ORIGINAL ARTICLE

A convective-scale 1000-member ensemble simulation and potential applications

Tobias Necker[1]* | Stefan Geiss[1] | Martin Weissmann[2,3] | Juan Ruiz[4] | Takemasa Miyoshi[5] | Guo-Yuan Lien[5,6]

[1]Hans-Ertel Centre for Weather Research, Ludwig-Maximilians-Universität, Munich, Germany

[2]Hans-Ertel Centre for Weather Research, Deutscher Wetterdienst, Munich, Germany

[3]Institut für Meteorologie und Geophysik, Universität Wien, Vienna, Austria

[4]Centro de Investigaciones del Mar y la Atmosfera, CIME/CONICET-UBA, Buenos Aires, Argentina

[5]RIKEN Center for Computational Science, Kobe, Japan

[6]Central Weather Bureau, Taipei, Taiwan

Correspondence
Tobias Necker, Meteorological Institute, Ludwig-Maximilians-Univerisität, Theresienstr. 37, 80333 Munich, Germany
Email: tobias.m.necker@lmu.de

Present address
*Ludwig-Maximilians-Univerisität, Munich, Germany

Funding information
HErZ2 by BMVI (Federal Ministry of Transport, Building, and Urban Development), Grant/Award Number: DWD2014P8

This study presents the first convective-scale 1000-member ensemble simulation over central Europe, which provides a unique data set for various applications. A comparison to the operational regional 40-member ensemble of Deutscher Wetterdienst shows that the 1000-member simulation overall exhibits realistic spread properties. Based on this, we discuss two potential applications. At first, we quantify the sampling error of spatial covariances of smaller subsets compared to the 1000-member simulation. Knowledge about sampling errors and their dependence on ensemble size is crucial for ensemble and hybrid data assimilation and for developing better approaches for localization in this context. Secondly, we present an approach for estimating the relative potential impact of different observable quantities using ensemble sensitivity analysis. This shall provide the basis for consecutive studies developing future observation and data assimilation strategies. Sensitivity studies on the ensemble size indicate that about 200 ensemble members are required to estimate the potential impact of observable quantities with respect to precipitation forecasts.

KEYWORDS
data assimilation; convective-scale; covariance; sampling error;

*Equally contributing authors.

ensemble sensitivity analysis; observing system; localization

1 | INTRODUCTION

Over the past thirty years, the skill of numerical weather prediction (NWP) has improved tremendously. This progress results from both scientific and technological advances in various fields (Bauer et al., 2015). Advanced data assimilation (DA) methods and especially the incorporation of flow-dependent error covariances from ensembles was one major contributor (Bonavita et al., 2016; Bannister, 2017). Additionally, new computational resources have allowed higher resolution and regional NWP models, which nowadays usually resolve deep convection explicitly based on a grid spacing of a few kilometers. The chaotic nature and limited predictability of convection, however, also poses new challenges in terms of data assimilation. In particular, the higher resolution and low predictability calls for the assimilation of spatially and temporally highly resolved observations (Gustafsson et al., 2018). Consequently, substantial efforts have been made to assimilate high-resolution radar reflectivity and cloud-affected satellite observations (Miyoshi et al., 2016b; Harnisch et al., 2016; Scheck et al., 2018; Sawada et al., 2019). However, successfully assimilating such observations requires both accurate parameterizations as well as accurate estimates of highly flow-dependent error covariances (Houtekamer and Zhang, 2016).

Ensemble-based estimates of error covariances strongly depend on the available ensemble size. Current operational ensemble systems range from about 20 up to 250 members as the affordable ensemble size is restricted by computational cost. Given that the number of ensemble members is therefore much smaller than the number of degrees of freedom of the model implies several challenges: On the one hand, a small ensemble does not sample all possible states. On the other hand, the estimates of error covariances are substantially affected by sampling errors leading to spurious correlations. To reduce sampling errors, localization is usually applied, which damps correlations after a certain distance (Gaspari and Cohn, 1999). However, finding appropriate localization scales is an important challenge as real correlations can extend over thousands of kilometers in the horizontal and throughout the entire troposphere in the vertical (Caron and Buehner, 2018; Lei et al., 2018). Furthermore, satellite observations often provide vertically integrated information that can not be assigned to a single level. Thus, it is crucial to understand error covariances better, to quantify sampling errors depending on the ensemble size and to develop improved techniques for sampling error correction and localization.

A major challenge is the development of observation and data assimilation strategies for high-resolution NWP given a vast amount of potentially available information in developed countries (Gustafsson et al., 2018). First, NWP centers do not have the human resources to incorporate all these often complex sources of information at the same time. Secondly, new observation selection strategies are especially required considering the vast amount of unused observations provided by radars, satellites, ground-based profilers or community observations (e.g., smartphones, webcams, and renewable power production). Last but not least, technological advances have led to novel and much cheaper remote-sensing instruments that could be deployed in the future. Therefore, better knowledge is needed on what observations are most important for convective-scale NWP and where to put priorities and resources.

Several large-ensemble assimilation and forecast studies have been conducted to address these challenges, but mostly using lower-resolution or idealized models. The latest generation of supercomputers allows one to perform high-resolution big ensemble forecasts with a frequent update cycling (Miyoshi et al., 2015, 2016a). First experiments using a 10240-member global ensemble showed that large ensembles can be applied to learn about sampling errors, non-Gaussianity (Miyoshi et al., 2014) or to improve covariance localization (Kondo and Miyoshi, 2016). Furthermore, a study by Jacques and Zawadzki (2015) computed 1000 convective-scale forecasts to investigate background errors for

radar data assimilation.

In this study, we calculated a set of ten ensemble forecasts with 1000 members each using a full-physics non-hydrostatic regional model (SCALE-RM) with a horizontal grid spacing of 3 km. First, we compare the 1000-member ensemble simulation against the operational convective-scale 40-member ensemble system of Deutscher Wetterdienst (COSMO/KENDA; Baldauf et al. (2011); Schraff et al. (2016)). This comparison is done to show that the 1000-member ensemble performs reasonably well and exhibits realistic properties despite differences in the DA and modelling systems. The large ensemble is afterwards used to derive realistic spatial and spatiotemporal correlations. These correlations serve as truth for quantifying the error that would be made with smaller subsets of the full ensemble. Previous studies used the same assumption as a basis for studying sampling errors but with smaller ensemble sizes (Hamill et al., 2001; Bannister et al., 2017). The present 1000-member ensemble is also used for a more detailed evaluation of sampling errors and correction methods in Necker et al. (2019).

As a second step, we present an approach to develop observation and DA strategies based on ensemble sensitivity analysis (ESA; Ancell and Hakim (2007)). This approach uses spatiotemporal correlations as a proxy for the potential impact of observable quantities. The main focus of our study is to assess the sensitivity of precipitation to a selection of model quantities. Precipitation is chosen as it is a primary forecast quantity of convective-scale forecasting systems. ESA has successfully been used in various synoptic-scale (Hakim and Torn, 2008; Torn and Hakim, 2009; Torn, 2010; Kunkel et al., 2013; Barrett et al., 2015) and convective-scale studies (Bednarczyk and Ancell, 2015; Wile et al., 2015; Hill et al., 2016; Limpert and Houston, 2018). However, most ESA studies relied on fairly small ensembles sizes and focused on the qualitative interpretation of sensitivities. The large ensemble simulation of the present study can be used to quantify the contribution of sampling errors for ESA.

This manuscript is outlined as follows: Section 2 provides details on the 1000-member ensemble experimental setup and introduces how the potential impact can be estimated using ESA. Section 3 splits into three parts: First, we compare the 1000-member ensemble simulations to a smaller, but well-tuned independent operational modelling system. Second, we analyze sampling errors and discuss localization in DA using spatial correlations. Third, we show how spatiotemporal correlations can be used as a proxy for the potential impact of observable quantities and discuss sampling errors in this context. A summary with conclusions follows in Section 4.

2 | EXPERIMENTAL SETUP AND METHODS

2.1 | Ensemble simulations

2.1.1 | SCALE-RM 1000-member ensemble

Our experiment comprises a set of ten ensemble forecasts during summer 2016 with 1000 ensemble members and forecast lead times of 14h. Forecasts are generated coupling two domains through an offline nesting approach (see flow-chart in Fig. 1a). The outer domain is used for the 15-km grid spacing cycled ensemble data assimilation and driven by Global Ensemble Forecast System (GEFS[1]; NCEP) boundary conditions. Initial conditions for the inner domain are obtained by downscaling from 15-km to 3-km grid spacing. The convective-scale forecasts are driven by additional forecasts performed in the 15-km mesh size outer domain. The GEFS system consists of 20 ensemble members generated using an ensemble transform with scaling approach (Ma et al., 2014) and stochastic total tendency perturbation to represent model errors. Ensemble members are integrated with the Global Forecast System (GFS) model with an spectral resolution of T564 and 64 vertical levels. Output data is available every 6 hours in a regular grid

[1] https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/global-ensemble-forecast-system-gefs

with a horizontal resolution of 0.5 degree.

In detail, initial conditions for the 15 km cycled experiment on 28 May 2016 00 UTC are taken from a previous 1000-member DA experiment over the same domain that has been spun-up for one week. The 3-h cycling is continued for one week till 03 June 2016 and includes 56 cycles. Perturbed boundary conditions are provided every 6 hours and are generated combining the GEFS 20-member analysis ensemble with 1000 randomly generated perturbations. The *i-th* random perturbation is added to the *j-th* GEFS ensemble member, where $j = i - 20 * floor(\frac{i-1}{20})$, so the *i-th* ensemble boundary condition is a combination of a unique random perturbation with the *j-th* GEFS ensemble member. At the beginning of the cycle, these perturbations are obtained as the difference between two random atmospheric states that correspond to the same season and time of the day. In the following cycles, the perturbations are updated using the following rule: If at time *t* the *i-th* random perturbation is generated by computing the difference between atmospheric states at times $t^i{}_1$ *and* $t^i{}_2$, the perturbation at time *t+dt* is computed as the difference between atmospheric states at times $t^i{}_1 + dt$ and $t^i{}_2 + dt$. This guarantees a smooth evolution of the random perturbations. Before applying the perturbations to the boundary conditions, their amplitudes are re-scaled by a multiplicative factor equal to 0.1. This re-scaling factor is chosen to significantly reduce the amplitude of the perturbations which, otherwise will be equal to twice the climatological variability of the state. Atmospheric states for the computation of random perturbations are obtained from the Climate Forecast System Reanalysis (CFSR) data-set (Saha et al., 2010) in the period between 2006 and 2009.

Our simulation applies the SCALE-LETKF DA system (Lien et al. (2017)). The SCALE-LETKF system combines the open source Scalable Computing for Advanced Library and Environment - Regional Model (SCALE-RM; version 5.1.2) (Nishizawa et al., 2015; Sato et al., 2015; Nishizawa and Kitamura, 2018) and a Localized Ensemble Transform Kalman Filter (LETKF) (Hunt et al., 2007). The LETKF assimilates conventional observations using a 3-hourly assimilation window on the 15 km grid. The localization is done with an R-localization approach (Greybush et al., 2011) using a Gaussian function with a fixed localization scale of 120 km in the horizontal and $0.3/ln(p)$ in the vertical and a cut-off radius equal to $2\sqrt{10/3}$ times the localization scale. The ensemble spread is inflated using relaxation to prior spread (RTPS) with a relaxation coefficient of 0.8 (Whitaker and Hamill, 2012). Figure 1b shows the 15-km mesh size cycling domain that is centered over Germany. The outer domain extends over an area of 100×100 grid points and exhibits 31 vertical levels. The model physics configuration is similar as in (Lien et al., 2017; Honda et al., 2018). All the experiments use the Tomita (2008) single-moment bulk microphysics scheme, the Mellor-Yamada-Nakanishi-Niino 2.5 closure boundary layer scheme (Nakanishi and Niino, 2004), the Model Simulation Radiation Transfer code for the representation of radiative fluxes (Sekiguchi and Nakajima, 2008) and the Beljaars-type surface model (Beljaars and Holtslag, 1991) for the computation of soil variables and surface fluxes.

The convective-scale 14-h forecasts are computed in the inner forecast domain (Fig. 1b). The convective-scale domain measures 350×250 grid points with a 3 km mesh size and 30 vertical levels. The initial conditions for each forecast are downscaled from the 15 km analysis to 3 km mesh size (cold-start approach). Additional lower-resolution 15 km mesh size 15-h forecasts provide frequent (hourly) boundary conditions for the convective-scale forecasts (Fig. 1a). Our study analyzes a total of ten consecutive 1000-member ensemble forecasts that are initialized by downscaling every 12 hours from 00 UTC 29 May to 12 UTC 02 June 2016. All simulations have been performed on the K-computer at the RIKEN Center for Computational Science in Kobe, Japan (Miyoshi et al., 2016a,b).

2.1.2 | COSMO-DE 40-member ensemble

The SCALE-RM convective-scale 1000-member ensemble simulation is compared to regional forecasts computed with the operational forecasting system of Deutscher Wetterdienst (DWD). The DWD DA system is composed of

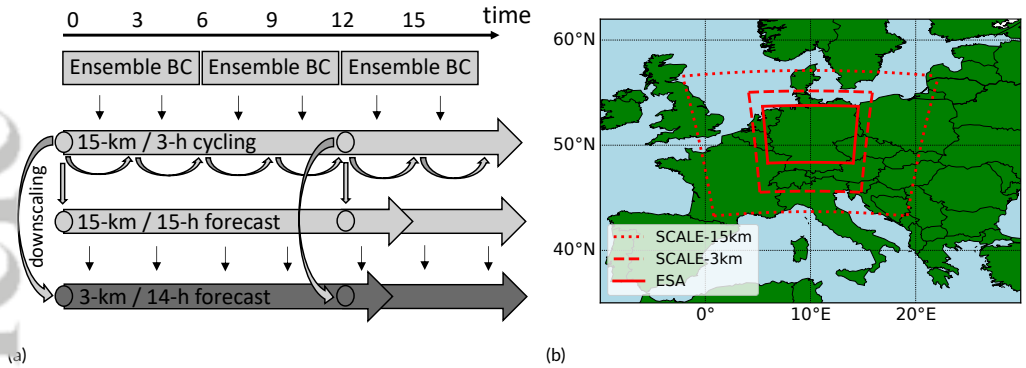


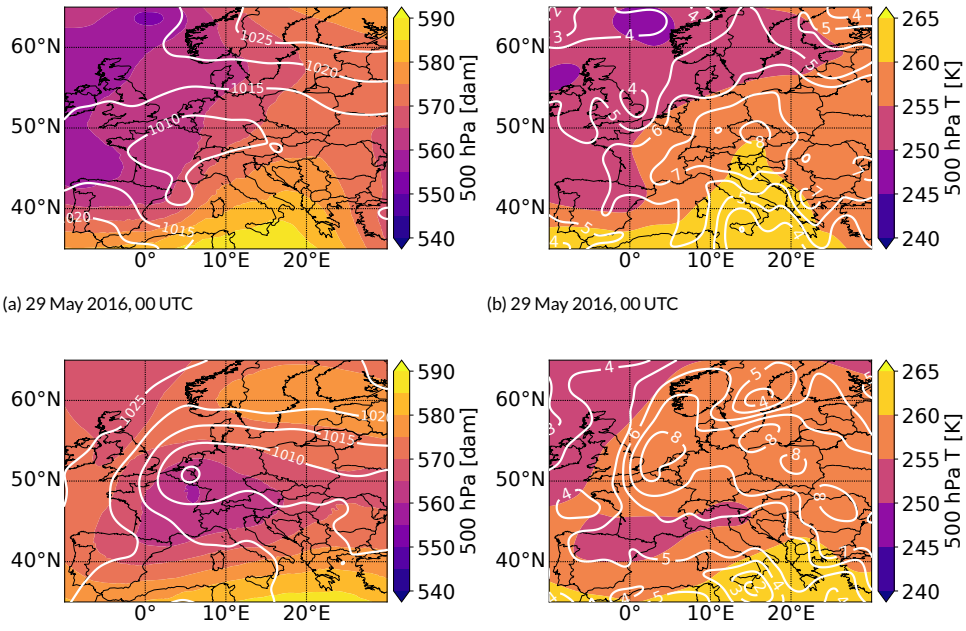
FIGURE 1 (a) Flow-chart of the 1000-member ensemble simulation setup. (b) Experimental domains used for 15 km cycling and forecasts (dotted), 3 km forecasts (dashed), and ensemble sensitivity analysis (solid).

the regional NWP model COSMO-DE (Baldauf et al., 2011) and the Km-scale ENsemble Data Assimilation (KENDA; Schraff et al. (2016)) system. KENDA is based on a 40-member LETKF. In contrast to the operational setup of DWD, latent heat nudging is switched off (no radar observations are used) and only conventional observations are assimilated. The horizontal (100 km) and vertical (ln(p)=0.3) localization scales are similar to the SCALE-LETKF system 1000-member experiment, but localization is done using a Gaspari-Cohn-function (Gaspari and Cohn, 1999). COSMO-DE is a non-hydrostatic limited-area forecast model and has a horizontal grid spacing of 2.8 km with 50 vertical levels. COSMO-KENDA simulations are driven by the global ICON ensemble (Zängl et al., 2015) that has a horizontal resolution of approximately 16 km. Similar to the 1000-member ensemble simulation, COSMO-DE 40-member 12-h forecasts are initialized twice a day at 0 and 12 UTC during the same five-day summer period in 2016, but COSMO-DE uses an LETKF instead of downscaling. The COSMO-DE domain extends over approximately 1200×1300 km and has the same domain center as the SCALE-RM domain. Both simulations are compared using an almost overlapping domain, which measures $\times 200$ grid points for both models (see the innermost domain in Fig. 1b). A difference of approximately 10% in the domain size originates from the unequal horizontal grid spacing of both models (2.8 km vs. 3 km). However, this should hardly affect the analysis carried out in this study as the domain size difference is small and the analysis is performed sufficiently far from the boundaries in both cases. Further details on the COSMO-KENDA simulation can be found in Necker et al. (2018).

2.2 | Synoptic situation

Figure 2 shows the general weather situation during the experiments. The five-day period was largely determined by an atmospheric blocking over the Atlantic ocean leading to a fairly stationary weather situation over central Europe (Piper et al., 2016). An upper-level trough accompanied by a shallow surface low was located over the experimental domain (Fig. 2a). The low-pressure system stayed almost stationary over France and Germany and reached its minimum pressure on 30 May (Fig. 2c). This led to a highly unstable environment with weak pressure gradients and synoptic-scale flow that changed from southerly (29./30. May) to easterly (31. May and 1./2. June).

At the beginning of the experimental period (Fig. 2b and 2d), the low-level advection of moist and warm air masses from southern Europe increased the thermal instability over Germany. Both strong surface heating, as well as convective instability, forced the development of deep convection and thunderstorms on all five days. In addition, low wind



(a) 29 May 2016, 00 UTC

(b) 29 May 2016, 00 UTC

(c) 31 May 2016, 00 UTC

(d) 31 May 2016, 00 UTC

FIGURE 2 (a, c) ECMWF IFS analysis of geopotential height at 500 hPa (shaded, dam) and contour lines of sea level pressure (white contour, hPa). (b, d) ECMWF IFS analysis of temperature at 500 hPa (shaded, K) and contour lines of specific humidity (white contour, g/kg).

speeds at 500 hPa led to several slow-moving cells causing locally extreme precipitation. The highest number of severe precipitation events occurred on 29 May 2016 (Fig. 3d) producing flash floods, landslides, hail, and tornadoes over southern Germany. In some regions, the rainfall exceeded an amount of 100 mm per day. Observed thunderstorms showed a distinct diurnal cycle peaking in the late afternoon. Overall, the five-day period and adjacent days were characterized by strong convective precipitation. This provides a unique period that was also investigated as a test case in several other studies using the COSMO-KENDA system (Rasp et al., 2018; Necker et al., 2018; Keil et al., 2019; Baur et al., 2018; Bachmann et al., 2019).

2 | Methods

2.1 | Sampling error correction (SEC)

Anderson (2012) introduced a statistical sampling error correction (SEC) that corrects for the overestimation of correlations due to spurious correlations. The sampling error corrected correlation r_{sec} can be obtained by

$$\hat{r}_{sec} = \gamma_{m,p}(\hat{r})\hat{r}, \tag{1}$$

where $\gamma_{m,p}(\hat{r})$ is provided by a look-up table for a given ensemble size m and a prior distribution p given the sample

correlation $\hat{r}$. The sampling error correction is calculated with an offline Monte Carlo technique assuming that the prior p is uniformly distributed in the range [-1, 1]. The SEC is only a function of the sample correlation $\hat{r}$, the ensemble size m and the prior p. A detailed description of the calculation of the look-up table is documented in Anderson (2012). Our study applies a SEC table that is provided within the Data Assimilation Research Testbed (DART; Anderson et al. (2009)) that was calculated independently of our simulation. Given that the look-up table has already been computed for the ensemble size, only the sample correlation $\hat{r}$ is required to obtain the corrected correlation r_{sec}.

2.3.2 | Ensemble sensitivity analysis and potential impact

Ancell and Hakim (2007) introduced ensemble sensitivity analysis (ESA) as the sensitivity S of a forecast response function $\mathbf{J}$ to the initial conditions $\mathbf{x}$

$$S = \frac{\partial J}{\partial x} \approx \frac{cov_m(\mathbf{J}, \mathbf{x})}{var_m(\mathbf{x})}, \tag{2}$$

where $\mathbf{J}$ and $\mathbf{x}$ are vectors consisting of m ensemble estimates of scalar quantities J and x. Here, cov_m denotes the sample covariance between two quantities and var_m denotes the sample variance of one quantity. A normalization of the sensitivity S with the ratio of the ensemble spread of the forecast response function $\mathbf{J}$ to the spread of the state variable of interest $\mathbf{x}$ provides the dimensionless correlation $\hat{r}$ that can be compared for different variables

$$\hat{r} = \frac{cov_m(\mathbf{J}, \mathbf{x})}{var_m(\mathbf{x})} \frac{\sqrt{var_m(\mathbf{x})}}{\sqrt{var_m(\mathbf{J})}} = \frac{cov_m(\mathbf{J}, \mathbf{x})}{\sqrt{var_m(\mathbf{J}) var_m(\mathbf{x})}}. \tag{3}$$

One goal of this study is to estimate the relative potential impact of observable quantities using spatiotemporal correlations. We take the squared correlations accumulated over the evaluation domain and all response functions of interest as proxy for the potential impact of the respective observable quantity on a precipitation forecast. This gives us the accumulated squared correlation (ASC):

$$ASC = \sum_{i=1}^{n} \sum_{I=1}^{N} \left(\hat{r}_{i,I}\right)^2. \tag{4}$$

where

$$\hat{r}_{i,I} = \frac{cov_m(\mathbf{J}_i, \mathbf{x}_I)}{\sqrt{var_m(\mathbf{J}_i) var_m(\mathbf{x}_I)}}. \tag{5}$$

with index $i = 1...n$ (n - number of forecast response functions) and index $I = 1...N$ (N - number of grid points). As forecast response function $\mathbf{J}$, we use the precipitation forecast spatially averaged over squares of 40×40 grid points. The squares/response functions do not overlap and cover the entire ESA domain.

2.3.3 | Confidence test (T95)

A confidence test is applied to detect and exclude insignificant correlations from ESA (Torn and Hakim, 2008). This is required to evaluate if a state variable $\mathbf{x}$ is able to cause a statistically significant change in the forecast response

function **J**

$$\left|\frac{cov_m(\mathbf{J},\mathbf{x})}{var_m(\mathbf{x})}\right| > \delta_s, \tag{6}$$

where δ_s is the confidence interval on the linear regression coefficient. For a given sample, we compute the sensitivity that allows us to reject the null hypothesis that there is no correlation between the response function **J** and the state variable **x** with a defined confidence level. In this study, we use a 95% confidence level (T95). Insignificant correlations are not considered in the computation of the ASC.

3 | EVALUATION OF THE ENSEMBLE SIMULATIONS AND ADDED VALUE BY LARGER ENSEMBLE SIZES

3.1 | Ensemble mean and spread

First, the 1000-member ensemble forecast is compared against the COSMO-DE forecast and independent radar-derived precipitation observations. The radar-based precipitation product (RADOLAN; EY-product) covers most parts of central Europe and delivers frequent observations over Germany (Fig. 3a). The main focus of the comparison is to assess if the 1000-member ensemble captures the precipitation and provides realistic spread as this is crucial for estimating potential impact using ensemble correlations.

3.1.1 | Precipitation

Examples of regional distribution of precipitation

Figure 3a displays a map of the 1-h accumulated precipitation observations for 29 May 2016, 04 UTC. The radar composite shows a precipitation event over northern Germany as well as several scattered smaller cells over France, Switzerland, and southern Germany. The 4-h COSMO-DE ensemble mean forecast captures the precipitation event over northern Germany while the precipitation over France and Switzerland is slightly overestimated and the precipitation in south-west Germany is underestimated (Fig. 3b). The SCALE-RM 1000-member ensemble forecast also predicts precipitation over northern Germany (Fig. 3c), but exhibits smaller precipitation amounts. Furthermore, SCALE-RM does not reveal enough precipitation over southern Germany even though some individual members showed precipitation in this area (not shown).

Figure 3d shows the precipitation observations for 29 May 2016 18 UTC, which was the strongest precipitation event occurring in the entire experimental period. On this day, both mesoscale and synoptic-scale lifting led to the development of severe thunderstorms that produced hail and rain-rates locally exceeding 20 mm per hour. The main precipitation event took place over southern Germany, although additional cells have been observed all over central Europe. The 6-h COSMO-DE ensemble mean precipitation forecast (Fig. 3e) covers the region of maximum precipitation but also predicts precipitation in many other parts of the domain. The region of severe precipitation is smaller, weaker, and slightly shifted to the north compared to the radar observation. The ensemble mean of SCALE-RM (Fig. 3f) underestimates the intensity of this unique event even more and shows larger precipitation over Switzerland and Austria. However, some members were at least able to produce precipitation rates close to the ensemble mean precipitation of COSMO-DE in the area of the maximum observed precipitation (not shown).

Figure 3g shows the precipitation observations at 30 May 2016, 16 UTC. At that time, an elongated precipitation region is visible over northern Germany. COSMO-DE can predict the approximate structure and intensity of the precipitation

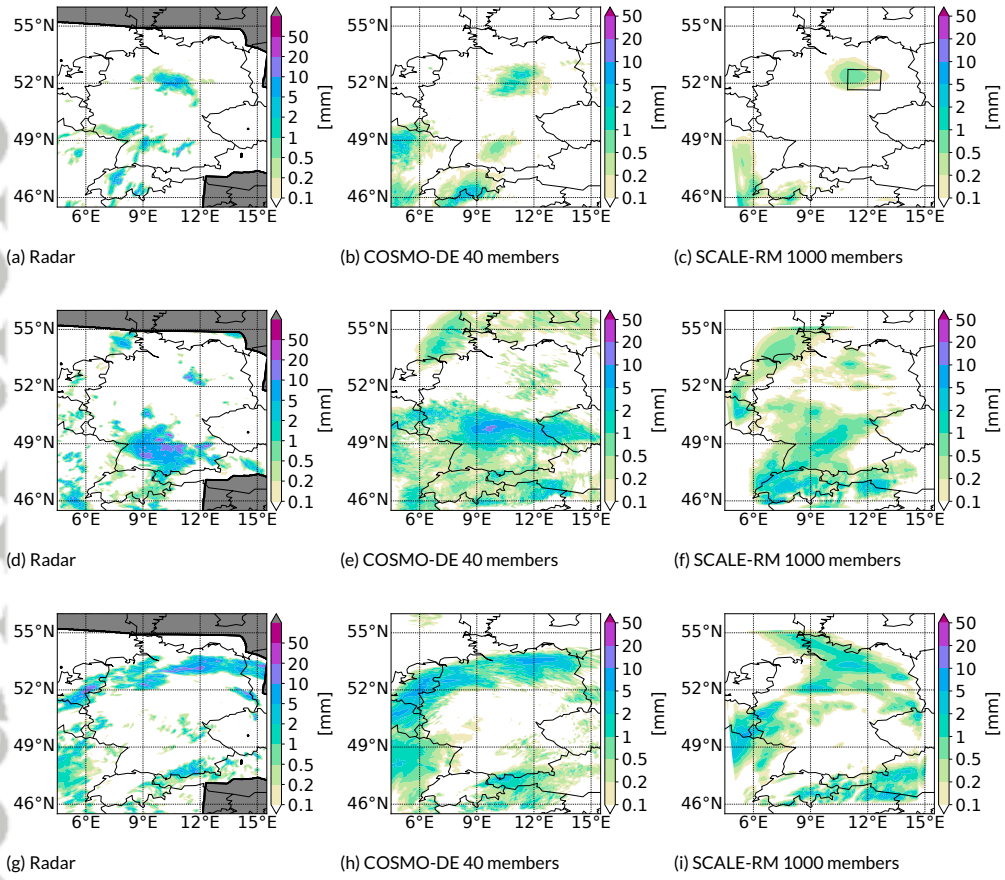


FIGURE 3 Hourly accumulated precipitation as estimated by the radar network (a,d,g) as well as COSMO-DE -member (b,e,h) and SCALE-RM 1000-member (c,f,i) ensemble mean precipitation for 29 May 2016 04 UTC (top row), 18 UTC (middle row) and 30 May 2016 16 UTC (bottom row). The forecast lengths for the model simulations -DE and SCALE-RM) are 4-h (b,c,h,i) and 6-h (e,f), respectively.

ent, but there is some uncertainty on the exact position among the ensemble members (Fig. 3h). In the SCALE-RM simulation (Fig. 3i), the precipitation band moved too slowly and is located approximately 100 km south of the observed tion. Nevertheless, SCALE-RM is able to overall capture the precipitation band as well as the precipitation over France and Austria.

summary, COSMO-DE provides more accurate precipitation forecasts than SCALE-RM, which is likely due to the high-resolution data assimilation incorporated in COSMO-KENDA and better tuning for the region of interest. Nevertheless, the SCALE-RM forecasts overall provide realistic precipitation amounts and patterns, which is an important prerequisite for studying spatial and spatiotemporal correlations based on this data set. A more detailed analysis of this ensemble simulation including the investigation on non-Gaussianity and more sphisticated measures will be performed in a subsequent study.

Temporal evolution

Figure 4a shows the temporal evolution of the domain mean precipitation during all ten forecasts. Both ensemble mean and spread are investigated for the innermost domain (see Fig. 1b). The radar-derived domain mean precipitation is again used as a reference for both ensemble simulations. All five days featured strong precipitation events and showed a diurnal cycle in the precipitation amount peaking in the afternoon. As discussed previously, most severe thunderstorms occurred in the afternoon of 29 May 2016 indicated by the highest domain average precipitation. COSMO-DE well captures the temporal evolution of the precipitation peaking in the afternoon at a similar time and with a similar intensity as in the radar observation. Nevertheless, the COSMO-DE ensemble is not able to predict the intensity of the severe rainfall on 29 May 2016.

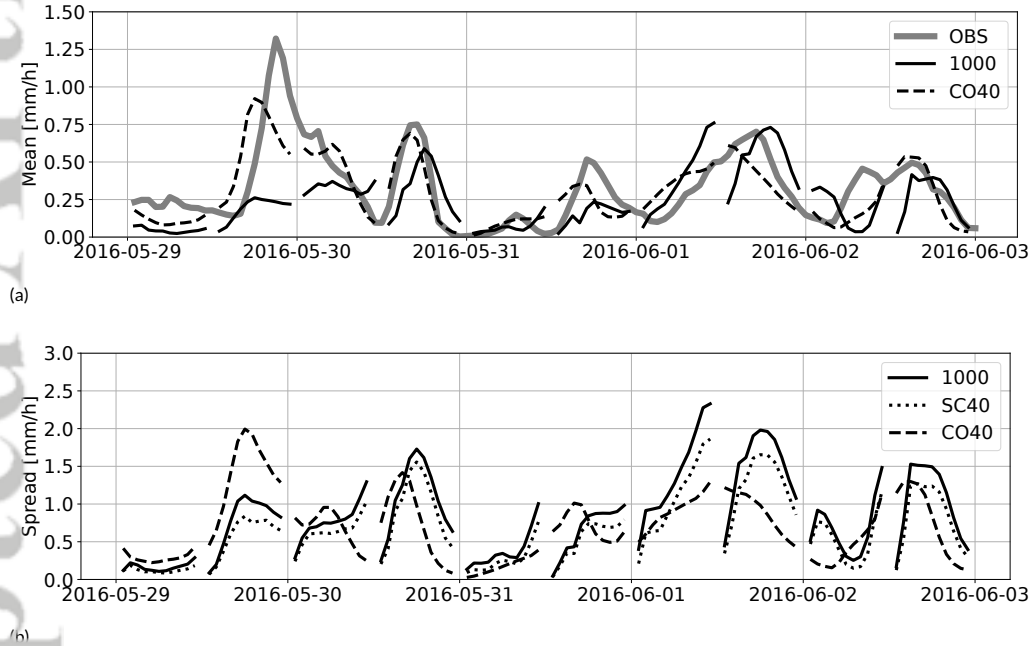


FIGURE 4 (a) Domain mean hourly accumulated radar derived precipitation observation (solid grey) as well as SCALE-RM 1000-member (solid black) and COSMO-DE 40-member (dashed black) 12-h ensemble precipitation forecasts, 29 May 2016 00 UTC till 03 June 2016 00 UTC. (b) Domain mean spread of the hourly accumulated precipitation forecasts for different ensemble samples and the same period (SCALE-RM 40-member ensemble, dotted black).

SCALE-RM reproduces the diurnal cycle of precipitation similarly to COSMO-DE but less accurately. Both timing and amplitude of the peaks are slightly different from the radar observations, especially at the beginning of the experimental period. As discussed previously, one reason is that SCALE-RM was not able to fully predict the correct intensity of the severe thunderstorms over southern Germany. Additionally, most members exhibited their strongest precipitation over the Alps, but this region is not included in the verification domain (Fig. 1b and 3f). Nevertheless, some members revealed a three times stronger precipitation than the ensemble mean.

Overall, the first forecast hour after each analysis should be treated with caution (see Fig. 4). The re-initialization of a

new forecast in some cases leads to an underestimation of precipitation at the beginning of the forecast. This spin-up effect is especially visible for the SCALE-LETKF system at the beginning of June. The cause of this effect seems to originate from the downscaling, which is required to obtain the high-resolution initial conditions. SCALE-RM requires a few model iterations to develop small-scale structures, as well as to diagnose sufficient precipitation from the prognostic variables, while in the COSMO-KENDA system the precipitation amount is almost preserved after the analysis. This is reasonable recalling that for the COSMO-KENDA system the analysis is obtained on the same resolution as the forecast (warm-start initialization; 2.8 km), while the SCALE-RM simulation is based on a cold-start approach that includes a downscaling of the initial condition from a coarser grid.

Figure 4b shows the temporal evolution of the domain mean spread of hourly precipitation. The diurnal cycle is visible in the spread peaking in the late afternoon and showing a smaller amplitude during the night. Except for the first day of the experimental period, the SCALE-RM 1000-member ensemble exhibits the largest spread of all ensemble simulations. A random 40-member subset of the 1000-member ensemble is additionally included in the comparison to assess if the ensemble spread strongly changes with the ensemble size as well as to compare COSMO-DE and SCALE-RM using an equal ensemble size. The SCALE-RM 40-member ensemble usually reveals a smaller spread resulting from an under-sampling of the true variance. Consequently, the spread of the SCALE-RM 40-member ensemble is often closer to the spread of COSMO-DE. As discussed previously, initializing SCALE-RM from the downscaled analysis reduces the spread at the beginning of most forecasts.

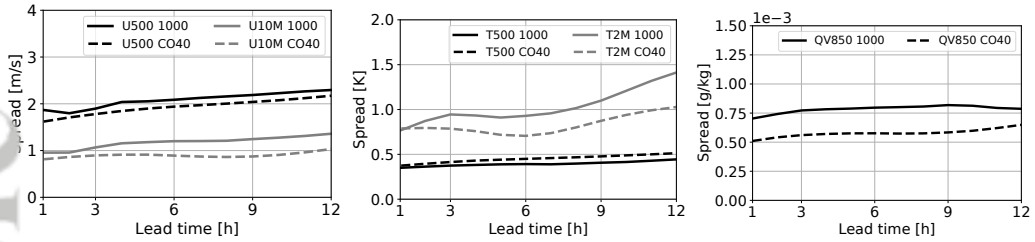
Overall, the SCALE-RM 1000-member ensemble delivers fairly realistic precipitation forecasts regarding ensemble mean and spread. The amount and timing of precipitation events do not necessarily need to coincide with an operational forecasting system or observations as ensemble sensitivity analysis or the analysis of sampling errors does not incorporate observations and therefore only requires realistic scenarios. The first forecast hour has been ignored for the ensemble evaluation to exclude potential spin-up effects originating from downscaling. For this reason, the 1-h forecast is used as an initial state x to compute ensemble sensitivities with respect to precipitation.

3.1.2 | Growth of spread for prognostic model variables

Figure 5 displays the evolution of the domain mean ensemble spread with forecast lead time for different prognostic model variables. For simplicity and as the focus is on the growth of the ensemble spread, the available ten forecasts have been averaged temporally. The 1000-member ensemble spread of 10-m and 500 hPa zonal wind (Fig. 5a) is slightly larger than for COSMO-DE. For both simulations, the ensemble spread increases equally fast throughout the forecast, while the upper-air spread is larger than that close to the surface. The ensemble spread of 500 hPa temperature (Fig. 5b) hardly increases with lead time and coincides roughly for both simulations. In contrast to the zonal wind, the ensemble spread close to the surface is larger than in the middle troposphere. Initially SCALE-RM and COSMO-DE exhibit a similar surface temperature spread, which increases stronger in the 1000-member ensemble simulation. The ensemble spread for 850 hPa specific humidity (Fig. 5c) is also higher for the SCALE-RM 1000-member ensemble, while the COSMO-DE ensemble spread increases slightly faster. As indicated by surface temperature and 850 hPa specific humidity, SCALE-RM exhibits a larger variability close to the surface compared to COSMO-DE. For all prognostic model variables, the evolution of the ensemble spread is reasonable and well simulated by the SCALE-RM ensemble.

3.1.3 | Spectral analysis of variances

The spectral variance is analyzed to examine the structure and amplitude of the ensemble perturbations and their dependence on different boundary perturbation and data assimilation schemes. Figure 6 compares spectra of variance



(a) Zonal wind (b) Temperature (c) Specific humidity

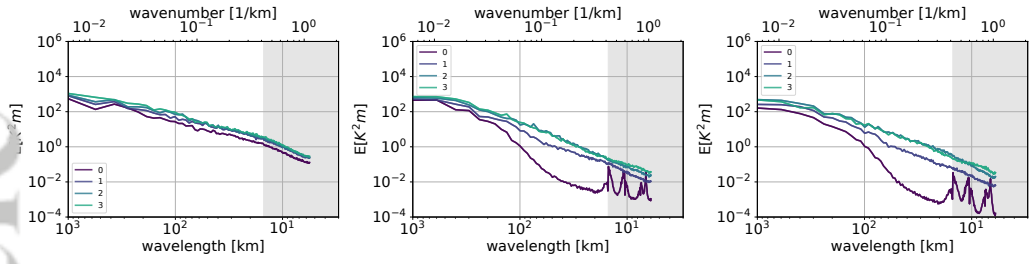
FIGURE 5 Time-averaged domain mean spread as function of lead time for prognostic variables, 29 May to 03 June 2016. Comparison between SCALE-RM 1000-member and COSMO-DE 40-member ensemble. (a) 500 hPa and 10-m zonal wind (b) 500 hPa and 2-m temperature, and (c) 850 hPa specific humidity.

fields of 850 hPa temperature for three different ensemble setups and four forecast lead times. The spectra are obtained for the innermost square ESA domain (see Fig. 1b). The one-dimensional spectra $E(k)$ are computed performing a 2D Fourier transform on variance fields. The Fourier coefficients are summed up over annuli in wavenumber space. All spectra are averaged over six different 1000-member ensemble forecasts that are initialized from 29 to 31 May 2016 every day at 0 and 12 UTC. The analysis only includes six forecasts to allow a comparison of three different ensemble simulations: the COSMO-DE 40-member ensemble (Fig. 6a), the SCALE-RM 1000-member ensemble including (Fig. 6b) and excluding (Fig. 6c) GEFS perturbations. The latter comparison is done to examine the impact of both random/climatological perturbations and GEFS perturbations on the SCALE-RM 1000-member ensemble simulation. The analysis is exemplarily discussed for 850 hPa temperature similar to Kuehnlein et al. (2014).

Figure 6a displays the temporally averaged variance spectra of the COSMO-DE 40-member ensemble. Given a convective-scale DA system, the shape of the variance spectra for the first model integration (0 h forecast) is already fairly similar to longer lead times. In contrast, the SCALE 1000-member ensemble exhibits a more pronounced spin-up (Fig. 6b), which is related to using downscaled initial conditions. The spectrum computed for the first model integration (0 h forecast) shows a clear lack of small-scale variability. Furthermore, the spectrum contains several peaks at scales smaller than 15 km that also seem to arise from the downscaling from 15 km to 3 km mesh size. The model requires approximately two hours to fully develop small-scale perturbations. However, most of the spin-up is already finished within the first forecast hour. The SCALE-RM simulation has a slightly smaller variability on meso and convective scales compared to COSMO, which partly seems to be a consequence of the downscaling.

Figure 6c displays the spectral variance for a second SCALE-RM 1000-member ensemble simulation that only applies random/climatological boundary perturbations. The comparison to the main SCALE-RM simulation (that applies both random and GEFS perturbations, Fig. 6b) shows that including the GEFS perturbations adds variability on all scales. Consequently, with the GEFS perturbations, the spin-up time is reduced and realistic small-scale perturbations develop earlier. After the spin-up, the differences due to the different perturbation settings seem to be small.

Overall, the spectral analysis shows the benefit of combining both GEFS and random perturbations to achieve realistic ensemble perturbations comparable to COSMO-DE. For most quantities, realistic small-scale perturbations develop within the first forecast hour. This growth saturates within two hours lead time. To exclude forecast steps that are affected by the spin-up and to ensure realistic ensemble perturbations, the 1h-forecast is used as initial state for the computation of spatial and spatio-temporal correlations.



(a) COSMO-DE 40 (b) SCALE-RM 1000 (c) SCALE-RM 1000 - no GEFS perturbations

FIGURE 6 Time-averaged spectral variance of temperature at 850 hPa for different forecast lead times (0, 1, 2, and 3 h) and ensemble setups, 29 to 31 May 2016. The 0-h forecast represents the first model integration. The gray shading indicates spatial scales smaller than 15 km. (a) COSMO-DE 40-member ensemble, SCALE-RM 1000-member ensemble (b) including and (c) excluding GEFS perturbations.

3.2 | Spatial correlations

As a second step, we focus on spatial correlations. Such correlations are crucial for hybrid and ensemble DA systems, which rely on accurate correlation estimates for spreading the information from observations spatially and among different variables. We show results for three different ensemble sizes: A small (40 members), medium-sized (200 members) and large (1000 members) ensemble. The subsets are drawn from the 1000-member ensemble with the constraints that each member of the 20-member GEFS boundary perturbations is represented equally often and that the 40-member subset is included in the 200-member subset. Spatial correlations between different grid points and variables are calculated using the 1-h forecast state, which is similar to using the first guess during hourly cycling.

3.2.1 | Horizontal correlation

Figure 7 displays the mean absolute correlation and error as a function of spatial distance. Here, the mean absolute correlation (MAC_m) and error (MAE_m) for a target distance are given by

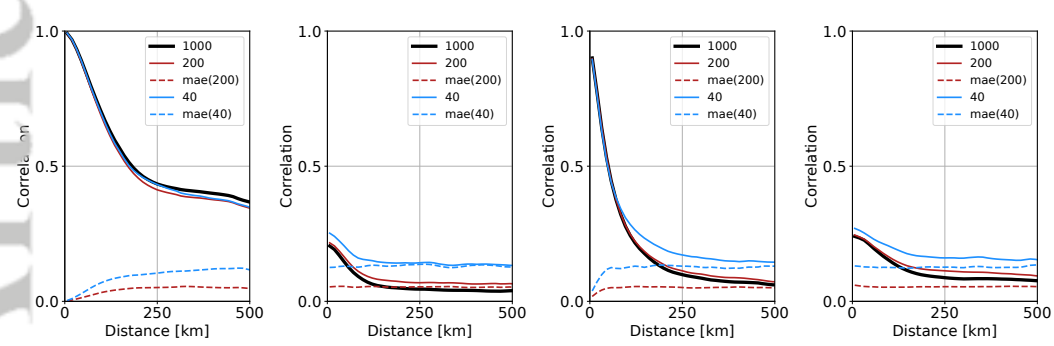
$$MAC_m = \frac{1}{N}\sum_{n=1}^{N}\left|r_m^n\right|$$

and

$$MAE_m = \frac{1}{N}\sum_{n=1}^{N}\left|r_m^n - r_{1000}^n\right|,$$

where m describes the ensemble size and N specifies the number of grid points in a defined distance range that are used to compute the mean correlation and error. Distances are binned in steps of 13 km. The error here is the error in the correlation assuming the 1000-member correlation as the truth. For each forecast, we evaluate correlations from nine different grid points to all other grid points in the domain. These nine grid points are evenly distributed in the domain and lie at least 150 km apart from each other. The error at each grid point is calculated with respect to the 1000-member ensemble correlation and results are also averaged over all ten forecasts.

Fig. 7a shows the mean absolute correlation of the full 1000-member ensemble (MAC_{1000}) of 500 hPa temperature to 500 hPa temperature itself as a function of horizontal distance. On short distances, tropospheric temperatures are highly correlated. The MAC_{1000} decreases to 0.5 at a distance of 200 km. From 200 to 500 km, the spatial de-correlation continues but with a weaker gradient. Both, the MAC_{40} and MAC_{200} coincide with the MAC_{1000}, slightly underestimating the correlation at large distances. The 200-member subset exhibits approximately half the mean absolute error (MAE) of the 40-member subset, but the error of both samples is much smaller than the MAC. The MAE_{200} increases slower with distance than MAE_{40} and seems to be saturated after about 150 km.



(a) T 500 hPa to T 500 hPa (b) T 500 hPa to QV 500 hPa (c) T 2 m to T 2 m (d) T 2 m to U 10 m

FIGURE 7 Mean absolute correlation (MAC; solid) and error (MAE; dashed) as function of spatial distance [km] for 1000, 200 and 40 members. Correlations of 500 hPa temperature to (a) 500 hPa temperature and (b) 500 hPa specific humidity as well as correlations of 2-m temperature to (c) 2-m temperature, and (d) 10-m zonal wind.

Cross-correlations of temperature to specific humidity (Fig. 7b) are substantially weaker. The MAC_{1000} exhibits a maximum value of about 0.22, decreases up to a distance of approximately 100 km and remains constant farther away. The 200-member ensemble roughly estimates the shape of the MAC_{1000}, while 40 members substantially overestimate the true correlation due to spurious correlations. The MAC_{40} is approximately three times larger compared to the MAC_{1000} after a distance of 150 km. For the 40-member ensemble, correlations of distances longer than 50 km are not trustworthy as their error exceeds the absolute value of the MAC_{1000}. This cross-over point roughly indicates a suitable choice as localization scale in data assimilation, but the applied localization scale may also be restricted by other considerations as e.g. consistency for different variables or the number of observations within the localization scale for an LETKF. Using 200 members almost doubles the distance of this cross-over point compared to 40 members.

Figure 7c shows the spatial correlation of 2-m temperature to 2-m temperature itself. As for upper air temperatures, the MAC_{1000} is large on short distances but decreases faster and is weaker at longer distances. The 200-member ensemble almost coincides with the MAC_{1000} and the error does not exceed the MAC_{1000} before reaching 500 km distance. The MAC_{40} agrees with the MAC_{1000} up to a distance of about 100 km, but around 200 km the error starts to get larger than the absolute value.

The MAE of cross-correlations of 2-m temperature to 10-m zonal wind (Fig. 7d) is fairly constant at all distances, while the corresponding MAC is much smaller than for the correlation of 2-m temperature to 2-m temperature. As a consequence, the MAE_{40} exceeds the MAC_{1000} at a distance of slightly over 100 km, while the MAE_{200} remains below up to a distance of 500 km.

In summary, these examples show that the 1000-member ensemble can be used to quantify sampling errors as well

as to investigate suitable choices for localization length scales in convective-scale NWP. The different results for different variables highlight that it would be desirable to select very different scales for different model variables and combinations of variables. Furthermore, the results show a big advantage of correlations from the 200-member subset compared to 40 members.

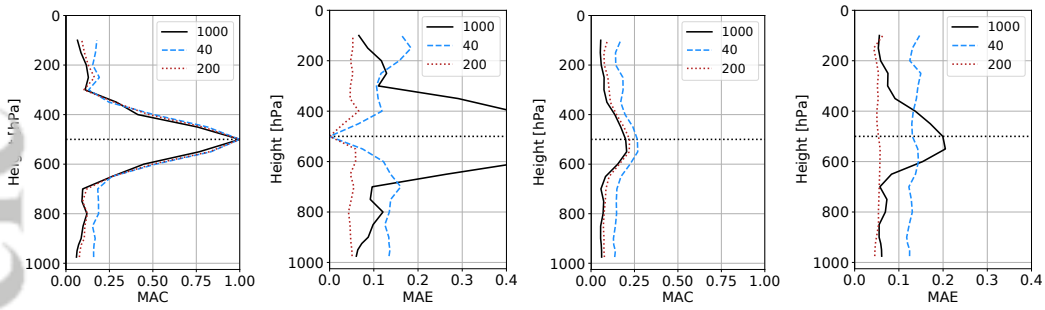
3.2.2 | Vertical correlation

Vertical correlations are evaluated using one 1000-member ensemble forecast at 30 May 2016 and vertical profiles at all grid points in the domain. Given a high number of vertical correlations in the domain under various atmospheric conditions (40.000 vertical columns), the analysis is presented for a single date. Figure 8a shows the mean absolute correlation of temperature at 500 hPa to temperature at other levels. The MAC_{1000} exhibits a correlation of 1 at the response level and rapidly decreases to a value of 0.25 reaching a vertical distance of 150 hPa. Levels close to the ground and tropopause are hardly correlated with 500 hPa. The 200-member ensemble again roughly coincides with 1000 members with only a slight overestimation of the true correlation. The 40-member ensemble gives similar results close to the response level but overestimates the absolute correlation above and below by a factor two.

Next, we examine the horizontally averaged mean absolute error (MAE) of the correlation (assuming the 1000-member correlation as the truth) as a function of height (Fig. 8b). The 200-member ensemble exhibits a small sampling error of about 0.05, except for the response level and the two neighboring levels above and below. The MAE_{40} exhibits a substantially higher error with values that are up to three times higher at distances of more than 100 hPa. Comparing the amplitudes of sampling errors for both subsets with the MAC_{1000}, the MAE_{200} hardly exceed the true correlation. In contrast, the MAE_{40} increases faster with distance and exceeds the MAC_{1000} 200 hPa above and below the response level. Consequently, using a 40-member ensemble would require a narrow vertical localization to reduce the impact of spurious correlations. For temperature, the width (in hPa) of the required vertical localization hardly changes with the height of the chosen response level (not shown).

Figure 8c displays vertical cross-correlations of 500 hPa temperature with specific humidity at other levels, and Figure 8d shows its sampling errors. The MAC_{1000} is generally weak and exhibits a maximum of 0.2 around 500 hPa. The 40-member ensemble overestimates the MAC by 0.1 independently of the height. The 200-member ensemble only slightly overestimates the MAC. As for temperature, the vertical extent of the area of increased correlation is approximately 200 hPa. Nevertheless, the relative error is larger for cross-correlations as the 1000-member correlation is much weaker (Fig. 8d). Using 200 members, correlations for distances larger than about 150 hPa are not trustworthy. For the 40-member ensemble, the error strongly exceeds the 1000-member correlation at nearly all levels except for a narrow band around the response level. Thus a strong localization would be required to reduce sampling errors.

Overall, 200 members appear sufficient to estimate vertical correlations of temperature, while 40 members require a narrow vertical localization of less than 200 hPa vertical extent. In general, estimating vertical cross-correlations is more demanding than estimating horizontal correlations. Especially, spurious correlations in combination with weak correlations are an issue as large relative errors emphasize the need for localization for both investigated ensemble sizes. Principally, localizing after a certain distance is potentially dangerous in case of long-range correlations. For instance, clouds and hydrometeors are correlated with temperatures near the surface due to radiative processes (not shown). Similarly, satellite observations often provide integrated information on the entire vertical profile. A possible solution for this issue are statistical sampling error correction (SEC) approaches that aim to correct for spurious correlations without damping correlations after a certain distance (Anderson, 2012, 2016). Further investigation of this approach is provided in Necker et al. (2019).



(a) T 500 hPa to T (b) T 500 hPa to T (c) T 500 hPa to QV (d) T 500 hPa to QV

FIGURE 8 Mean absolute correlation (MAC) and error (MAE) as function of vertical distance [hPa] for differently sized ensembles. Correlations are calculated from 500 hPa response level (horizontal dotted line) to all other levels. Correlations of 500 hPa temperature to (a) temperature and to (c) specific humidity. Corresponding MAE for temperature (b) and specific humidity (d). Note: The black solid line in (b,d) displays the MAC_{1000} as shown in (a,c). Vertical correlations are evaluated for the 1-h forecast initialized on 30 May 2016, 12 UTC.

3.3 | Spatiotemporal correlations

3.3.1 | Example of correlation fields

Ensemble sensitivity analysis is used to compute spatiotemporal correlations for the 1000-member ensemble as well as two random subsets. Ensemble subsets are generated identically as for spatial correlations (Sec. 3.2), and we focus on short-range forecasts with a lead time of 3-h. The response function is fixed at 4-h lead time, and the 1-h forecast is used as the initial state. Figure 9a shows the 1000-member ensemble mean precipitation forecast at 29 May 2016 4 UTC including streamlines of 500 hPa wind. The small black box marks the position of the response function (precipitation spatially averaged over 40×40 grid points) that is used to calculate the spatiotemporal correlations for the investigation time sensitivity of this precipitation system.

The precipitation forecast and initial sea-level pressure field are negatively correlated (Fig. 9b). This means that lower pressure coincides with stronger precipitation in the ensemble. A small-scale structure with correlation values near embedded slightly south of the response function within the relatively smooth large-scale correlation field. This small-scale structure roughly matches the position of the precipitating system at the beginning of the forecast and likely corresponds to surface cooling due to evaporating precipitation. The correlation field of initial 500 hPa zonal wind (Fig. 9c) exhibits a dipole structure. In this case, the dipole seems to indicate stronger cyclonic shear in the south of the box.

Figure 9d shows the spatiotemporal correlation of precipitation inside the response function to earlier precipitation. Precipitation is positively correlated with itself as initially stronger precipitation correlates with increased precipitation three hours later. A similar correlation signal can be observed for hydrometeors (Fig. 9e). Here, hydrometeors are composed of specific cloud water, rain, ice, snow, and graupel content. For hydrometeors, the region of maximum correlation is slightly shifted northwards compared to the precipitation. This could originate from accumulation of the precipitation over one forecast hour or the fact that hydrometeors appear before precipitation is observed at the ground. Furthermore, both precipitation and hydrometeors exhibit a weak positive correlation signal over south-west Germany, which is caused by precipitation in this region in some of the 1000 members.

Similar to sea-level pressure, the upper-air temperature (Fig. 10a) reveals a rather smooth and large-scale correlation

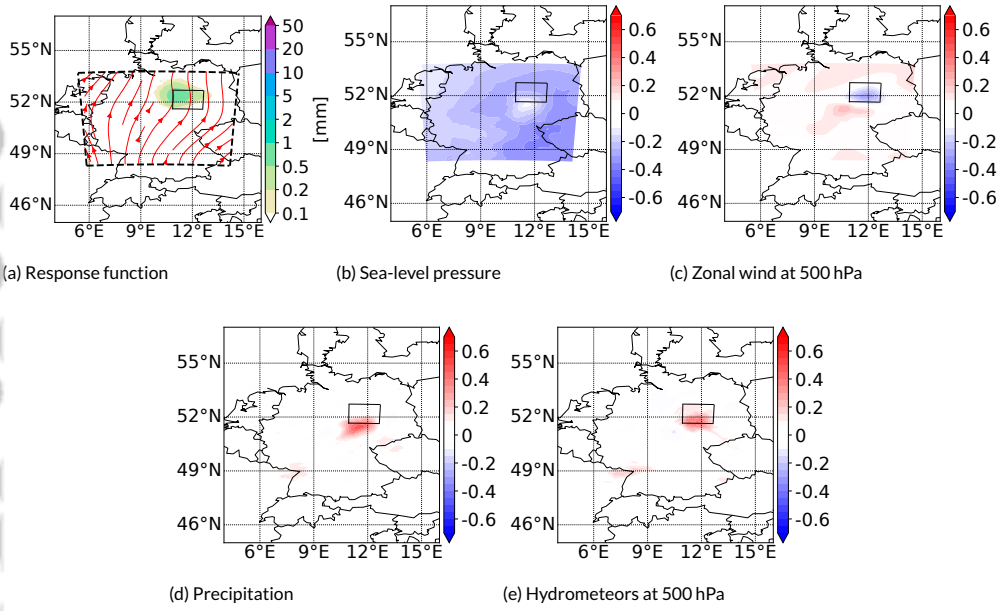


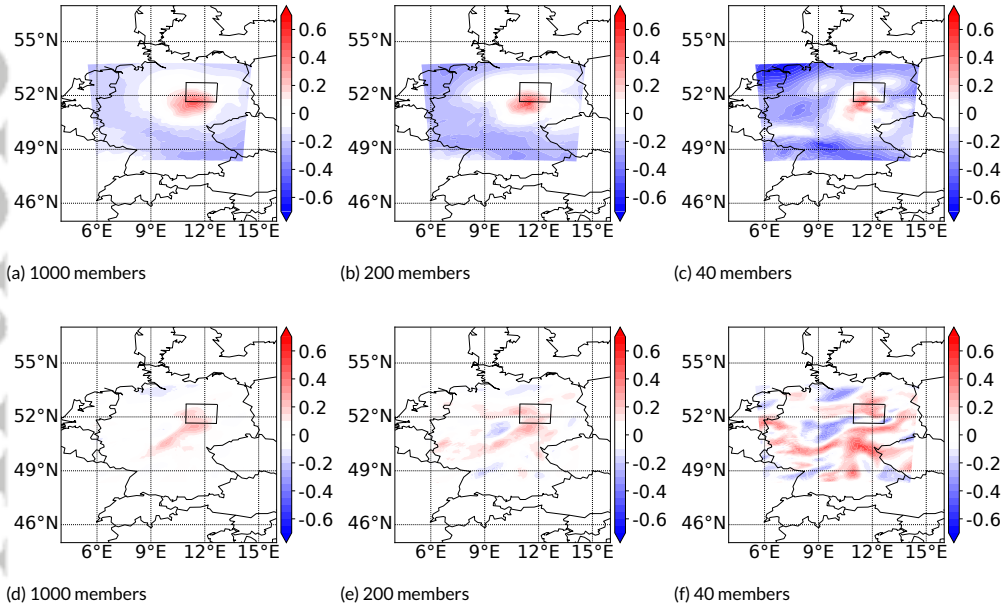
FIGURE 9 (a) Same as Fig. 3c but 1000-member ensemble mean precipitation forecast including streamlines of 500 hPa wind (red solid) in the ESA domain (black dashed), 29 May 2016 04 UTC. The small black box indicates the precipitation region that is used for the sensitivity analysis. (b-e) Sensitivity of the 3 h precipitation forecast inside the box to different initial model fields using the 1000-member ensemble, 29 May 2016 01 UTC.

pattern with negative values, but positive correlation values in the vicinity of the precipitating system that are likely related to the release of latent heat in the precipitating system. The correlation of the specific humidity at 850 hPa (Fig. 10d) is weaker and only extends over a smaller area compared to temperature. The elongated tail roughly marks the track of the precipitating system during the night indicated by the streamlines in Fig. 9a. It seems that the humidity signal reflects precipitation that took place already before analysis time. Interestingly, this feature is not visible in the sensitivity to hydrometeors (Fig. 9e). This may by related to the shorter presence of hydrometeors compared to their longer lasting effect on humidity. The maximum correlation is located in the same region as for temperature, showing a positive correlation of specific humidity and precipitation intensity. Overall, the correlations obtained from the 1000-member ensemble depict physical processes that contribute to the evolution of the precipitating system. If the initial conditions are warmer, more humid and exhibit a higher amount of hydrometers at 500 hPa, the resulting precipitation is more intense. The same applies to initially lower pressure and stronger precipitation.

3.3.2 | Properties of sampling errors

In this section, we present examples of sampling errors for spatiotemporal correlations of precipitation to two representative variables (500 hPa temperature and 850 hPa specific humidity) using two ensemble subsets (200 and 40 members). The 500 hPa temperature correlation pattern is exemplary for other variables with large-scale correlation patterns (e.g., pressure), whereas 850 hPa specific humidity is representative for variables that exhibit small-scale structures in the correlation field (e.g., surface quantities or hydrometeors and precipitation). The 1000-member

ensemble again serves as a reference.



(a) 1000 members (b) 200 members (c) 40 members

(d) 1000 members (e) 200 members (f) 40 members

FIGURE 10 Same as Fig. 9, but for the sensitivity of the 3 h precipitation forecast to temperature at 500 hPa (top row) and specific humidity at 850 hPa (bottom row) for different ensemble sizes, 29 May 2016 01 UTC.

Reducing the ensemble size from 1000 to 200 members only leads to moderate changes for the correlation to 500 hPa temperature (Fig. 10b). The region of positive correlation still looks fairly similar regarding position and magnitude, but negative correlations farther away are systematically larger in magnitude due to spurious oscillations. Differences moving to a 40-member ensemble are substantially larger (Fig. 10c). The local positive correlation pattern lost its shape and negative correlations farther away intensified even further due to sampling errors. Nevertheless, the 40-member ensemble still provides qualitative information as it captures the overall structure and sign of the correlation field for 500 hPa temperature.

Figure 10e shows the sensitivity of precipitation to specific humidity at 850 hPa using a 200-member ensemble. Similar to temperature, weak spurious correlations appear in large parts of the domain, but the region of maximum correlation as well as its elongated tail are well-captured. Lowering the ensemble size to 40 members (Fig. 10f) substantially increases sampling errors and the correlation field is now dominated by spurious correlations.

These results suggest that the 40-member ensemble can provide qualitative information for large-scale patterns, but struggles to estimate correlations for more variable fields as for example 850 hPa humidity or hydrometeors. The 200-member ensemble provides reasonable correlation patterns for all variables, but the fields are still affected by spurious correlations and caution is necessary when using correlations in a quantitative sense as in the following section.

3.4 | Estimating potential impact

This section discusses an approach for investigating the relative potential of observable quantities for improving precipitation forecasts. Again, we focus on spatiotemporal correlations of precipitation obtained for 3-h lead time forecasts. As introduced in Section 2.3.2, we use the accumulated squared correlation (ASC) as a proxy for the potential impact. This means that we accumulate the squared correlation of the precipitation forecast with an initial condition variable over the whole domain and all available forecasts to estimate the relative importance of that variable for data assimilation.

Figure 11 shows the time-averaged ASC as a function of ensemble size between the 3-h precipitation forecast and zonal wind at 500 hPa (Fig. 11a), 2-m temperature (Fig. 11b) and precipitation (Fig. 11c), respectively, at the initial time. For all variables, the ASC using small ensembles is strongly overestimated due to spurious correlations. For instance, the ASC_{1000} is overestimated by more than 200 % using a 40-member ensemble. Generally, the ASC strongly decreases with increased ensemble size, but hardly changes from 600 to 1000 members. This saturation for large samples indicates that spatiotemporal correlations obtained with a 1000-member ensemble are presumably reliable estimates. However, it should also be noted that the larger subsamples overlap with the 1000-member ensemble to a larger degree.

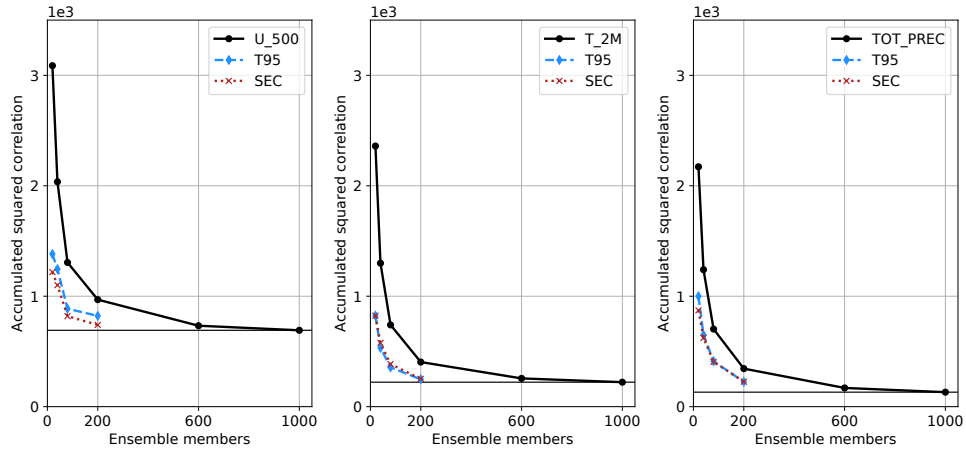
Furthermore, we tested two different approaches to mitigate sampling errors. The first is a confidence test (T95) that excludes insignificant correlations and has been used in several previous ESA studies (Torn and Hakim, 2008). The second is a statistical sampling error correction (SEC; Anderson (2012)) that is based on a Monte-Carlo approach. Necker et al. (2019) shows that the SEC substantially reduces sampling errors for spatial correlations for DA as well as for spatiotemporal correlations as they are calculated within ESA. Compared to the T95, the SEC does not fully exclude small correlations, but systematically reduces the correlation values to account for the overestimation of correlations due to spurious correlations. Examining 500 hPa zonal wind (Fig. 11a), both approaches substantially improve the ASC estimate for small samples. The SEC performs slightly better compared to the T95 and results for the sampling error corrected 200-member ensemble are already close to the ASC_{1000}. Improvements are similar for other variables (Fig. 11b and Fig. 11c). Overall, 200 members including SEC seem to be a reasonable choice for estimating the ASC if no 1000-member ensemble is available. However, it should be noted that there is a small remaining error in the estimate and that the relative error differs for different variables as sampling errors tend to be higher for smaller-scale fields. This can lead to a systematic over- or underestimation of the relative potential of the respective variable. For smaller ensembles (e.g., 40 members), this effect is even larger, and it seems questionable if smaller ensembles are applicable for such a quantitative evaluation of correlations.

Figure 12a shows the time-averaged ASC_{1000} for seven different variables using a precipitation response function and the 1000-member ensemble. Before the discussion, it should be noted that the primary purpose of this study is the discussion of the appropriate ensemble size for such an application. The potential of different observable quantities will be analyzed in more detail in a subsequent study that will also separate different scales.

Surface-level pressure (PS) exhibits the largest ASC_{1000}, followed by wind at 500 hPa and 10-m height. Precipitation has a smaller sensitivity to initial perturbations of temperature and humidity. The smallest ASC is found for precipitation. Applying a confidence test to the 1000-member correlations hardly changes the ASC (Fig. 12a). This confirms the reliability of the results obtained for the 1000-member ensemble.

Using 200 members, the ASC is overestimated for all variables (Fig. 12b). The largest differences are visible for wind and precipitation. As found before, both the T95 and the SEC substantially improve the ASC (Fig. 12b and 12c). Again, the SEC performs slightly better than the T95. The results including the SEC are fairly close to the ASC_{1000}, but there are still some small differences as for example an overestimation of the ASC for precipitation.

Using a 40-member ensemble (Fig. 12d), the ASC is strongly overestimated and the ranking changes compared to



(a) 500 hPa zonal wind (b) 2-m temperature (c) Precipitation

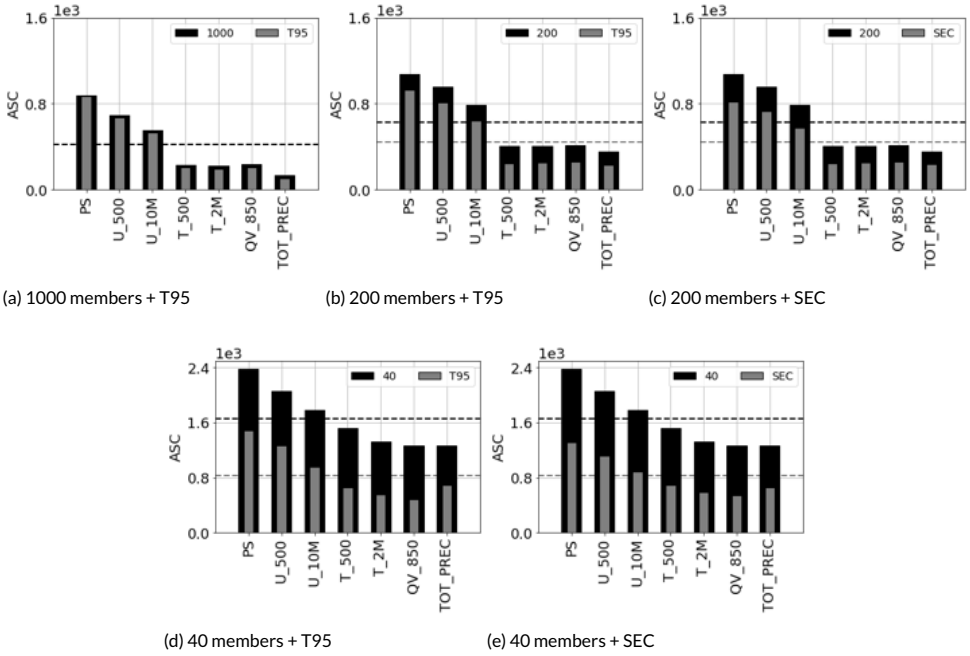
FIGURE 11 Time-averaged accumulated squared correlation (solid, black) as function of ensemble size for different variables, 29 May to 03 June 2016. ASC including confidence test with 95 % confidence level (dashed, blue), sampling error correction (dotted, red) and ASC_{1000} (thin horizontal line, black). Spatiotemporal correlations of precipitation to (a) 500 hPa zonal wind, (b) 2-m temperature and (c) precipitation.

mining the ASC_{1000}, even when the SEC is included. For example, the ASC for precipitation now has an equally large or higher impact as specific humidity or surface temperature. Overall, under-sampling causes an overestimation of the ASC, even if the T95 or the SEC are applied (Fig. 12d and 12e). The 40-member ensemble is therefore not reliable for estimating the ASC or relative ASC of one quantity to another. Nevertheless, it can provide some qualitative guidance.

| CONCLUSIONS

Our study presents the first convective-scale 1000-member ensemble simulation over central Europe and discusses ies and potential applications of this unique data set. The 1000-member ensemble simulation couples two domains through an offline nesting approach and applies the SCALE-LETKF data assimilation system. The data assimilation cycling is performed on a 15 km mesh size and ensemble boundary conditions are obtained from the GEFS system combined with random perturbations. The initial conditions for the 3 km mesh size 1000-member forecasts are obtained by downscaling. The experimental domain is centered over Germany and the five-day period has been chosen because of exceptionally strong summertime convective precipitation.

The simulation is compared to observations as well as to ensemble forecasts of similar resolution computed with the operational COSMO-KENDA system of Deutscher Wetterdienst. COSMO-KENDA forecasts that incorporate high-resolution data assimilation are more skillful than the downscaled SCALE-RM forecasts, but the SCALE-RM 1000-member ensemble overall provides realistic precipitation patterns. Both models reproduce the diurnal cycle of precipitation moderately well in comparison to radar observations. Spin-up effects for precipitation resulting from the downscaling are identified during the first few model integrations. However, these effects are negligible excluding the first forecast hour from the performed analysis. Overall, the 1000-member ensemble exhibits a realistic evolution of the ensemble spread of precipitation and other variables.



(a) 1000 members + T95 (b) 200 members + T95 (c) 200 members + SEC

(d) 40 members + T95 (e) 40 members + SEC

FIGURE 12 Time-averaged ASC using a precipitation response function for all 3-h forecasts, 29 May to 03 June 2016. 1000-member ensemble (a) as well as sub-sampled 200 (b,c) and 40 (d,e) member ensemble including sampling error correction (SEC) or confidence test (T95). Average ASC of all variables (dashed line).

Spatial correlations are calculated using the 1000-member ensemble as well as two ensemble subsets (200 and 40 members). The comparison of these two subsets against the 1000-member ensemble is used to investigate error correlations and the effect of sampling error for convective-scale data assimilation. Horizontally, a 40-member ensemble in most cases strongly overestimates the true absolute correlation due to spurious correlations. Cross-correlations of temperature to other variables as well as near-surface correlations would, therefore, require a narrow horizontal localization of less than 150 km. For all variables, using 200 members largely reduces spurious correlations, and the mean absolute error is more than halved. Consequently, a much broader localization can be applied. Overall, the results suggest that different model variables and combinations of variables require very different localization scales. This emphasizes the advantage of different localization for different variables. Such a variable-dependent localization, however, is not straightforward to implement in some variations of ensemble DA such as the LETKF for example, where localization is applied in observation space. Furthermore, the results show a big advantage of using 200 members compared to 40 members.

Vertically, temperature correlations strongly decrease to a distance of about 200 hPa. Considering cross-correlations of temperature to other variables, the 40-member ensemble struggles to produce reliable estimates of error covariances, and consequently, a strong vertical localization would be required. Using 200 members substantially improves estimated error covariances for convective-scale DA. However, damping correlations after a certain distance appears potentially dangerous considering that pressure, clouds or satellite observations contain integrated information or exhibit long-range correlations. Mitigation of such issues could be achieved using a statistical sampling error correction as suggested

by Anderson (2012). Necker et al. (2019) evaluate this look-up table based sampling error correction (SEC) using the presented 1000-member ensemble simulation.

Spatiotemporal correlations of precipitation with different initial condition variables are discussed for a nocturnal precipitation event on 29 May 2016. Such spatiotemporal correlations are the basis for ensemble sensitivity analysis (ESA) that is often used to investigate atmospheric dynamics on various scales. The example shows that ESA using a 1000-member ensemble is able to return realistic spatiotemporal correlations with respect to precipitation. The 1000-member ensemble can highlight small-scale features that are traceable in space and time. Sensitivity studies on the ensemble size suggest that a 40-member ensemble can provide some qualitative guidance for large-scale patterns. However, about 200 members are required to detect small-scale structures reliably.

The accumulated squared correlation (ASC) is presented as an approach to investigate the relative potential of observable quantities for data assimilation. Sensitivity studies on ensemble size indicate that a 1000-member ensemble returns reliable estimates of the ASC. A 200-member ensemble can provide fairly reliable estimates of the ASC if a confidence test or sampling error correction is included. However, some differences to the 1000-member ensemble still occurred for highly variable fields for example precipitation. Smaller ensembles are not able to estimate the correct amplitude of the ASC but were able to distinguish variables with considerable differences of the ASC. Overall, this study aims to provide the basis for subsequent research on observing and data assimilation strategies for convective-scale NWP. Further investigation is particularly required to separate different scales in this context. Such a scale-analysis and the investigation of vertical localization for satellite data assimilation, is currently ongoing in concurrent projects that build upon the simulation presented in this study.

ACKNOWLEDGEMENTS

The authors wish to thank the RIKEN DA group for their support with the RIKEN/K-computer system as well as Yvonne Ruckstuhl, Leonhard Scheck and George Craig at LMU for helpful discussions. We are also grateful to the reviewers for their suggestions, which helped to improve the manuscript. The open source project and Python package 'xarray' (Hoyer and Hamman, 2017) has been used to post-process ensemble data. Furthermore, we appreciate that Greg Hakim and Julia Keller provided their code for ensemble sensitivity analysis. This study was carried out in the Hans-Ertel Centre for Weather Research (Weissmann et al. (2014); Simmer et al. (2016)), which is a German research network of universities, research institutes and DWD funded by the BMVI (Federal Ministry of Transport and Digital Infrastructure). Furthermore, we acknowledge support from the Transregional Collaborative Research Center SFB/TRR 165 Waves to Weather funded by the German Science Foundation (DFG). This study used computational resources of the K computer provided by the RIKEN Center for Computational Science through the HPCI System Research project (Project ID:ra000015, ra001011).

REFERENCES

Ancell, B. and Hakim, G. J. (2007) Comparing Adjoint- and Ensemble-Sensitivity Analysis with Applications to Observation Targeting. *Monthly Weather Review*, **135**, 4117–4134.

Anderson, J., Hoar, T., Raeder, K., Liu, H., Collins, N., Torn, R. and Avellano, A. (2009) The Data Assimilation Research Testbed: A Community Facility. *Bulletin of the American Meteorological Society*, **90**, 1283–1296.

Anderson, J. L. (2012) Localization and Sampling Error Correction in Ensemble Kalman Filter Data Assimilation. *Monthly Weather Review*, **140**, 2359–2371.

— (2016) Reducing Correlation Sampling Error in Ensemble Kalman Filter Data Assimilation. *Monthly Weather Review*, **144**, 913–925.

Bachmann, K., Keil, C., Craig, G. C., Weissmann, M. and Welzbacher, C. A. (2019) Predictability of Deep Convection in Idealized and Operational Forecasts: Effects of Radar Data Assimilation, Orography and Synoptic Weather Regime. *Monthly Weather Review*.

Baldauf, M., Seifert, A., Förstner, J., Majewski, D., Raschendorfer, M. and Reinhardt, T. (2011) Operational convective-scale numerical weather prediction with the COSMO model: description and sensitivities. *Mon. Weather Rev.*, **139**, 3887–3905.

Bannister, R. N. (2017) A review of operational methods of variational and ensemble-variational data assimilation. *Quarterly Journal of the Royal Meteorological Society*, **143**, 607–633.

Bannister, R. N., Migliorini, S., Rudd, A. C. and Baker, L. H. (2017) Methods of investigating forecast error sensitivity to ensemble size in a limited-area convection-permitting ensemble. *Geoscientific Model Development Discussions*, **2017**, 1–38.

Barrett, A. I., Gray, S. L., Kirshbaum, D. J., Roberts, N. M., Schultz, D. M. and Fairman, J. G. (2015) Synoptic versus orographic control on stationary convective banding. *Quarterly Journal of the Royal Meteorological Society*, **141**, 1101–1113.

Bauer, P., Thorpe, A. and Brunet, G. (2015) The quiet revolution of numerical weather prediction. *Nature*, **525**, 47.

Baur, F., Keil, C. and Craig, G. C. (2018) Soil moisture–precipitation coupling over Central Europe: Interactions between surface anomalies at different scales and the dynamical implication. *Quarterly Journal of the Royal Meteorological Society*, **144**, 2863–2875.

Bednarczyk, C. N. and Ancell, B. C. (2015) Ensemble Sensitivity Analysis Applied to a Southern Plains Convective Event. *Monthly Weather Review*, **143**, 230–249.

Beljaars, A. C. M. and Holtslag, A. A. M. (1991) Flux Parameterization over Land Surfaces for Atmospheric Models. *Journal of Applied Meteorology*, **30**, 327–341.

Bonavita, M., Hólm, E., Isaksen, L. and Fisher, M. (2016) The evolution of the ECMWF hybrid data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, **142**, 287–303.

Caron, J.-F. and Buehner, M. (2018) Scale-Dependent Background Error Covariance Localization: Evaluation in a Global Deterministic Weather Forecasting System. *Monthly Weather Review*, **146**, 1367–1381.

Gaspari, G. and Cohn, S. E. (1999) Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, **125**, 723–757.

Greybush, S. J., Kalnay, E., Miyoshi, T., Ide, K. and Hunt, B. R. (2011) Balance and Ensemble Kalman Filter Localization Techniques. *Monthly Weather Review*, **139**, 511–522.

Gustafsson, N., Janjić, T., Schraff, C., Leuenberger, D., Weissman, M., Reich, H., Brousseau, P., Montmerle, T., Wattrelot, E., Bučánek, A., Mile, M., Hamdi, R., Lindskog, M., Barkmeijer, J., Dahlbom, M., Macpherson, B., Ballard, S., Inverarity, G., Carley, J., Alexander, C., Dowell, D., Liu, S., Ikuta, Y. and Fujita, T. (2018) Survey of data assimilation methods for convective-scale numerical weather prediction at operational centres. *Quarterly Journal of the Royal Meteorological Society*, **144**, 1218–1256.

Hakim, G. J. and Torn, R. D. (2008) Ensemble Synoptic Analysis. *Meteorological Monographs*, **33**, 147–162.

Hamill, T., Whitaker, J. and Snyder, C. (2001) Distance-Dependent Filtering of Background Error Covariance Estimates in an Ensemble Kalman Filter. *Monthly Weather Review*, **129**, 2776–2790.

Hanley, K. E., Kirshbaum, D. J., Roberts, N. M. and Leoncini, G. (2013) Sensitivities of a Squall Line over Central Europe in a Convective-Scale Ensemble. *Monthly Weather Review*, **141**, 112–133.

Harnisch, F., Weissmann, M. and Perianez, A. (2016) Error model for the assimilation of cloud-affected infrared satellite observations in an ensemble data assimilation system. *Q. J. R. Meteorol. Soc.*, **142**, 1797–1808.

Hill, A. J., Weiss, C. C. and Ancell, B. C. (2016) Ensemble Sensitivity Analysis for Mesoscale Forecasts of Dryline Convection Initiation. *Monthly Weather Review*, **144**, 4161–4182.

H da, T., Miyoshi, T., Lien, G.-Y., Nishizawa, S., Yoshida, R., Adachi, S. A., Terasaki, K., Okamoto, K., Tomita, H. and Bessho, K. (2018) Assimilating All-Sky Himawari-8 Satellite Infrared Radiances: A Case of Typhoon Soudelor (2015). *Monthly Weather Review*, **146**, 213–229.

Ho tekamer, P. L. and Zhang, F. (2016) Review of the Ensemble Kalman Filter for Atmospheric Data Assimilation. *Monthly Weather Review*, **144**, 4489–4532.

er, S. and Hamman, J. J. (2017) xarray: N-D labeled Arrays and Datasets in Python. *Journal of Open Research Software*, **5**, 1–6.

Hunt, B. R., Kostelich, E. J. and Szunyogh, I. (2007) Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D: Nonlinear Phenomena*, **230**, 112–126.

Jacques, D. and Zawadzki, I. (2015) The Impacts of Representing the Correlation of Errors in Radar Data Assimilation. Part II: Model Output as Background Estimates. *Monthly Weather Review*, **143**, 2637–2656.

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*.

Kondo, K. and Miyoshi, T. (2016) Impact of Removing Covariance Localization in an Ensemble Kalman Filter: Experiments with 10 240 Members Using an Intermediate AGCM. *Monthly Weather Review*, **144**, 4849–4865.

hnlein, 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.

, L., Whitaker, J. S. and Bishop, C. (2018) Improving Assimilation of Radiance Observations by Implementing Model Space Localization in an Ensemble Kalman Filter. *Journal of Advances in Modeling Earth Systems*, **10**, 3221–3232.

Lien, G.-Y., Miyoshi, T., Nishizawa, S., Yoshida, R., Yashiro, H., Adachi, S. A., Yamaura, T. and Tomita, H. (2017) The Near-Real-Time SCALE-LETKF System: A Case of the September 2015 Kanto-Tohoku Heavy Rainfall. *SOLA*, **13**, 1–6.

G. L. and Houston, A. L. (2018) Ensemble Sensitivity Analysis for Targeted Observations of Supercell Thunderstorms. *Monthly Weather Review*, **146**, 1705–1721.

J., Zhu, Y., Hou, D., Zhou, X. and Peña, M. (2014) Ensemble Transform with 3D Rescaling Initialization Method. *Monthly Weather Review*, **142**, 4053–4073.

oshi, T., Kondo, K. and Imamura, T. (2014) The 10,240-member ensemble Kalman filtering with an intermediate AGCM. *Geophysical Research Letters*, **41**, 5264–5271.

oshi, T., Kondo, K. and Terasaki, K. (2015) Big Ensemble Data Assimilation in Numerical Weather Prediction. *Computer*, **48**, 15–21.

iyoshi, T., Kunii, M., Ruiz, J., Lien, G.-Y., Satoh, S., Ushio, T., Bessho, K., Seko, H., Tomita, H. and Ishikawa, Y. (2016a) Big Data Assimilation Revolutionizing Severe Weather Prediction. *Bulletin of the American Meteorological Society*, **97**, 1347–1354.

Miyoshi, T., Lien, G., Satoh, S., Ushio, T., Bessho, K., Tomita, H., Nishizawa, S., Yoshida, R., Adachi, S. A., Liao, J., Gerofi, B., Ishikawa, Y., Kunii, M., Ruiz, J., Maejima, Y., Otsuka, S., Otsuka, M., Okamoto, K. and Seko, H. (2016b) "Big Data Assimilation" Toward Post-Petascale Severe Weather Prediction: An Overview and Progress. *Proceedings of the IEEE*, **104**, 2155–2179.

Nakanishi, M. and Niino, H. (2004) An Improved Mellor–Yamada Level-3 Model with Condensation Physics: Its Design and Verification. *Boundary-Layer Meteorology*, **112**, 1–31.

Necker, T., Weissmann, M., Ruckstuhl, Y., Anderson, J. and Miyoshi, T. (2019) Sampling error correction evaluated using a convective-scale 1000-member ensemble. *Monthly Weather Review*.

Necker, T., Weissmann, M. and Sommer, M. (2018) The importance of appropriate verification metrics for the assessment of observation impact in a convection-permitting modelling system. *Quarterly Journal of the Royal Meteorological Society*, **144**, 1667–1680.

Nishizawa, S. and Kitamura, Y. (2018) A Surface Flux Scheme Based on the Monin-Obukhov Similarity for Finite Volume Models. *Journal of Advances in Modeling Earth Systems*, **10**, 3159–3175.

Nishizawa, S., Yashiro, H., Sato, Y., Miyamoto, Y. and Tomita, H. (2015) Influence of grid aspect ratio on planetary boundary layer turbulence in large-eddy simulations. *Geoscientific Model Development*, **8**, 3393–3419.

Piper, D., Kunz, M., Ehmele, F., Mohr, S., Mühr, B., Kron, A. and Daniell, J. (2016) Exceptional sequence of severe thunderstorms and related flash floods in May and June 2016 in Germany - Part 1: Meteorological background. *Natural Hazards and Earth System Sciences*, **16**, 2835–2850.

Rasp, S., Selz, T. and Craig, G. C. (2018) Variability and Clustering of Midlatitude Summertime Convection: Testing the Craig and Cohen Theory in a Convection-Permitting Ensemble with Stochastic Boundary Layer Perturbations. *Journal of the Atmospheric Sciences*, **75**, 691–706.

Saha, S., Moorthi, S., Pan, H.-L., Wu, X., Wang, J., Nadiga, S., Tripp, P., Kistler, R., Woollen, J., Behringer, D., Liu, H., Stokes, D., Grumbine, R., Gayno, G., Wang, J., Hou, Y.-T., Chuang, H.-y., Juang, H.-M. H., Sela, J., Iredell, M., Treadon, R., Kleist, D., Van Delst, P., Keyser, D., Derber, J., Ek, M., Meng, J., Wei, H., Yang, R., Lord, S., van den Dool, H., Kumar, A., Wang, W., Long, C., Chelliah, M., Xue, Y., Huang, B., Schemm, J.-K., Ebisuzaki, W., Lin, R., Xie, P., Chen, M., Zhou, S., Higgins, W., Zou, C.-Z., Liu, Q., Chen, Y., Han, Y., Cucurull, L., Reynolds, R. W., Rutledge, G. and Goldberg, M. (2010) The NCEP climate forecast system reanalysis. *Bulletin of the American Meteorological Society*, **91**, 1015–1058.

Satoh, Y., Nishizawa, S., Yashiro, H., Miyamoto, Y., Kajikawa, Y. and Tomita, H. (2015) Impacts of cloud microphysics on trade wind cumulus: which cloud microphysics processes contribute to the diversity in a large eddy simulation? *Progress in Earth and Planetary Science*, **2**, 23.

Sawada, Y., Okamoto, K., Kunii, M. and Miyoshi, T. (2019) Assimilating every-10-minute himawari-8 infrared radiances to improve convective predictability. *Journal of Geophysical Research: Atmospheres*, **124**, 2546–2561.

Scheck, L., Weissmann, M. and Mayer, B. (2018) Efficient Methods to Account for Cloud-Top Inclination and Cloud Overlap in Synthetic Visible Satellite Images. *J. Atmos. Oceanic Technol.*, **35**, 665–685.

Schraff, C., Reich, H., Rhodin, A., Schomburg, A., Stephan, K., Periáñez, A. and Potthast, R. (2016) Kilometre-scale ensemble data assimilation for the COSMO model (KENDA). *Quarterly Journal of the Royal Meteorological Society*, **142**, 1453–1472.

Sekiguchi, M. and Nakajima, T. (2008) A k-distribution-based radiation code and its computational optimization for an atmospheric general circulation model. *Journal of Quantitative Spectroscopy and Radiative Transfer*, **109**, 2779 – 2793.

Skamarock, C., Adrian, G., Jones, S., Wirth, V., Göber, M., Hohenegger, C., Janjic, T., Keller, J., Ohlwein, C., Seifert, A., Trömel, S., Ulbrich, T., Wapler, K., Weissmann, M., Keller, J., , Masbou, M., Meilinger, S., Riß, N., Schomburg, A., Vormann, A. and Weingärtner, C. (2016) HErZ: The German Hans-Ertel Centre for Weather Research. *Bulletin of the American Meteorological Society*, **97**, 1057–1068.

Tomita, H. (2008) New microphysical schemes with five and six categories by diagnostic generation of cloud ice. *Journal of the Meteorological Society of Japan. Ser. II*, **86A**, 121–142.

Torn, R. D. (2010) Ensemble-Based Sensitivity Analysis Applied to African Easterly Waves. *Weather and Forecasting*, **25**, 61–78.

Torn, R. D. and Hakim, G. J. (2008) Ensemble-Based Sensitivity Analysis. *Monthly Weather Review*, **136**, 663–677.

— (2009) Initial Condition Sensitivity of Western Pacific Extratropical Transitions Determined Using Ensemble-Based Sensitivity Analysis. *Monthly Weather Review*, **137**, 3388–3406.

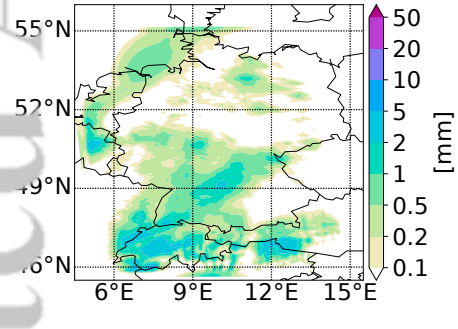
Weissmann, M., Göber, M., Hohenegger, C., Janjic, T., Keller, J., Ohlwein, C., Seifert, A., Trömel, S., Ulbrich, T., Wapler, K., Zollmeyer, C. and Deneke, H. (2014) Initial phase of the hans-ertel centre for weather researchigures – a virtual centre at the interface of basic and applied weather and climate research. *Meteorologische Zeitschrift*, **23**, 193–208.

Whitaker, J. S. and Hamill, T. M. (2012) Evaluating Methods to Account for System Errors in Ensemble Data Assimilation. *Monthly Weather Review*, **140**, 3078–3089.

Wile, S. M., Hacker, J. P. and Chilcoat, K. H. (2015) The Potential Utility of High-Resolution Ensemble Sensitivity Analysis for Observation Placement during Weak Flow in Complex Terrain. *Weather and Forecasting*, **30**, 1521–1536.

Zängl, G., Reinert, D., Rípodas, P. and Baldauf, M. (2015) The ICON (ICOsahedral Non-hydrostatic) modelling framework of DWD and MPI-M: Description of the non-hydrostatic dynamical core. *Q. J. R. Meteorol. Soc.*, **141**, 563–579.

GRAPHICAL ABSTRACT



This study presents a unique convective-scale 1000-member ensemble simulation over central Europe. As a first step, the large ensemble is used to investigate sampling errors and localization in convective-scale data assimilation. Furthermore, we introduce how the relative potential impact of observable quantities can be estimated using spatiotemporal correlations obtained with ensemble sensitivity analysis.