastic boundary-layer perturbation scheme, and varied aerosol concentrations representing microphysical uncertainties on the diurnal cycle of convective precipitation and its spatial variability. They observe that, in the Consortium for Small-scale Modeling (COSMO) model, the stochastic boundary-layer perturbations lead to the largest spatial variability, impacting precipitation from the initial time onwards with an amplitude comparable to the operational ensemble spread. Similarly, the results of Jankov *et al.* (2017) indicate that a Weather Research and Forecasting (WRF) ensemble combining three stochastic methods consistently produces the best spread–skill ratio and generally outperforms the multiphysics ensemble (see also Jankov *et al.*, 2019), suggesting that using a single-physics ensemble together with stochastic methods should be considered in the design of future high-resolution regional and global ensembles.

In recent years, efforts have been made to develop and test stochastic boundary-layer turbulence schemes that reintroduce the missing small-scale variability. Kober and Craig (2016) developed a physically based stochastic perturbation (PSP) scheme that uses turbulent kinetic energy and flux information from the model's turbulence parameterization to compute the corresponding variances in temperature, moisture, and vertical velocity. Spatially and temporally correlated stochastic increments are then added to the model fields to introduce the resolved portion of this turbulent variability. Using the scheme in the COSMO model, they find that stochastic perturbations lead to triggering of additional convective cells and improve precipitation amounts in simulations of two days with weak synoptic forcing of convection. In a case with strong forcing, the boundary-layer perturbations have little impact, as expected, since the amount of precipitation is controlled by the mesoscale and synoptic environment. The PSP scheme has been revised and improved by Hirt *et al.* (2019), whose version is used in this work (for details see Section 2.2).

Clark *et al.* (2021) implemented a similar, physically consistent stochastic boundary-layer scheme in the Met Office's Unified Model, which introduces temporally correlated multiplicative Poisson noise with a scale-dependent distribution. They evaluate the scheme using small ensemble forecasts of two case studies of severe convective storms over the UK. They find that, with horizontal grid lengths around 1 km, temporal correlation is far more important than spatial. They also show that the scheme produces sufficient differences between ensemble members at the scale of convective cells. Fleury *et al.* (2022) test two process-oriented perturbation schemes in a single-column version of the convection-permitting AROME model. They study three idealized boundary-layer cases using a planetary boundary layer (PBL) turbulence scheme and a shallow convection scheme. They find that these schemes do not produce enough spread to represent the small-scale variability in temperature and humidity seen in large eddy simulations for the same cases. For wind, the variability compares favourably, due to perturbations generated by the stochastic turbulence scheme.

In this study, we use the physically based stochastic perturbation scheme PSP to represent small-scale model error in the boundary layer in the operational ICON-D2-EPS for a three-month period in summer 2021 over Germany. The large number of forecasts in the parallel trial allows for a systematic analysis of the impact of the scheme and thereby lets us infer the properties of forecast uncertainty in different weather regimes. A better understanding of these properties and their dependence on the weather regime will allow us to set up a more optimal prediction system to represent the future state of the atmosphere and the uncertainty associated with it.

The article is structured as follows. Section 2 introduces the experimental setup, the simulation period, and the perturbation scheme used. In Section 3, the effect of the scheme on the diurnal cycle of precipitation in different forcing regimes is presented, including its beneficial impact on the spread in weak forcing conditions. Then the effect on spatial uncertainty of precipitation is shown, as measured by the Fractions Skill Score (FSS), as well as the probabilistic verification of other near-surface variables, which indicates a general improvement in the spread-to-skill relationship. Section 4 summarizes the conclusions of this work, discusses its limitations, and offers a basis for future investigations.

2 | MODEL AND METHODS

2.1 | Model and simulation setup

The ICON model was used for the experiment, more specifically the ICON-D2-EPS of Deutscher Wetterdienst (DWD, Reinert *et al.*, 2021), operational since February 2021, having a horizontal grid spacing of approximately 2.2 km, 65 vertical levels, 20 ensemble members with initial conditions from the Kilometre-scale Ensemble Data Assimilation (KENDA) system (Schraff *et al.*, 2016), and lateral boundary conditions from the operational ICON-EU ensemble. The model runs for the 24-hr forecasts were initialized at 0000 UTC, using the 2100 UTC runs of ICON-EU-EPS from the day before for the lateral boundary conditions. The output was saved hourly for the main prognostic variables. The trial period spans three months of summer 2021, from May 26 to August 31, 2021, which means 98 days of 24-hr forecasts in total.

The trial simulation consists of two separate experiments with slightly different setups: the reference run and the stochastic run with the PSP scheme turned on. The only representation of model uncertainty in the reference run, which mirrors the operational setup, is the parameter perturbations, which are constant in forecast lead time and space but vary among the ensemble members and between forecast runs. In the "psp" run, the PSP scheme is applied with a different random seed to each ensemble member, on top of parameter perturbations. This scheme is described now.

2.2 | Physically based stochastic perturbation scheme

The PSP scheme addresses the grey zone effects of subgrid-scale turbulence for kilometre-scale models (Kober and Craig, 2016). Traditional turbulence parameterizations are typically designed for models with larger grid spacing, where the resolved scales are well separated from the turbulent eddy scale. For the same resolved state, the impact of many subgrid-scale eddies on the resolved state is assumed to be identical, independently of their small-scale, unresolved variability. For kilometre-scale grid spacing, however, this assumption is no longer valid, as only a few eddies can be included within a single grid box. Then, given identical resolved states, the impact of the eddies can vary and follows a distribution. This variability is of specific relevance for convective initiation, when convective inhibition can be overcome only by the stronger eddies. The PSP scheme reintroduces the variability by means of stochastic perturbations that are scaled according to the turbulence variability. The following stochastic

perturbations are added to the temperature, humidity, and vertical velocity tendencies with $\Phi \in \{T, q_v, w\}$:

$$\partial_t \Phi|_{\text{PSP}} = f_z \alpha_{\text{tuning}} \; \eta \; \frac{1}{\tau_{\text{eddy}}} \; \frac{l_{\text{eddy}}}{\Delta x_{\text{eff}}} \sqrt{\overline{\Phi'^2}}, \tag{1}$$

where

$$f_z(z) = \begin{cases} \frac{z}{z_0}, & 0 \le z < z_0, \\ 1, & z_0 \le z \le H_{\text{PBL}}, \\ 1 - \frac{z - H_{\text{PBL}}}{z_0}, & H_{\text{PBL}} < z < H_{\text{PBL}} + z_0, \\ 0, & z \ge H_{\text{PBL}} + z_0, \end{cases} \quad z_0 = 500 \text{ m}. \tag{2}$$

The perturbations are based on a horizontal random field $\eta(x, y, t | \tau_{\text{eddy}}, \Delta x_{\text{eff}})$. It evolves over time by an autoregressive process with a time correlation corresponding to the characteristic lifetime of turbulent eddies $\tau_{\text{eddy}} = 10$ min, thereby allowing for memory effects due to missing scale separation (Berner *et al.*, 2017). The random field η also has a spatial correlation of Δx_{eff} (via an approximate Gaussian convolution), which corresponds to the smallest effectively resolved scale, $\Delta x_{\text{eff}} = 5\Delta x$ (see e.g., Bierdel *et al.*, 2012). Importantly, the random field is scaled according to the subgrid standard deviation $\sqrt{\overline{\Phi'^2}}$ computed by the deterministic turbulence parameterization (Raschendorfer, 2001). Furthermore, the scheme becomes scale-adaptive by multiplying the perturbations with $1/\sqrt{N_{\text{eddy}}}$, since the number of eddies of scale $l_{\text{eddy}} = 1000$ m in a grid box, $N_{\text{eddy}} = \Delta x^2 / l_{\text{eddy}}^2$, characterizes the variance of the subgrid scale impact (Craig and Cohen, 2006). In the vertical, we taper the perturbations linearly to zero above the top of the boundary layer (H_{PBL}), as well as in the lowest part of the boundary layer, near the surface, as described by factor f_z. Finally, the parameter α_{tuning} should be of order one and independent of weather regimes or model resolution. Here, α_{tuning} is set to 5. The implementation of PSP into ICON closely follows the version from Hirt *et al.* (2019) implemented in the COSMO model, including the autoregressive process and the tapering of the perturbations at the top of the boundary layer, but excluding the perturbations of the horizontal wind.

Figure 1 shows the observed 24-hr accumulated precipitation on one day and the corresponding fields for one ensemble member, drawn from the reference and the PSP experiments. The scheme triggers convection more efficiently, which is visible over central–south Germany, where more convective cells appear throughout the day, compared with the reference, where there is a "hole" in the precipitation pattern over that region. This feature does
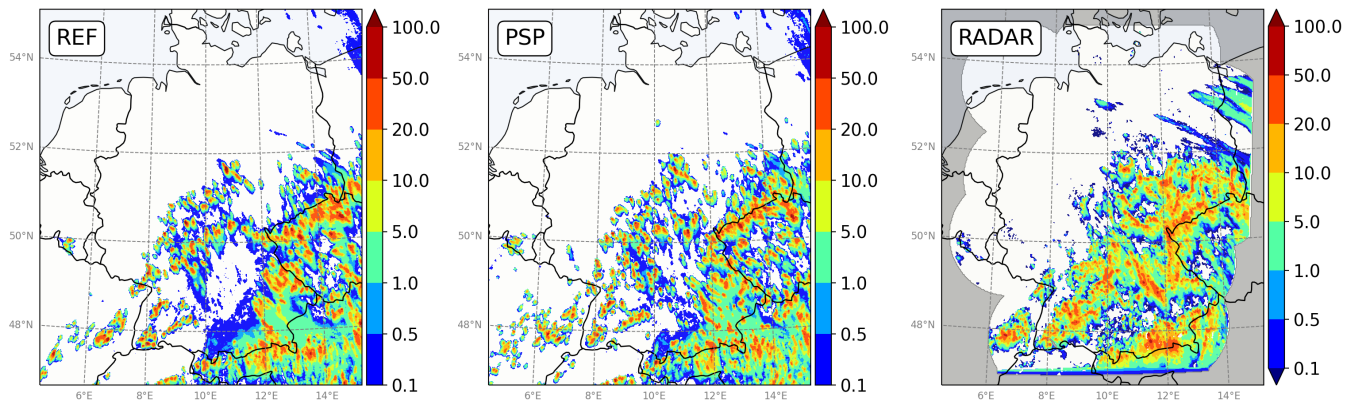
FIGURE 1   Total daily precipitation for member 19 of the ensemble for June 10, 2021: reference experiment (left), PSP experiment (centre), and estimate of accumulated precipitation from radar observations (right). [Colour figure can be viewed at wileyonlinelibrary.com]
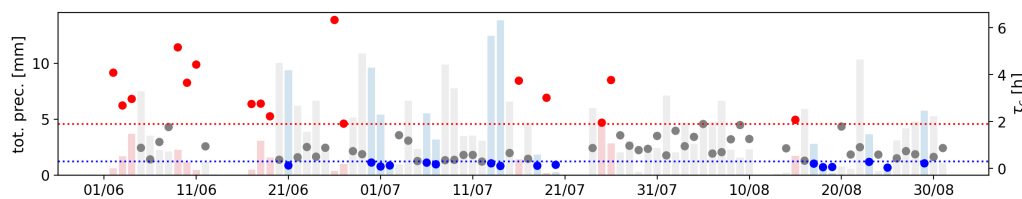

FIGURE 2   Time series of daily total precipitation (bars) and daily averaged convective adjustment timescale (dots), both for the reference experiment, averaged over Germany, for summer 2021. Red indicates weakly forced days, blue strongly forced days. The horizontal dotted lines show the threshold value of the convective adjustment timescale for weak forcing (red) and strong forcing (blue). See section 3.1 for details on the classification. [Colour figure can be viewed at wileyonlinelibrary.com]

not appear in the radar-derived precipitation map. At the same time, however, the PSP scheme does not trigger spurious convective cells in the rest of the domain, where there is virtually no precipitation on that day. The same conclusions can be drawn by looking at other ensemble members (not shown).

3 | RESULTS AND DISCUSSION

3.1 | Synoptic forcing regime classification

From the perspective of forecast uncertainty at the convective scale, the type of convective forcing is important. Hence, studying the strong and weak forcing regimes separately allows us to infer the properties of forecast error and uncertainty evolution conditional to the weather regime. To make the distinction between strong and weak synoptic forcing, we applied the convective adjustment timescale τ_c. It is an estimate of the timescale for the removal of conditional instability (measured by convective available potential energy (CAPE)) by convective heating (Keil *et al.*, 2014). A small value of τ_c compared with the timescale of the synoptic flow (about 12 hr) means that CAPE is removed by convection as soon as it is created

and the large-scale flow controls the amount of convection. If the value of τ_c is large, the removal of the CAPE by convection is too slow, so small scale factors drive the convection. The computation of τ_c is only done on days when the threshold of $1 \, \text{mm·hr}^{-1}$ was exceeded at least once in more than 100 grid points over Germany (as in Kühnlein *et al.*, 2014).

In summer 2021, the weather was characterised by abundant precipitation, with the largest accumulations in the last 10 years on average over Germany (DWD, 2021). Several high-impact weather events occurred, including floods in western Germany (July 13–14) and hailstorms in southern Germany in the last third of June, including a squall line with widespread severe winds on June 29. Daily average values of the convective adjustment timescale, averaged over Germany, vary from less than an hour to more than 5 hr (dots in Figure 2). Most of the strongly forced days have large amounts of domain-averaged accumulated precipitation. In contrast, weakly forced days typically feature smaller domain-averaged precipitation sums, while its spatial distribution is highly variable (see Figure 1). To partition out the days with strongest and weakest synoptic control we take the lowest 20% of average τ_c values and classify these as strong forcing, while the highest 20% of daily values are considered as weak forcing. An advantage of this approach over setting certain fixed
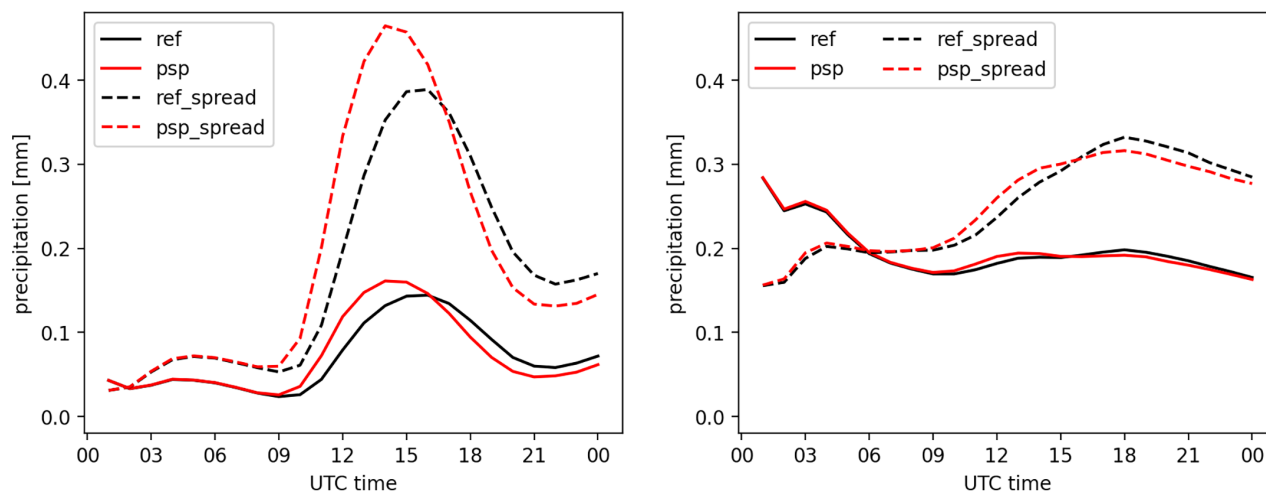
FIGURE 3   Composite time series of hourly precipitation amount (continuous lines) and spread (dashed lines) for weak forcing (left) and strong forcing days (right), for the reference experiment (black) and the PSP experiment (red), averaged over Germany. [Colour figure can be viewed at wileyonlinelibrary.com]

thresholds is the creation of equally populated samples containing 16 days for each regime.

3.2 | Diurnal cycle of precipitation

One of the key challenges in convective-scale weather prediction is an accurate forecast of precipitation amount, timing, and uncertainty. This holds especially true in the absence of larger-scale forcing, when local processes in the boundary layer drive the convection. Figure 3 shows composite time series of domain-averaged precipitation amounts and its variability for both regimes based on 16 strongly and 16 weakly forced days. The PSP scheme has an overall higher impact during weak forcing, in terms of both average precipitation amounts and ensemble spread of precipitation. During weak forcing days, the diurnal cycle is clearly evident and peaks in the afternoon (around 1500 UTC). The PSP scheme shifts the maximum about an hour earlier due to more efficient triggering of convection, caused by buoyant air bubbles in the boundary layer, created by the PSP scheme. The onset of perturbations is directly connected to the subgrid standard deviation of selected variables, which increases as the solar radiation heats the surface. Hence, convection is formed earlier than in the absence of the PSP scheme, which is one of the goals of the scheme. For a more detailed discussion about the role of PSP in triggering mechanisms, the reader is referred to Hirt *et al.* (2019).

The earlier shift of the diurnal cycle of precipitation is beneficial, since precipitation in convective-scale models usually lags the observed precipitation maximum in these flow conditions (e.g., Keil *et al.*, 2019). Moreover, the magnitude of the peak is higher, especially in spread. Thus,

physically based perturbations in the boundary layer lead to a reduction of the underdispersion of precipitation (not shown), which is a general issue of convection-permitting ensemble prediction systems. This includes the ICON-D2 EPS as well, according to spread–skill ratios, which are mostly below 0.5 in the DWD operational verification based on rain-gauge data.

On strong forcing days, there is no clear diurnal cycle in precipitation amount and the PSP perturbations have little effect. While the average amount of precipitation is higher in strong forcing, the magnitude of the spread is higher in weak forcing. These results are consistent with those of earlier case studies using the PSP scheme in the COSMO model (Kober and Craig, 2016; Hirt *et al.*, 2019; Keil *et al.*, 2019).

3.3 | Spatial uncertainty of precipitation

To assess the predictive skill of the precipitation forecasts, we apply a spatial verification method to account for the spatiotemporal highly variable nature of precipitation. The widely used FSS compares the fractional coverage of the events directly in windows surrounding the observations and forecasts (Roberts and Lean, 2008). Observations are provided by the quality-controlled precipitation field estimated by the German radar network on a 1-km grid every 5 min. The observed data are upscaled to the ICON-D2 grid and accumulated to hourly values for comparison with the model output.

In Figure 4, the FSS is shown for each ensemble member (thin lines) of both ensembles (PSP experiment in red and reference in black) as a function of the lead time on days classified as weakly forced according to the convective
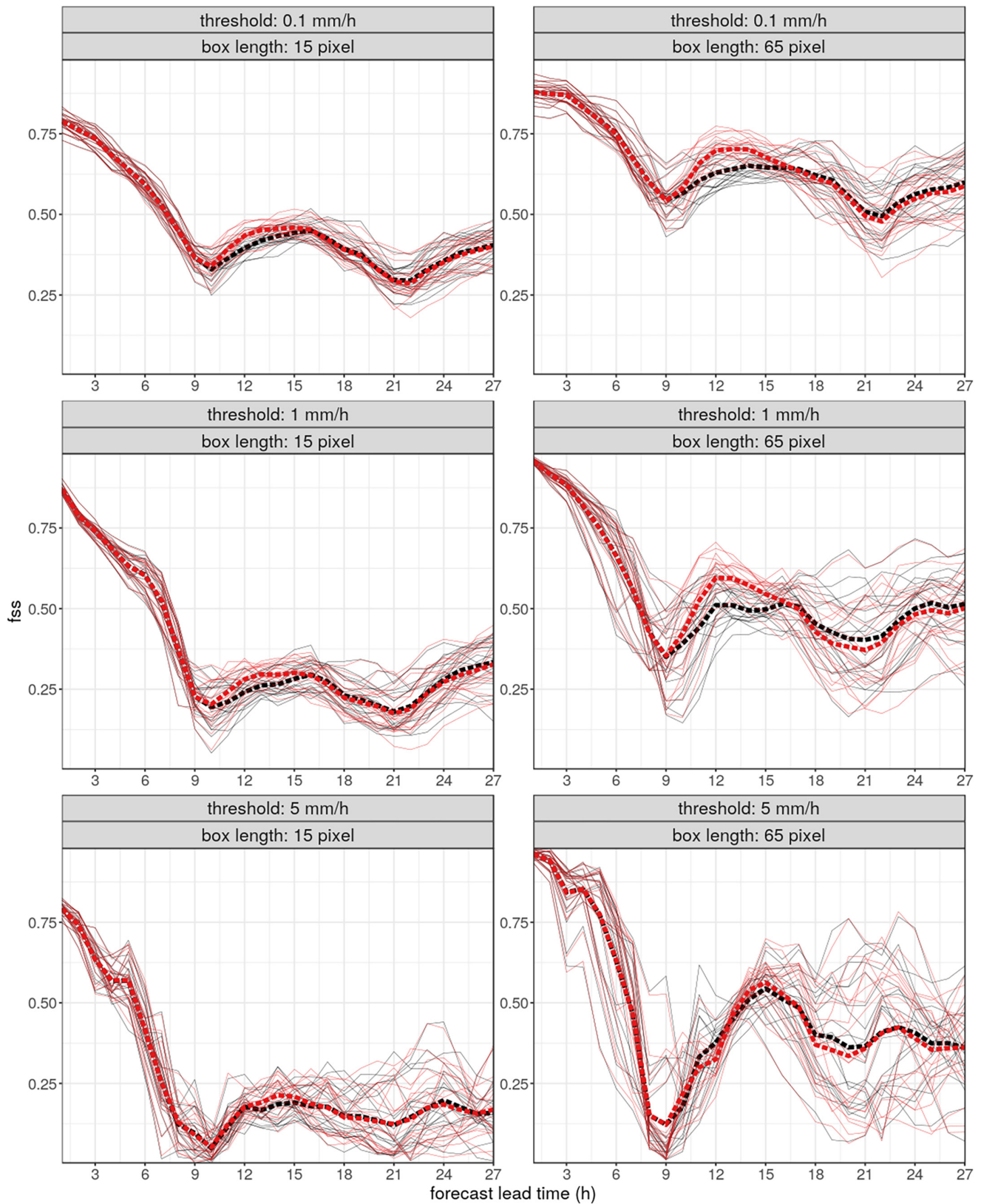
FIGURE 4 Spatial forecast skill as measured by the FSS for each ensemble member (thin lines) of the two ensembles (PSP in red and reference in black) as a function of the lead time averaged over the 16 days with weakly forced convection. The mean over the members is plotted as a thick dashed line. The scores are shown for exceedances of 0.1 (top), 1 (middle), and 5 mm·hr⁻¹ (bottom) and for two aggregation window sizes with dimensions of 15 pixel (about 30 km, left) and 65 pixel (about 140 km, right). [Colour figure can be viewed at wileyonlinelibrary.com]

adjustment timescale (see Figure 2). The scores are shown for the exceedance of three selected hourly precipitation thresholds (0.1, 1, and 5 mm·hr^{-1}: upper, middle, and lower row, respectively) and for two aggregation window sizes, with dimensions of 15 pixel (left column), corresponding to about 30 km, and 65 pixel (right column), corresponding to about 140 km. For ease of reading, the average value of the FSS of the members is also plotted, as a thick dashed line, for both the PSP experiment and the reference one.

Generally, the FSS is higher at short forecast lead times and decreases over time. However, after convective initiation and the generation of precipitation from 0900 UTC onwards, the FSS increases and attains higher FSS values in the central part of the day, between 0900 and 1800 UTC, when the maximum of convection occurs. A relative minimum is observed around 2100 UTC. At the smaller aggregation scale (left column), the lines related to individual members of both experiments tend to stay close together, showing similar performance of the two experiments. Between 1200 and 1500 UTC, the period of most active convection, the mean score shows a slightly better performance for the PSP experiment. Both ensembles become more disperse at longer forecast lead times with a higher precipitation threshold, as shown by the larger difference in FSS between the members.

When the larger aggregation window is considered (right column), the difference in performance of the two ensembles becomes more marked. The individual ensembles are more disperse, and the PSP experiment outperforms the reference one, as shown by the larger mean FSS value, in particular for the 0.1 and 1 mm·hr^{-1} thresholds. The difference is clearly evident during strong convection between 1200 and 1500 UTC. Interestingly, the FSS of the 5 mm·hr^{-1} threshold is slightly higher than that of the 1 mm·hr^{-1} threshold at peak precipitation at 1500 UTC. This is presumably caused by the sample size and averaging effects. After 1800 UTC, the FSS of the PSP experiment is slightly lower than that of the reference run in the last part of the day, in agreement with the behaviour shown when representing the diurnal cycle of the spread (see Figure 3). During strong forcing, the time series of the FSS do not show a significant impact of the PSP scheme and predominantly show a steady decrease with lead time (not shown).

3.4 | Probabilistic verification of near-surface variables

An objective verification of the performance of both experiments has been carried out for a wide range of meteorological variables using a standard set of indices for the whole trial period. The results are shown in Figure 5 for a selection of variables: wind speed at 10 m above the ground (first column), cloud cover for low clouds (second column), temperature, and dew-point temperature at 2 m above the ground (third and fourth column, respectively). The scores are computed against observations at the SYNOP stations over Germany. Due to the representativity error of these observations, we exclude the verification of precipitation in this section. In the first row, the continuous ranked probability score (CRPS) is shown as a measure of quality of the ensemble forecasts (negatively oriented), in the second row the root-mean-square error of the ensemble mean (RMSE), in the third row the mean error of the ensemble mean (ME), and in the fourth row the ensemble standard deviation (SPREAD). Finally, in the fifth row the spread-to-skill relationship is shown, expressed as the ratio between the spread and the standard deviation (SD) of the error of the ensemble mean. This measure has been chosen because the ensemble spread should match the random component of the RMSE of the ensemble mean, having subtracted the bias, so that values less than one indicate underdispersion. For a description of the indices, the reader is referred to Wilks (2019). In each plot, the red colour is for the PSP experiment, while the black colour is for the reference experiment. The dots are filled with colour when the difference between the scores of the two experiments is significant, as computed following a bootstrap method. No significance estimation has been performed for the spread-to-skill relationship.

Probabilistic verification scores over the whole period are improved for a wide range of variables when using the PSP scheme, especially the spread and spread-to-skill relation. The CRPS is also slightly improved for the PSP experiment, compared with the reference, with the exception of the 2-m dewpoint temperature between 1800 and 2400 UTC. The 2-m dewpoint temperature scores are also slightly deteriorated in terms of RMSE in this part of the day, while for other variables the RMSE is either smaller (low cloud cover) or not significantly different for the PSP experiment compared with the reference experiment. The combination of larger spread and equal or smaller RMSE leads to an increased, beneficial spread/skill relation. This is a particularly positive result, given the general issue of models being underdispersive for near-surface variables.

However, a more detailed look at the verification scores shows some issues, specifically for 2-m dewpoint temperature and low cloud cover. The mean error of the 2-m dewpoint temperature points to a marked drying effect in the period considered. Moreover, the (at first sight) beneficial behaviour of the PSP scheme causing a reduction of the mean error of low clouds turns out to be mainly caused by increased patchiness of the low cloud field on nonrainy days, resulting in a reduced mean error. Visual
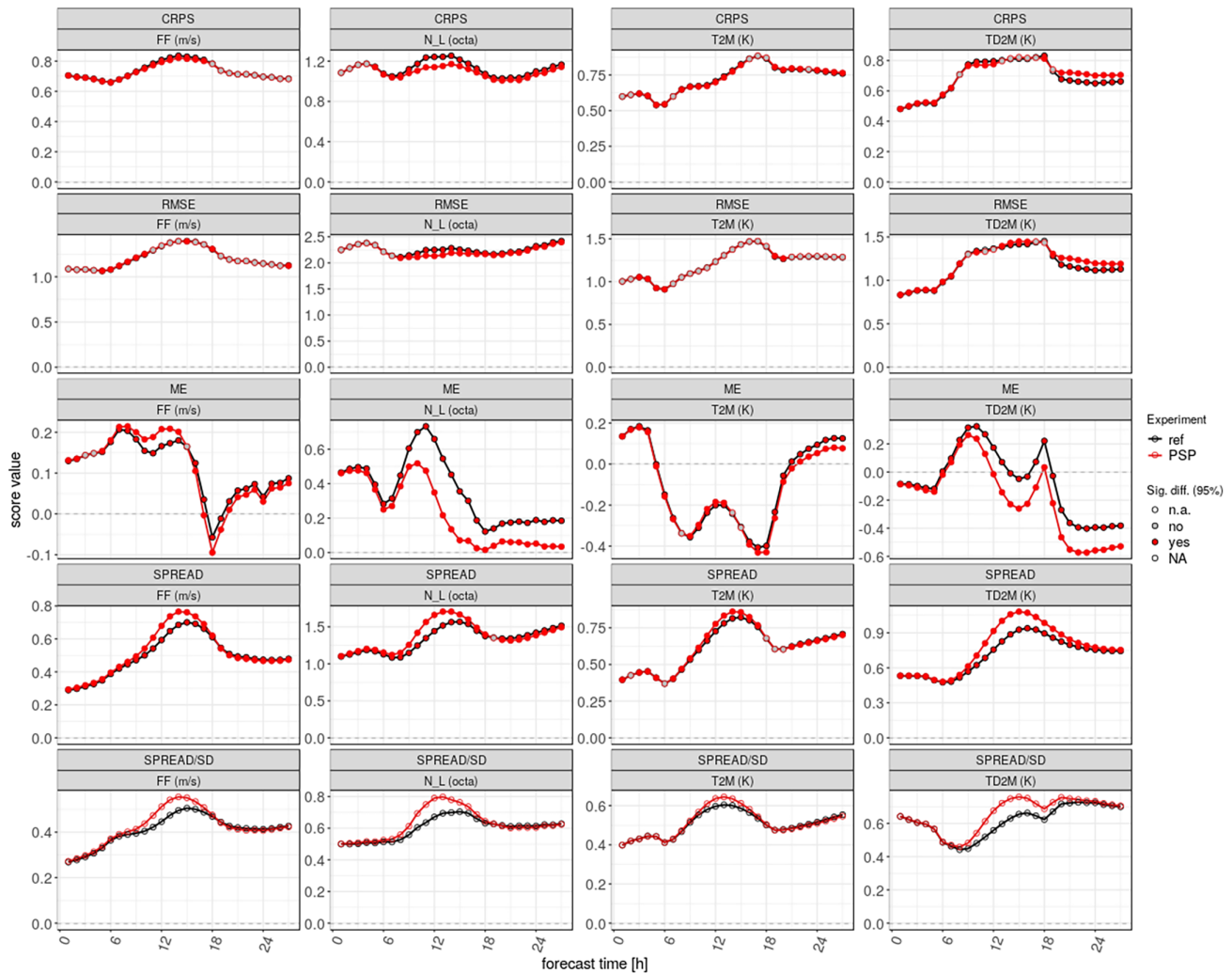
FIGURE 5 Diurnal cycle of domain-averaged CRPS, RMSE, mean error, ensemble spread, and spread-to-skill relationship (see text for details) for 10-m wind (FF), low cloud cover (NL), 2-m temperature (T2M), and 2-m dewpoint temperature (TD2M) of the reference (black) and PSP experiments (red), verified against SYNOP observations, for the period between May 26 and August 31, 2021. [Colour figure can be viewed at wileyonlinelibrary.com]

inspection indicates that the increased patchiness is unrealistic (not shown). Preliminary inspection of a few case studies with low cloud cover, but no rain, indicates that the current vertical profile of the perturbations in PSP entrains too much dry air from aloft into the boundary layer, causing a dry bias in the boundary layer (see the mean error of the 2-m dewpoint temperature, too). A modification of the vertical profile at the upper boundary of the PBL determining the perturbation strength of PSP leads to improved results for a nonrainy and rainy case studies and will be pursued in future PSP applications (a systematic investigation is beyond the scope of the present trial). For medium and high clouds, the impact of PSP is almost neutral (not shown).

The PSP scheme also shows an increase of the mean error of 10-m wind speed, which was also detected in the verification of wind gusts (not shown). This is likely caused

by a double counting of turbulence effects that are taken into account when diagnosing near-surface wind speed. We should point out that the verification in Figure 5 was performed against SYNOP observations, an observation type traditionally used for near-surface variables that has limitations when estimating forecast errors of cloud cover and wind gusts. Therefore, a different kind of observation should be used in future to improve the diagnostics, including, for example, satellite observations to compare visible reflectances directly (using a satellite forward operator, e.g., Scheck *et al.*, 2020).

4 | CONCLUSIONS

Increasing the resolution of NWP models makes traditional parameterizations of boundary-layer turbulence

inadequate, because the assumption that the grid-box size is much larger than the size of eddies does not hold anymore in kilometre-scale models. In this study, we use the recently implemented physically based stochastic perturbation scheme PSP in ICON-D2-EPS as a representation of model error originating from the subgrid scale in the boundary layer, but affecting the smallest resolved scales. The experimental period spans a whole summer season, which allows for a systematic analysis of the impact of the scheme in different synoptic forcing conditions. The main conclusions of this work are the following.

1. The PSP scheme provides a good representation of the effect of subgrid-scale turbulence in ICON-D2 and has realistic, beneficial effects in ensemble forecasts, especially on the ensemble spread. It helps trigger convection, while preserving the intensity of single convective cells, and does not produce spurious convection. PSP reduces the underdispersion of precipitation.

2. Small-scale perturbations, introduced by PSP, have a larger impact on convective precipitation in weak than in strong synoptic forcing, especially on its spread. This is in line with the hypothesis of local processes in the boundary layer driving the convection on weakly forced days, whereas on strong forcing days the synoptic pattern controls the convection.

3. The PSP scheme improves the spatial distribution of precipitation (FSS) slightly around the peak of its diurnal cycle in weak synoptic forcing, compared with radar observations. Its impact is neutral during strong forcing.

4. The probabilistic verification of near-surface variables predominantly shows a neutral to slightly beneficial forecast performance. The systematic assessment indicates a few issues that deserve further research (namely the 2-m dewpoint temperature and wind gusts at the surface) on the way towards operational implementation of PSP in ICON-D2-EPS.

General issues with physically based schemes are interactions between different schemes and the double counting of physical processes. Further work will therefore examine the effects of combining PSP with the stochastic shallow convection scheme developed by Sakradzija and Klocke (2018) in ICON. The two schemes should act independently by design, although both would likely affect the layers around the top of the PBL, where we found a detrimental impact of PSP in its current implementation.

The results of this work are encouraging: PSP improves the spread-to-skill ratio of the ensemble for several variables, especially those near the surface, for which the forecast is often underdispersive. This is a promising step on the way to operational use of the scheme and in general for the development of physically based stochastic schemes. The limitations of certain observation types and the deterioration of the forecasts for a few variables provide a basis for further research.

AUTHOR CONTRIBUTIONS

Matjaž Puh: conceptualization; data curation; formal analysis; investigation; methodology; software; validation; visualization; writing – original draft; writing – review and editing. **Christian Keil:** conceptualization; formal analysis; funding acquisition; investigation; methodology; project administration; resources; supervision; validation; writing – original draft; writing – review and editing. **Christoph Gebhardt:** conceptualization; data curation; formal analysis; investigation; methodology; resources; software; validation; visualization; writing – original draft; writing – review and editing. **Chiara Marsigli:** conceptualization; data curation; formal analysis; investigation; methodology; resources; software; validation; visualization; writing – original draft; writing – review and editing. **Mirjam Hirt:** methodology; writing – original draft; writing – review and editing. **Fabian Jakub:** methodology; software; writing – review and editing. **George C. Craig:** conceptualization; formal analysis; funding acquisition; methodology; project administration; resources; supervision; validation; writing – review and editing.

ACKNOWLEDGEMENTS

DATA AVAILABILITY STATEMENT

The computations in this study have been performed using the ICON model available at https://gitlab.dkrz.de/icon/icon-nwp/-/tree/icon-nwp/icon-nwp-psp with revision 546be640a57c53a4565a899f233fac79f01738a1. Given that that access to the ICON code repository is restricted, the interested reader may request a copy of the PSP scheme source code from the authors.

ORCID

Matjaž Puh https://orcid.org/0000-0002-0954-8940
Christian Keil https://orcid.org/0000-0003-2736-4309
Chiara Marsigli https://orcid.org/0000-0003-3376-1896
Mirjam Hirt https://orcid.org/0000-0002-5650-7594
George C. Craig https://orcid.org/0000-0002-7431-8164

REFERENCES

Berner, J., Achatz, U., Batté, L., Bengtsson, L., de la Cámara, A., Christensen, H.M., Colangeli, M., Coleman, D.R.B., Crommelin, D., Dolaptchiev, S.I., Franzke, C.L.E., Friederichs, P., Imkeller, P., Järvinen, H., Juricke, S., Kitsios, V., Lott, F., Lucarini, V., Mahajan, S., Palmer, T.N., Penland, C., Sakradzija, M., von Storch, J.-S., Weisheimer, A., Weniger, M., Williams, P.D. and Yano, J.-I. (2017) Stochastic parameterization: toward a new view of weather and climate models. *Bulletin of the American Meteorological Society*, 98, 565–588 https://journals.ametsoc.org/view/journals/bams/98/3/bams-d-15-00268.1.xml.

Bierdel, L., Friederichs, P. and Bentzien, S. (2012) Spatial kinetic energy spectra in the convection-permitting limited-area NWP model COSMO-DE. *Meteorologische Zeitschrift*, 21, 245–258.

Bouttier, F., Vié, B., Nuissier, O. and Raynaud, L. (2012) Impact of stochastic physics in a convection-permitting ensemble. *Monthly Weather Review*, 140, 3706–3721.

Clark, P., Roberts, N., Lean, H., Ballard, S.P. and Charlton-Perez, C. (2016) Convection-permitting models: a step-change in rainfall forecasting. *Meteorological Applications*, 23, 165–181.

Clark, P.A., Halliwell, C.E. and Flack, D.L.A. (2021) A physically based stochastic boundary layer perturbation scheme. Part i: formulation and evaluation in a convection-permitting model. *Journal of the Atmospheric Sciences*, 78, 727–746 https://journals.ametsoc.org/view/journals/atsc/78/3/JAS-D-19-0291.1.xml.

Craig, G.C. and Cohen, B.G. (2006) Fluctuations in an equilibrium convective ensemble. Part I: theoretical formulation. *Journal of the Atmospheric Sciences*, 63, 1996–2004.

DWD. (2021) Deutschlandwetter im Sommer 2021. https://www.dwd.de/DE/presse/pressemitteilungen/DE/2021/20210830_deutschlandwetter_sommer2021_news.html.

Fleury, A., Bouttier, F. and Couvreux, F. (2022) Process-oriented stochastic perturbations applied to the parametrization of turbulence and shallow convection for ensemble prediction. *Quarterly Journal of the Royal Meteorological Society*, 148, 981–1000 https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.4242.

Hirt, M. and Craig, G.C. (2021) A cold pool perturbation scheme to improve convective initiation in convection-permitting models. *Quarterly Journal of the Royal Meteorological Society*, 147, 2429–2447 https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.4032.

Hirt, M., Rasp, S., Blahak, U. and Craig, G.C. (2019) Stochastic parameterization of processes leading to convective initiation in kilometer-scale models. *Monthly Weather Review*, 147, 3917–3934 https://journals.ametsoc.org/view/journals/mwre/147/11/mwr-d-19-0060.1.xml.

Hohenegger, C. and Schär, C. (2007) Predictability and error growth dynamics in cloud-resolving models. *Journal of the Atmospheric Sciences*, 64, 4467–4478 https://journals.ametsoc.org/view/journals/atsc/64/12/2007jas2143.1.xml.

Jankov, I., Beck, J., Wolff, J., Harrold, M., Olson, J.B., Smirnova, T., Alexander, C. and Berner, J. (2019) Stochastically perturbed parameterizations in an hrrr-based ensemble. *Monthly Weather Review*, 147, 153–173 https://journals.ametsoc.org/view/journals/mwre/147/1/mwr-d-18-0092.1.xml.

Jankov, I., Berner, J., Beck, J., Jiang, H., Olson, J.B., Grell, G., Smirnova, T.G., Benjamin, S.G. and Brown, J.M. (2017) A performance comparison between Multiphysics and stochastic approaches within a north American RAP ensemble. *Monthly Weather Review*, 145, 1161–1179.

Keil, C., Baur, F., Bachmann, K., Rasp, S., Schneider, L. and Barthlott, C. (2019) Relative contribution of soil moisture, boundary-layer and microphysical perturbations on convective predictability in different weather regimes. *Quarterly Journal of the Royal Meteorological Society*, 145, 3102–3115 https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3607.

Keil, C., Heinlein, F. and Craig, G.C. (2014) The convective adjustment time-scale as indicator of predictability of convective precipitation. *Quarterly Journal of the Royal Meteorological Society*, 140, 480–490 https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.2143.

Kober, K. and Craig, G.C. (2016) Physically based stochastic perturbations (psp) in the boundary layer to represent uncertainty in convective initiation. *Journal of the Atmospheric Sciences*, 73, 2893–2911 https://journals.ametsoc.org/view/journals/atsc/73/7/jas-d-15-0144.1.xml.

Kühnlein, C., Keil, C., Craig, G.C. and Gebhardt, C. (2014) The impact of downscaled initial condition perturbations on convective-scale ensemble forecasts of precipitation. *Quarterly Journal of the Royal Meteorological Society*, 140, 1552–1562.

Lorenz, E.N. (1965) A study of the predictability of a 28-variable atmospheric model. *Tellus*, 17, 321–333 https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2153-3490.1965.tb01424.x.

Matsunobu, T., Zarboo, A., Barthlott, C. and Keil, C. (2022) Impact of combined microphysical uncertainties on convective clouds and precipitation in icon-d2-eps forecasts during different synoptic control. *Weather and Climate Dynamics Discussions*, 2022, 1–25 https://wcd.copernicus.org/preprints/wcd-2022-17/.

Raschendorfer, M. (2001) The new turbulence parameterization of LM. COSMO Newsletter No. 1, 89-97. http://www.cosmo-model.org/.

Reinert, D., Prill, F., Denhard, H.F.M., Baldauf, M., Schraff, C., Marsigli, C.G. and Zängl, G. (2021) *DWD Database Reference for the Global and Regional ICON and ICON-EPS Forecasting System*. Offenbach, Germany: DWD.

Roberts, N.M. and Lean, H.W. (2008) Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, 136, 78–97 https://journals.ametsoc.org/view/journals/mwre/136/1/2007mwr2123.1.xml.

Sakradzija, M. and Klocke, D. (2018) Physically constrained stochastic shallow convection in realistic kilometer-scale simulations. *Journal of Advances in Modeling Earth Systems*, 10, 2755–2776 https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018MS001358.

Scheck, L., Weissmann, M. and Bach, L. (2020) Assimilating visible satellite images for convective-scale numerical weather prediction: a case-study. *Quarterly Journal of the Royal Meteorological Society*, 146, 3165–3186 https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.3840.

Schraff, C., Reich, H., Rhodin, A., Schomburg, A., Stephan, K. and Periáñez, A. (2016) Kilometre-scale ensemble data assimilation

for the COSMO model (KENDA). *Quarterly Journal of the Royal Meteorological Society*, 142, 1453–1472.

Thompson, G., Berner, J., Frediani, M., Otkin, J.A. and Griffin, S.M. (2021) A stochastic parameter perturbation method to represent uncertainty in a microphysics scheme. *Monthly Weather Review*, 149, 1481–1497 https://journals.ametsoc.org/view/journals/mwre/149/5/MWR-D-20-0077.1.xml.

Wilks, D.S. (2019) Chapter 9-forecast verification. In: Wilks, D.S. (Ed.) *Statistical Methods in the Atmospheric Sciences*, fourth edition. Elsevier, pp. 369–483 https://doi.org/10.1016/B978-0-12-815823-4.00009-2.