Local Quantitative Precipitation Forecast with Minimal Data Requirement—An Ensemble Approach

RAUL FERNANDO MENDEZ TURRUBIATES, MARKUS GROSS, AND VANESA MAGAR

CICESE, Ensenada, Baja California, Mexico

(Manuscript received 11 April 2019, in final form 18 February 2020)

ABSTRACT

Operational weather forecasts are routinely performed at convection-allowing resolutions, and thus these forecasts generate weather features that appear to be realistic. However, at times the comparison of the forecast to observations is less favorable, particularly at grid scales. This lack of skill is partly due to the chaotic system underlying the weather. Another problem is that it is impossible to evaluate the risk of making decisions based on deterministic forecasts. However, running global high-resolution ensembles involves substantial computational resources. A 555-m resolution WRF ensemble based on stochastic perturbations of a deterministic forecast of the North American Mesoscale model was created. Observations are used to constrain the ensemble and improve the skill. This method increases the skill for forecasting 60-h accumulated precipitation in five standard, statistical metrics: bias, false alarm ratio, threat score, probability of detection, and success ratio. Furthermore, the ensemble continuous ranked probability score (CRPS) will be compared to a poor man's ensemble. The forecast error is generally smaller in more than 70% of the case studies performed when compared to nine deterministic model forecasts. The ensemble-enhanced mesoscale system presented can help to determine the most likely scenario without the significant computational requirement of global ensembles and is expected to be useful when global high-resolution ensembles are not available.

1. Introduction

Mesoscale model forecasts provide remarkable detail and realism in the resolved convective systems, as well as in the temperature and wind distributions. Less dependence on parameterizations of physical processes and more resolved physics lead to a representation of weather features that appear convincing, and their value has been reported in numerous studies (Done et al. 2004; Kain et al. 2006, 2008; Weisman et al. 2008; Schwartz et al. 2009). The resolution of mesoscale models and the associated reduced need for parameterization lead to a more accurate depiction of the physics (Done et al. 2004). When objectively compared to observation, however, it is frequently found that locality, timing, and intensity are not improved (Mass et al. 2002). While high resolution enhances the forecast skill (Yan and Gallus

2016), in general, it does not necessarily provide more accurate point forecasts. With the resolved local features, such as convective cells, small displacement errors can produce differences in, for example, rainfall rate at individual grid points (Schwartz and Sobash 2017).

The lack of objective skill then raises the question, how much confidence can be placed on this forecast? This is a question that a single deterministic forecast cannot answer. For example, during a week in January 2016 severe rainfall was predicted for the region of Ensenada, Mexico. Some of the available deterministic models at the time predicted accumulated precipitation of 70 mm or more over 1 or 2 days. The amount of 70 mm is more than the mean monthly accumulated precipitation during the rainy season months. This forecast created anxiety among the vulnerable public. Also, emergency responders were communicating these high values, demonstrating the high level of confidence attributed to these forecasts. The tension was increased in particular because the events occurred during the peak of an extreme El Niño event, following months of alarming speculation about the potential impacts of expected severe precipitation events commonly associated

Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/WAF-D-19-0077.s1>.

Corresponding author: Markus Gross, mgross@cicese.mx

with El Niño in this region (Pavia et al. 2016). Comparing the forecast with available ensemble prediction systems, such as the North American Ensemble Forecasting System (Candille 2009), showed that these 70-mm amounts were likely a severe overestimate of a single deterministic forecast.

For risk assessment and planning, knowledge of the probabilities associated with forecast events is essential (Palmer 2002; Thielen et al. 2009). Attempts can be made to compute probabilities of precipitation by evaluating neighboring grid cells of the deterministic forecast (Theis et al. 2005). Applying the neighborhood approach to a single forecast is preferable to communicating forecast results purely deterministically. Further improvements can be achieved by using an ensemble system. However, the current global ensemble systems are running at ≈ 20 -km grid resolution with an effective resolution of 60–80 km (Abdalla et al. 2013; Klaver et al. 2020). Therefore, while producing some confidence in their forecasts, spatial uncertainty can still lead to sizable local forecast errors. Weidle et al. (2016) downscale the forecast from the European Centre for Medium-Range Weather Forecasts Ensemble Prediction System (ECMWF-EPS) and the Global Ensemble Forecasting System (GEFS) from the National Centers for Environmental Prediction (NCEP). However, the downscaling from the ECMWF-EPS 32 km is a computationally intense job, both from the calculation and data transfer perspective. To reduce the computational work, the system presented in this article starts using the highest-resolution mesoscale forecast available, perturbs the solution to create the initial condition for the ensemble, and evaluates the ensemble output.

Even if the number of ensemble members considered in the forecast is small, the high spatial resolution is expected to improve the forecast in the short range. Experimental short-range ensemble forecasts started at NCEP in 1995 (Du and Tracton 2001). Convection-permitting ensembles with a resolution of 1–4 km are now frequently used in operational centers (Gowan et al. 2018a), in particular for precipitation forecasts (Clark et al. 2018). Experiments with high-resolution convection-allowing modeling systems are performed by some national forecasting centers. For example, during the 2010 Spring Forecasting Experiment, a 1-km forecast was generated by the University of Oklahoma (Clark et al. 2012). This experiment required 10 000 computing cores. Not many centers can dedicate resources of that magnitude to a single project, not to mention the need for increased resolution for local impacts and performing ensemble runs. The cost becomes a limiting factor because it increases exponentially with resolution and nearly dominates discussions of optimum configurations and resolutions (cf. Fosser et al. 2019).

Hence local, less computationally intensive solutions are highly desirable.

Downscaling a mesoscale analysis (e.g., NAM analysis at 5-km resolution) to a higher resolution (555 m) is computationally less expensive than downscaling a coarse-resolution analysis (e.g., GFS) to the same resolution. This cost reduction partially offsets the higher cost of ensembles, which can provide insight into the uncertainty of the deterministic forecasts. It can be reasonably expected that the high-resolution nests spin up the small-scale structure missing in the coarser grids in a few hours (Skamarock 2004). This additional information will improve the representation of the underlying physics (e.g., resolved convection vs parameterized convection) and the atmosphere interacts with a more realistic orography, providing useful forecast data for the remainder of the forecast period. Also, if the generating mesoscale analysis was based on a deficient forecast of the actual condition, then the perturbations provide variation in the initial data such that the ensemble has a chance to recover from a suboptimal initial state. Such suboptimal initial states may, for example, occur in regions with few, or low-quality, observations.

Here, the authors use random perturbations applied to a deterministic analysis to generate initial conditions for a 555 m, limited-area ensemble forecast over Ensenada, Baja California, Mexico. The forecast for a single location is derived from ensemble subsets based on the performance during the first 24 h. Constraining the ensemble members in this way can increase the skill of the forecast (see, e.g., Dong and Zhang 2016; Ancell 2016). Grid points adjacent to the specific location of interest are also included in the analysis in what will be termed a geoensemble. This “ensemble-enhanced mesoscale forecast” can be desirable in at least three scenarios. First, there may be insufficient computational resources to run the required global ensemble or downscale each member of the global ensemble to the required local resolution. Second, the system could be useful when it is of the essence to provide a forecast in a timely fashion, for example, in emergency response situations or forecaster-requested on-demand modeling support for the weather desk, similar to the on-call fire nest of the NAM (Black et al. 2009). Finally, the proposed scheme may be attractive in the case of remote embedded forecast units [such as Weather Running Estimate–Nowcast (WRE-N); Perconti 2017; Robert et al. 2014], which may rely on satellite transmission of the initial conditions.

2. Study area

The case studies presented here have been performed for the Port of Ensenada, Baja California, Mexico.

TABLE 1. Model setup for the three domains d01, d02, and d03.

	d01	d02	d03
Resolution	5000 m	1666.66 m	555 m
Time step	20 s	20/3 s	20/9 s
Microphysics	WSM 3-class simple ice	Same as d01	Same as d01
Longwave radiation	RRTM	Same as d01	Same as d01
Shortwave radiation	Dudhia	Same as d01	Same as d01
Time between radiation	1 min	Same as d01	Same as d01
Surface layer	MM5 Monin–Obukhov	Same as d01	Same as d01
Land surface	Thermal diffusion	Same as d01	Same as d01
Boundary layer	YSU	Same as d01	Same as d01
Cumulus	Kain–Fritsch (new Eta)	Same as d01	No cumulus
Nonhydrostatic	.true.	.true.	.true.
dfi_opt	Twice digital filtering initialization (TDFI)		

The site is interesting due to the importance of precipitation forecasts in this area. Precipitation is usually sparse and, when it does occur, events are intense. The average temperature is 16.5°C, with maxima in August and minima in January and December. Average annual rainfall is 281 mm, with a maximum of around 60 mm in March and nearly no precipitation in June and July. When precipitation occurs, it is focused on only a few days, followed by dry periods of one to several weeks. The Köppen–Geiger climate classification (Köppen and Geiger 1936) for Ensenada is BSk, a cold semiarid climate (Kottke et al. 2006; Rubel and Kottke 2010). Preparation and planning are essential both for agriculture and, during extreme events, for emergency services. Moreover, the combination of a coastal setting with mountainous and desert regions nearby challenges the forecast models.

3. Methodology

The initial single deterministic mesoscale forecast will be perturbed with random noise, to create an ensemble of individual forecast runs. During these runs, the forecast is dynamically downscaled to increase the resolution. The performance of this ensemble is analyzed and a subensemble formed by selecting ensemble members. The performance of this subensemble is then evaluated by

- comparing the forecast with alternative deterministic models at varying resolution,
- evaluating the continuous rank probability score (CRPS) of the ensemble and the alternative poor man's ensemble (Ebert 2001) of other available deterministic models, and
- defining forecast success so that the ensemble can then be further evaluated using model bias, success ratio, probability of detection, threat score.

Details of this methodology are presented in turn next.

a. Initial and lateral boundary conditions

The initial data and boundary conditions were taken from the North American Mesoscale (NAM) model CONUS Nest (Tolman 2014), with 5-km grid spacing, as run by the U.S. National Weather Service NCEP (WMC) at 0000 UTC. A single deterministic forecast of the NAM model is perturbed, to generate initial conditions for a 20-member ensemble, by adding random noise to the analysis in control variable space.

b. Model configuration

These initial conditions are then dynamically downscaled to increase the resolution. The Weather Research and Forecasting (WRF) Model (Skamarock et al. 2008) version 3.8.1 was used for the downscaling. Each forecast was run as a cold start with two nests in two-way nesting mode. The model configuration is listed in Table 1. All domains, labeled d01, d02, and d03, had 241×241 grid points each, and a grid spacing of 5 km, 1.666 km, and 555 m, respectively; they all had 30 vertical levels. On the d03 domain, corresponding to the highest-resolution domain in the computations of the ensemble members, no cumulus parameterization was used. The grid is centered over Ensenada, Baja California, Mexico, with a reference latitude of 31.9394°N and reference longitude of 116.5863°W.

c. Ensemble generation

The ensemble consists of 20 individual model runs. These model runs are the ensemble members. The initial conditions and lateral boundary conditions of each ensemble member of streamfunction, potential velocity, temperature, surface pressure, and pseudo-relative humidity are perturbed in physical space using spatially correlated noise and the NCEP background error covariance matrix (Parrish and Derber 1992). This noise is

generated using a random number generator, which was seeded from the computer's pseudo-random number generators. The pseudo-random number generators compute random numbers from the computer's environmental noise, such as device drivers. This seeding is different from the standard initialization in WRF data assimilation (WRFDA), which uses a constant seed. The modified routine of WRFDA is available in the online supplemental material.

The ensemble can be extended by treating the neighboring grid points as if they were independent solutions for the grid point of interest. This extension appreciates that significant uncertainty exists in orography, surface roughness, and the precipitation field itself (Tapiador et al. 2019). This uncertainty implies that a location in the model is not necessarily the best representation of the physical location. A neighboring grid point may be more appropriate. In the present work, this procedure is termed the geoensemble (GE), to distinguish this procedure from other neighborhood approaches (see Schwartz and Sobash (2017) for an excellent review). The neighborhood is defined by a circle with a radius of $r\Delta x$, where Δx is the grid spacing and r a constant factor. GE r , the geoensemble with radius $r\Delta x$, now represents a new ensemble. The individual time series make up the geoensemble and are termed geomembers. This term has been chosen to avoid confusion with the original members of the ensemble (i.e., the independent forecast runs). A circle with two grid spacings radius encloses 13 grid points. Therefore, a GE2, as illustrated in Fig. 1, based on an ensemble of 20 perturbed forecast runs (members) generates a geoensemble with 13×20 geomembers.

d. Downscaling

The downscaling is performed with WRF using the configuration detailed above. The boundary update interval is 1 h for the first 36 h and then 3 h until forecast hour 60. No observations were assimilated into the downscaling runs because the initial condition is provided by a numerical weather prediction model, which already assimilates observations. Digital filtering initialization (Peckham et al. 2016) was used to reduce the initial transients during the spinup of the model. The validity of the downscaling model setup was tested in two ways. First, a forecast with 5-km resolution at forecast hour 24 was compared to the analysis of the subsequent day. The RMS error of this forecast, computed across the whole domain, was compared to the RMS error of the NAM forecast to the subsequent analysis. It was found that the RMS error for wind, temperature, and relative humidity was either comparable or smaller than the error of the NAM forecast to its

FIG. 1. Selection of grid point in the geoensemble. All ensemble members of each grid point on or inside the circle, here 13 for a GE2, are combined into one geoensemble.

subsequent analysis. A second test was performed to evaluate the behavior of the solution in the higher-resolution nests. Here it is expected that the nests generate a solution, which is different from a mere interpolation onto the finer grid in the nest, due to the improved resolution of the physical processes. If the nests only provide an interpolation, the additional effort of running the forecast model at that resolution would not be justified. In general, high-resolution nests always produce “fine structure.” However, in the context of this work that is only guaranteed in two scenarios: the simulation period is long enough, and its parent does not artificially constrain the nest (i.e., the nesting is sufficiently large). The former is constrained by the forecast period, which here is short and not changeable; the latter is often constrained by computational power. Too frequently are nests one or both of the following: too small or too close to the boundaries of its parent. If the nest is too close to the boundary of its parent, the parent is not able to develop its solution at all the scales it theoretically supports. Hence it cannot pass this information on to the even finer nest, turning the nesting into mere interpolation. A series of experiments was carried out without nests, with one nest, and finally with both nests. The final accumulated precipitation patterns were re-gridded to the finest grid, using the conservative remapping (Jones 1999) implemented in the Earth System Modeling Framework (ESMF) provided by the NCAR Command Language (NCL). From these interpolated fields, the Pearson correlation was evaluated, comparing

the two fields on the d03 domain. The direction of the interpolation is critical. If the nest does not add any new information, then its results will be only an interpolation. If the nest generates its fine structure, then interpolating this back to a coarser grid would eliminate most of this. Generally, the correlation between the interpolated field from the coarser grid and the solution at the fine grid dropped from initially 1, due to the same initial data, to a value of around 0.6, demonstrating that the solution in the nested domain is different from merely interpolating from a coarse grid into a fine grid.

In some cases, the difference was less clear. However, this test is weather dependent. If the precipitation pattern is dominated by large-scale features, the solution using nesting will not necessarily differ between the coarse and the fine grid.

e. Forecast performance analysis

For the analysis, the model output of the highest-resolution nest was sampled at the grid point closest to the weather station of the Mexican weather service (SMN), station name: P. LOPEZ ZAMORA, Ensenada, Baja California, Mexico, latitude: 31°53'29"N, longitude: 116°36'12"W, altitude: 32 m. All analysis performed focuses on the accumulated precipitation predicted by the model.

1) DESCRIPTION OF THE GEOMEMBER SELECTION PROCESS

When the geoensemble forecasts are compared to observations, some forecasts are closer to the observations than others. A subensemble can then be created using geomembers whose forecasts are close to the observations in the first 24 h, to improve the forecast skill over the remaining 36 h. This selection procedure is similar to the observation-based ensemble subsetting technique (OBEST) tropical cyclone track prediction (Qi et al. 2013; Dong and Zhang 2016), where the short-term track error was used to select a subensemble for the remaining forecast period. The geomember selection criteria is based on accumulated precipitation over the period from forecast start to time t (p_{acc}) and its standard deviation σ :

$$\sigma(t) = \sqrt{\frac{1}{M} \sum_{m=1}^M [p_{\text{acc},m}(t) - \overline{p_{\text{acc},m}}(t)]^2}, \quad \text{with} \quad \overline{p_{\text{acc},m}}(t) = \frac{1}{M} \sum_{m=1}^M p_{\text{acc},m}(t), \quad (1)$$

where M denotes the total number of geomembers and m the individual geomember. A conformity index CI is then defined as the number of 1-h periods

where the precipitation accumulated over the period predicted by the geomember ($p_{\text{acc},m}$) at time t fulfills the criterion:

$$[p_{\text{acc,obs}}(t) - \gamma\sigma(t)] \leq p_{\text{acc},m}(t) \leq [p_{\text{acc,obs}}(t) + \gamma\sigma(t)], \quad (2)$$

where $p_{\text{acc,obs}}$ is the observed precipitation accumulated over the period and γ is a scaling parameter. This procedure is illustrated using a case study in Fig. 2 for the case of a geoensemble 3 (GE3). In that figure, the shown accumulated precipitation of geomember 80 fulfills the selection criterion seven times, therefore scores a CI of 7. Geomember 200 has a CI of 10 and geomember 250 scores 5. Cases where the criterion is fulfilled are indicated by filled dots. Using the minimum CI ($\text{CI}_{\min}$) as a selection criterion reduces the final subensemble geomember count as illustrated in Fig. 2b) of the same figure. The rationale for this procedure is that if a forecast does not produce a satisfactory precipitation forecast in the first 24 h, there is no physical reason other than chance why the remaining period of the forecast should be better. This statement is of course slightly exaggerated and simplistic, as will be discussed further below, but in the first principle sufficient. Conversely, if the forecast performed well over the first 24 h, then the likelihood that it continues to do so is higher than just a random guess. In the words of Tapiador et al. (2019): it is unlikely to get instantaneous precipitation right for the wrong reasons.

2) SKILL EVALUATION

The mean ensemble continuous rank probability score (Gneiting and Raftery 2007) is computed as

$$\overline{\text{CRPS}} = \frac{1}{I} \sum_{i=1}^I \int_{-\infty}^{\infty} [F_i^f(x) - F_i^o(x)]^2 dx, \quad (3)$$

where I , F^f , and F^o denote number of case studies, forecast cumulative probability density function (cdf), and observed precipitation expressed as a cdf, respectively.

The forecasts are then evaluated using

- Bias: The degree of correspondence between the mean forecast and the mean observation, stating how well the forecast frequency of events compared to the observed frequency of events.
- Probability of detection (POD): The fraction of predicted events that occurred.
- False alarm ratio (FAR): The fraction of the predicted events that did not occur.
- Success ratio (SR): The fraction of the forecast and observed events, and

FIG. 2. Selection of geomembers. Geomembers are selected based on their hourly accumulated precipitation value. With 20 perturbed initial conditions, the GE3 has 29×20 geomembers. (a) If the value falls inside the blue shaded region, its CI increases by one. The blue region extends in time over the first 24 h of the forecast run. In the vertical, it extends γ standard deviations, σ of the full ensemble above and below the observed accumulated precipitation value. Only three geomembers are depicted. The selection area is spanned by observation plus standard deviation and observation minus standard deviation. The number of times the ensemble geomember produces hourly forecasts inside this area defines the CI. (b) The top panel shows how the accumulated precipitation at the end of the 60-h forecast period approaches the observed accumulated precipitation with increasingly strict selection criterion for various values of γ . The bottom half shows how the number of geomembers is reduced due to selection.

- Threat score (TS) or critical success index (CSI): Correspondence of the forecast events to the observed events. The CSI is a biased score that is dependent upon the frequency of the event (Schaefer 1990).

However, before the objective skill of the forecast can be evaluated using bias, POD, SR, TS, and FAR, forecast success and failure need to be defined. The definition of success depends crucially on the target use of the forecast. Here forecast success is defined using a parameter N , and the accumulated precipitation at the end of the forecast period. N is bounded between zero and one, allowing for up to 100% error. If the precipitation accumulated over the forecast period is within

$$(1 - N)p_{acc,obs} \leq \frac{1}{M} \sum_{m=1}^M p_{acc,m} \leq (1 + N)p_{acc,obs} \quad (4)$$

the forecast is counted as a success (hit h). If the accumulated precipitation is below $(1 - N)p_{acc,obs}$, it is counted as miss, otherwise, if it is above $(1 + N)p_{acc,obs}$ it is counted as false alarm (fa). Using this classification the performance of the forecasts can be expressed using

$$\begin{aligned} \text{BIAS} &= \frac{h + fa}{h + miss}, & \text{POD} &= \frac{h}{h + miss}, \\ \text{FAR} &= \frac{fa}{h + fa}, & \text{SR} &= \frac{h}{h + fa} = 1 - \text{FAR}, \quad \text{and} \\ \text{TS} &= \frac{h}{h + miss + fa}. \end{aligned} \quad (5)$$

4. Results

A total of 28 case studies are presented. Each case study includes a 20-member ensemble and a deterministic “rerun” (i.e., a repetition of the NAM forecast using the NAM as initial data and boundary condition without nesting at d01 resolution). The case studies cover the most significant precipitation events of the years 2016 and 2017, as listed in Table 2.

a. Ensemble, selected ensemble, or deterministic forecast?

The geoensemble results are compared against results from six deterministic models, listed in Table 3, and the performance of the deterministic models relative to the geoensemble are presented in Table 4. In this table the percentage of case studies with a forecast error higher than the selected geoensemble mean is listed. The selected geoensemble mean was taken from the $CI_{min} = 13$ geoensemble GE3, with $\gamma = 1$. If that was not available, because there may not have been

TABLE 2. Case study dates.

Days	Month	Year
4, 5	Jan	2016
10	Apr	2016
21, 22, 24, 31	Dec	2016
1, 11, 12, 13, 18, 19, 21, 22, 23, 24	Jan	2017
6, 7, 17, 18, 19, 26, 27, 28	Feb	2017
21, 23	Mar	2017
6	May	2017

13 members within the selection range for a specific date, $CI_{\min} = 5$ was used. If that was not available either,

the geoensemble mean was used as reference. This procedure simulates a real forecasting scenario. If at the time of geoensemble analysis, the desired $CI_{\min}$ cannot be realized, then the forecaster would have three possibilities. The forecaster could request to run more forecast members, but at this point, there is no time for that anymore. The forecaster could not issue a forecast, which usually is not an option, or the requirement on the forecast can be downgraded to the next realizable one. With $CI_{\max}$ denoting the maximum CI supported by the geoensemble for a specific date, the geoensemble forecast error ε_g can be defined as

$$\varepsilon_g = \begin{cases} \left| p_{\text{acc,obs}} - \overline{p_{\text{acc},m}(t=60)}^{M(CI_{\min}=13)} \right|, & \text{if } CI_{\max} \geq 13 \\ \left| p_{\text{acc,obs}} - \overline{p_{\text{acc},m}(t=60)}^{M(CI_{\min}=5)} \right|, & \text{if } 5 \leq CI_{\max} < 13, \\ \left| p_{\text{acc,obs}} - \overline{p_{\text{acc},m}(t=60)}^M \right|, & \text{if } CI_{\max} < 5 \end{cases} \quad (6)$$

where $M(CI_{\min} = n)$ defines the subset of members that are selected for a given n . The percentage of larger forecast errors $P_{>}$ and larger or equal $P_{\geq}$ is then

$$P_{>} = \frac{C - \sum_{c=1}^C \delta \left[\max \left(0, \frac{|p_{\text{acc,obs}} - p_{\text{acc},d}|}{\varepsilon_g} - 1 \right) \right]}{C}, \quad (7)$$

and

$$P_{\geq} = \frac{\sum_{c=1}^C H \left(\frac{|p_{\text{acc,obs}} - p_{\text{acc},d}|}{\varepsilon_g} - 1 \right)}{C}, \quad (8)$$

respectively, where C defines the total number of case studies, δ is the Dirac delta function, H is the Heaviside

step function with $H(0) \equiv 1$, and $p_{\text{acc},d}$ is the accumulated precipitation of the respective deterministic forecast d .

Note that this is not a comparison of forecast means across all case studies, which would reduce the impact of extreme events to the advantage of the ensemble. Indeed, the pure ensemble has a larger error in 50% of the forecasts. Extreme events will favor the deterministic forecast if the deterministic model forecasts them correctly. They will improve the scores of the selected ensemble, if, due to the selection procedure, only the members forecasting the extreme event accurately, were chosen. If instead the precipitation of all case studies were averaged, then this test would bias the results in favor of the ensemble, because the averaging would diminish the impact of extreme events. Table 4 shows that

TABLE 3. Deterministic model forecasts used in the verification. The forecast type indicates if the data are from a single forecast run from one analysis (full forecast) or from sequential forecasts of 12-, 9-, or 1-h forecasts [i.e., a full forecast for a 60-h period, starting at date T0 uses forecast hours (fh) 1–60 from one initial date (id) with id = T0]. The 12-hourly forecast uses fh 12 from 6 different id: fh 12 from id = T0, fh 12 from id = T0 + 12, fh 12 from id = T0 + 24, etc. The 9-hourly forecast type uses 6 initial dates and the 1-h forecast finally uses 59 initial dates.

Model	Resolution	Forecast type
GFS (NOAA/NCEP/NWS 2015a)	0.25° (27.8 km)	Full forecast
CFSv2 (Saha et al. 2011)	0.2° (22.3 km)	9-hourly forecasts
ECMWF IFS	0.14° (15.6 km)	Full forecast (pre-March 2016)
NAM218 (NOAA/NCEP/NWS 2015b)	12.2 km	Full forecast
ECMWF IFS	0.07° (7.8 km)	Full forecast (8 Mar 2016 onward)
Rerun	5 km	Full forecast
NAM	5 km	Full forecast
HRRR12 (Blaylock et al. 2017)	3 km	12-hourly forecasts
HRRR (Blaylock et al. 2017)	3 km	1-h forecasts

TABLE 4. Evaluation of the forecast performance of available deterministic models vs selected geoensemble for a 60-h forecast of accumulated precipitation. The percentage indicates the fraction of case studies where the respective model has larger forecast errors than the selected ensemble, or as in the last column, equal or larger. HRRR12 uses five 12-h forecasts, whereas HRRR uses 60 (total) 1-hourly forecasts, with a significant increase in data assimilation. The deterministic forecasts have been evaluated at the nearest grid point. “Control” denotes the unperturbed ensemble member at the nearest grid point.

Model	Percentage of larger forecast errors, $P_{>}$	Percentage equal or larger error over last 36 h, $P_{\geq}$
NAM 5 km	61.5	76.9
Rerun	73.1	61.5
ECMWF IFS	72.0	53.8
NAM218	57.7	50.0
HRRR12	65.2	47.8
HRRR	47.8	52.2
CFSv2	80.8	56.0
GFS	61.5	50.0
Geoensemble mean	50.0	80.8
Geoensemble selected 13	0.0	100.0
Geoensemble selected 5	36.0	80.0
Control	76.9	65.4

the high-resolution ensemble mean performs better than all but the HRRR. The HRRR, however, has a clear advantage by using hourly data assimilation and accumulation of 1-h forecast periods. However, it is *outperformed by the selected geoensemble in 50% of the cases*. It is not only the high resolution of the geoensemble that is responsible for this improvement because the control is outperformed in nearly 77% of cases, only the CFSv2 performs worse.

The preceding analysis provides insights into the two main research questions associated with this publication:

- Is it necessary to run an ensemble instead of a deterministic forecast? The high-resolution control has larger precipitation errors in nearly 76.9% of cases. This reduces to 50% for the full geoensemble forecast. The only forecast that provides lower errors for more than half of the case studies performed is the 1-hourly HRRR, which benefits from moderately high resolution, and short data assimilation cycles. Furthermore, in the 1-hourly setup it cannot really be called a forecast, as the forecast period is only 1 h. This suggests that the ensemble is certainly useful.
- Is it necessary to perform the selection process? Comparing the forecast errors over the last 36 h in Table 4 shows that the selected ensemble performs as well or better than most of the deterministic forecasts, even over the period of the last 36 h. This period is more challenging because of its longer lead time, and

the fact that the evaluation is based on the variable used for selection, as discussed further below. Based on accumulated precipitation, the selected geoensemble performs equally or better in more than 80% of cases when compared to the geoensemble mean.

The above analysis has been carried out under a “simulated forecast” scenario. Here that is meant to mean that a preferred geoensemble configuration ($CI_{\min}$ and γ) was defined. If that was not fulfilled on any given day, these criteria were lowered to provide a forecast. The question may arise what the impact would be if the requirements were fixed and forecasts not complying with these criteria dropped from the evaluation. The results of this test are shown in Table 5. Using the geoensemble mean produces the lowest-performing forecasts against all models over the 36-h periods tested and against all but one model (NAM) for the 60-h periods. Requiring a $CI_{\min} = 5$ only outperforms the flexible scheme used above against two models, the GFS and NAM. Fixing the requirement to $CI_{\min} = 13$ then improves the forecasts against three models, the NAM, CFSv2, GFS, and control. However, insisting on a certain $CI_{\min}$ means not issuing a forecast when this requirement is not fulfilled, which is undesirable.

This simple analysis has given some insights into the potential usefulness of the system. However, the forecast system has several variables, which have not been considered in the preceding analysis. The selection procedure has two free variables, $CI_{\min}$ and γ . Furthermore, the size of the geoensemble can be varied. The following sections aim at analyzing the forecast performance taking these additional variables into account. The traditional skill scores complicate the analysis because forecast success needs to be defined. This will be addressed later. First the focus is on the CRPS, which does not require a definition of forecast success.

b. The impact of GEr , $CI_{\min}$, and γ

The CRPS analysis has only three free parameters, GEr , which is the geoensemble size, the $CI_{\min}$ required as a minimum for selection, and γ , the definition of the selection criterion. In Fig. 3, the mean CRPS for all case studies is shown. The mean accumulated precipitation of the selected geomembers was compared to the observation of accumulated precipitation at forecast hour 60. It can be seen clearly how the *skill improves with selection*. Selecting a too broad γ eliminates this improvement, requiring more coincidences generally improves the skill, but requiring too many again reduces it; as observed above, a too stringent criterion eliminates too many otherwise good geomembers. The “poor

TABLE 5. Evaluation of the forecast performance of available deterministic models vs selected geoensemble for a 60-h forecast of accumulated precipitation. The percentage indicates the fraction of case studies where the respective model has equal or larger forecast errors than the selected ensemble. The deterministic forecasts have been evaluated at the nearest grid point. “Control” denotes the unperturbed ensemble member at the nearest grid point. The sample size is given for every percentage in parentheses. The values larger than the ones obtained in the forecast simulation are indicated with an asterisk.

Model	$P_{\geq}$ for 60 h forecast			$P_{\geq}$ over last 36 h		
	$CI_{\min} = 13$	$CI_{\min} = 5$	Geoensemble mean	$CI_{\min} = 13$	$CI_{\min} = 5$	Geoensemble mean
NAM	63.6 (22)*	64.0 (25)*	65.4 (26)*	77.3 (22)*	76.0 (25)	76.9 (26)
Rerun	72.7 (22)	60.0 (25)	57.7 (26)	59.1 (22)	52.0 (25)	53.8 (26)
ECMWF	71.4 (21)	62.5 (24)	64.0 (25)	45.5 (22)	52.0 (25)	53.8 (26)
NAM218	54.5 (22)	48.0 (25)	46.2 (26)	50.0 (20)	47.8 (23)	41.7 (24)
HRRR12	65.0 (20)	54.5 (22)	52.2 (23)	45.0 (20)	45.5 (22)	43.5 (23)
HRRR	45.0 (20)	40.9 (22)	39.1 (23)	45.0 (20)	45.5 (22)	47.8 (23)
CFSv2	81.8 (22)*	76.0 (25)	73.1 (26)	52.4 (21)*	50.0 (24)	48.0 (25)
GFS	63.6 (22)*	64.0 (25)*	61.5 (26)	40.9 (22)	44.0 (25)	46.2 (26)
Geoensemble mean	90.9 (22)	92.0 (25)	100.0 (26)	77.3 (22)	88.0 (25)	100.0 (26)
$CI_{\min} = 13$	100.0 (22)	59.1 (22)	45.5 (22)	100.0 (22)	77.3 (22)	72.7 (22)
$CI_{\min} = 5$	81.8 (22)	100.0 (25)	68.0 (25)	77.3 (22)	100.0 (25)	80.0 (25)
Control	81.8 (22)*	72.0 (25)	61.5 (26)	63.6 (22)	64.0 (25)	57.7 (26)

man’s ensemble” was computed using the NAM 5 km, NAM 12 km, rerun, ECMWF, and GFS. The HRRR and CFSv2 have not been included as the available data do not span the 60-h forecast with one forecast run. The CRPS of the poor man’s ensemble is 8.5, higher than most of the selected geoensemble scores, shown in Fig. 3. Since the CRPS is negatively oriented, that is, lower is better, the selected geoensemble performs better.

It would be desirable if the selection process would be entirely independent of the variable for verification. Resolving this shortcoming is the focus of our current research in this area. It would be preferable if other variables could be identified as useful precursors, such as surface temperature or mean sea level pressure. Their more continuous nature and more reliable observation are just two desirable features. An in-depth analysis of what type of perturbation will produce good forecasts can also aid in selecting members. An example of how perturbations can lead to improved forecasts apart from just sampling the chaotic system are perturbations that correct or counteract biases of the driving models. The perturbations can also correct resolution-dependent shortcomings in the driving models, such as lack of orographic detail, or surface type miss classification, to name just two. It is expected that this will lead to better, more independent selection criteria.

c. Specific forecast performance in the context of the definition of forecast success

With a definition of forecast success, the forecast performance can be evaluated relative to the forecast target application. Some applications may need to minimize false alarms. In emergency response, for example,

too many false alarms will desensitize the public. Other applications place more emphasis on prediction without needing to know the precise amount. Such as the tourist sector, where it is more important to know whether it will rain or not, but not necessarily by how much.

Here, the results are further interpreted using the four following parameters:

- N in the definition of forecast success, which depends on the target use of the forecast,
- the CI is required as a minimum for selection,
- γ is the definition of the selection criterion, and
- GEr is the geoensemble size.

All analysis has been performed on the highest-resolution domain, which is d03 for the ensemble, and d01 for the rerun experiment. While the choice of N depends on the use of the forecast, the remaining three free parameters need to be determined by long-term evaluation.

For GE1, the number of false alarms, hits, and misses are visualized with respect to γ and $CI_{\min}$ in Fig. 4. The number of false alarms is mostly invariant to the requested $CI_{\min}$. However, the overall number of geomembers reduces monotonically with increasing $CI_{\min}$. These rejected geomembers come mainly from the hits and misses. The misses generally reduce with increasing $CI_{\min}$, a smaller γ resulting in a faster reduction of misses. Hits do mostly increase with increasing $CI_{\min}$, but there are also occasions where an increase of $CI_{\min}$ reduces the number of hits. This reduction of hits means that not only poorly performing geomembers are rejected but also geomembers that, at the end of the run at least, had an adequate amount of precipitation.

FIG. 3. Mean ensemble CRPS for the 60-h accumulated precipitation forecast over all case studies. The mean CRPS of an ensemble of NAM 5 km, NAM 12 km, rerun, ECMWF, and GFS scores a CRPS of 8.5, which is surpassed by all geoenvironmental forecasts with γ less than 2.5. It can be seen how generally the forecast improves with selection (higher CI_{min}) and larger geoenvironmental (GE). All figures use identical color mapping. The color bars represent the range of values present in each subpanel.

This can occur for mainly three reasons: the precipitation is correctly represented in space but delayed in time (which motivates the use of accumulated precipitation in this study), there is a small error in space, which means that the precipitation event is missed—albeit being present in the forecast (which inspired and probably partly explains the success of the geomembers), or, there is little or no precipitation in the first 24 h, where the method presented here fails. GE1 marginally improved the performance of the ensemble when compared to GE0 and GE2 results. However, the overall pattern does not change. Here only GE1 results are presented, the

results for all GE0, GE1, and GE2 can be found in the online supplemental material.

The selection period has to be kept in mind when comparing the forecasts. If, for example, the NAM forecast with initial time $t + 24$ performs better than the ensemble with initial time t and 24-h selection period, then it would not have been worthwhile to run the ensemble from the deterministic forecast point of view. The ensemble would, however, add a probabilistic component, irrespective of the performance against the deterministic forecast. This probabilistic component can be valuable in particular for extreme events.

FIG. 4. False alarms, hits, and misses for different γ .

In the following, for each case study, five forecasts are analyzed and the resulting skills compared. These forecasts are

- 1) ens, a subgeoensemble selected by the CI-based procedure,
- 2) NAM, the original NAM forecast evaluated at the nearest grid point,
- 3) rerun, a forecast rerun with the ensemble model configuration, but which is deterministic and has a resolution of only 5 km,
- 4) NAM24, the NAM forecast with initial date 24 h later than the initial date of the ensemble, and
- 5) rerun24, a deterministic forecast with the ensemble model configuration, but with an initial date 24 h later than the initial date of the ensemble.

The difference plots in Fig. 5, columns 2–5, from left to right, indicate an improvement in the geoensemble run in green tones and a deterioration of skill in magenta tones. The whole 60-h forecast period is evaluated (i.e., the accumulated precipitation at forecast hour 60 is compared to the observation over the same period; in contrast to Fig. 6, where the selection period has been excluded). The green dashed line is the zero isoline. When compared to the NAM forecast of the same initial date and time as the geoensemble, the geoensemble performs well, as shown in the second column from the left. Bias, SR, and TS show improvement through the

whole parameter range. Only for POD, the NAM provides slightly better results. The 5-km deterministic rerun results, fourth column, is at times better than the ensemble in terms of BIAS and TS. In POD, however, the ensemble improves, in particular for higher CI_{min} . This improvement relative to the NAM indicates that the model configuration of the ensemble is more suitable for the region than the NAM configuration. The comparison with the NAM24 forecast, third column, is expected to be superior, due to improved initial and boundary conditions from the data assimilation. However, only in POD and TS does this advantage materialize. BIAS and SR are still substantially enhanced by the ensemble. Compared with the GE1, the best forecast was made by the 5-km resolution deterministic “rerun24,” with the ensemble model configuration. Nevertheless, the BIAS, and SR still, at times, show a slight improvement from the ensemble.

Generally, the ensemble has worse POD scores, which is likely due to the comparison against models that seem to suffer from overforecasting precipitation events. The NAM and rerun models have frequency biases above one (Fig. 7) and higher than the ensemble, which leads to a larger amount of false alarms. Due to this the ensemble generally has lower FAR and larger SR scores but does not seem to show an improved TS except for a larger CI_{min} and N . Such an overprediction problem has been identified in the NAM

FIG. 5. Bias, probability of detection (POD), success ratio (SR), and threat score (TS) for the ensemble, NAM, NAM24, rerun, and rerun24, with $GE = 1$ and $\gamma = 1$. NAM and NAM24 have the same valid time, but their initial time is 24 h delayed; rerun and rerun24 are similar. Green shading in the difference plots (columns 2–5, from left to right) indicates a better performing ensemble, whereas magenta favors the respective deterministic run.

(Gowan et al. 2018b) and attributed to the particular convection scheme used in the NAM. The ensemble at 500-m resolution does not rely on parameterizations for convection. It has been demonstrated that such convection-permitting models generate superior results (Schwartz et al. 2009).

Based on the rerun24, for which the statistical metrics give excellent results, it would seem that the additional cost of the computations of the ensemble is less well justified. It is therefore essential to benefit from the ensemble early (i.e., to reduce the selection period). That this can be successful will be shown below on a particular case study.

The false discrimination of geomembers by the selection procedure and its detrimental effect becomes even more apparent when performing the same analysis but restricting the skill evaluation to the precipitation of the last 36 h of the 60-h forecast period, excluding the first 24 h that have been used for selecting geomembers. As shown in Fig. 6, the performance of the selected ensemble is worst around the $CI_{\min} = 15$ mark. However, when setting the $CI_{\min}$ to 23, the results improve markedly:

- while POD and TS are only slightly improved by the selection and the former is, as expected from the previous results, worse than the two 24-h delayed deterministic runs and TS is only better than the deterministic rerun24 for $N < 0.21$ and $N > 0.39$,
- the bias is closer to one than the nonselected ensemble with $N > 0.375$, and
- SR increased (FAR reduced, not shown).

Figure 7 shows this more readily in a line plot for different N .

d. Statistical significance

One route to establish the statistical significance of the presented results is to state the null hypothesis that no differences between the NAM24 and ensemble forecast exist. Then the aim would be to reject this hypothesis in favor of the alternative: A significant difference exists. This is achieved by computing the probability of the observed differences in the time series of hits and misses and performing a binomial test (Berger and Zhou 2005). Two independent tests are performed, one for hits, the other for misses. The hits and misses are interpreted as votes. If the NAM24 has a hit, it is a vote for the NAM24. If the ensemble has a hit, it is a vote for the ensemble. If neither or both have a hit, it is an invalid vote. Two events are defined, one is the NAM24 receiving the most votes, the other is the geoensemble receiving the most votes. The probability

of one of these events occurring by chance can be computed by

$$p = \sum_{k=s-1}^n \binom{n}{k} 0.5^n, \quad (9)$$

where s represents the number of votes for one candidate and n the total number of valid votes. If this probability is sufficiently low, the hypothesis that these results occur by chance can be rejected. A significance level of 0.15 was assumed, and only events with $n \geq 8$ were considered. Figure 8 shows the obtained p values in Figs. 8a and 8b and corresponding n values in Figs. 8c and 8d. In Fig. 8a the 0.15 isoline is plotted in red. All p values for the misses (Fig. 8b) are below 0.15, where the minimum of eight votes was satisfied. These results show significance at low N and high $CI_{\min}$ for the hits and nearly for all $CI_{\min}$ and low N in the misses. A larger sample size would, nevertheless, have been desirable here.

e. Case studies

To illustrate the presented method, two case studies have been selected based on the occurrence of significant precipitation in the first 24 h and the remaining 36-h period. Furthermore, to investigate the impact of the length of the selection period, a case study date with precipitation over nearly the whole 24-h selection period was chosen for the second case study.

1) PRECIPITATION RESULTS FOR THE 4 JANUARY 2016 CASE STUDY

On 4 January 2016, the ensemble geomembers fell into discrete bands, but none were close enough to the observations to satisfy the $\gamma = 0.5$ and $CI_{\min} = 20$ criteria. Therefore the criterion was relaxed to $\gamma = 1$. The geomembers for the different geoensembles with the respective mean forecasts and selection criterion are shown in Fig. 9. The NAM24 overestimated the observed precipitation of 31 mm by 14 mm to 44.9 mm. The ensemble corrected this forecast to 28.3 mm, only a 2.7-mm difference, which is even better than the rerun24, which predicted 26.95 mm. The GE3 result shows more clearly the clustering of the forecasts into similar families, from three to four between forecast hours 0 and 24 and two for forecast hours 24–60. However, the final prediction is not affected by a particular choice of GE. The choice of γ , however, in both cases influences the forecast with a too broad (large) γ resulting in a nearly 50% underestimation of precipitation because it leads to the inclusion of members who show no or little precipitation in between hours 24 and 45.

FIG. 6. As in Fig. 5, but only showing the differences against NAM24 and rerun24 (i.e., restricting the evaluation period to the forecast hours 24–60, excluding the selection period).

FIG. 7. The effect of geomember selection on ensemble performance on (a) bias, (b) probability of detection, (c) success ratio, (d) and threat score, when comparing the ensemble forecast hours 24–60 with the rerun24 and NAM24 at forecast hours 0–36.

2) PRECIPITATION RESULTS FOR THE 22 DECEMBER 2016 CASE STUDY

The use of precipitation as the selection variable limits the scope of the presented method. Naturally, it would be desirable to issue improved forecasts early, and not only after 24 h, a time at which improved deterministic forecasts are already available. To investigate the performance of the ensemble with less than 24-h selection period, a date with significant precipitation throughout both, the first 24 h and the subsequent 36 h, was selected. This forecast has an initial time of 0000 UTC 22 December 2016. The forecast accumulated precipitation in the geoensemble mean, the individual geomembers, NAM, and NAM24 over forecast hours 0–24 and 24–60 is shown in Fig. 10. The NAM forecast predicted more than 120-mm accumulated precipitation and more than half of that between 1500 and 1800 UTC. The observation does not verify this extreme forecast. As shown in Fig. 10, for the first 24 h the observation recorded around 40 mm of precipitation. The two

panels show the same information, apart from the geoensemble, which increases from Fig. 10a to Fig. 10b. The selection criteria of $\pm 0.5\sigma$, $\pm\sigma$, and $\pm 2\sigma$ are shown by the filled regions. Due to the shorter selection times it is not possible to obtain a $CI > 20$. Here the largest CI_{min} supported by the geomembers are selected. This means the required CI_{min} is increased until there is no geomember available anymore to support an even higher number, meaning that requesting a larger CI_{min} would reject all geomembers. The respective CI_{min} values used are shown in the figure legend with the respective selection time span (or lead). For the 6-, 12-, and 18-h lead the forecasts are nearly identical. For the first 24 h, the ensemble overestimates the precipitation, but only to approximately 2 times the observed values and not 3 times, as the NAM does.

For forecast hours 25–60 the 6-h lead time forecast is better than the 24-h lead forecast, which is plotted with a dashed line. The increasing standard deviation of the ensemble precipitation at later forecast hours means

FIG. 8. Probability p and event size n for hits and misses. The 0.15 confidence interval is indicated using a red line, where appropriate. The p values associated with events sizes of less than eight have been removed; $\gamma = 0.5$ and GE1.

that potentially more geomembers can be included in the selected ensemble. These geomembers, as can be seen in Fig. 10, mainly underestimate the precipitation. Hence the mean of the geomembers also decreases.

Interestingly, for the second period, forecast hours 24–60, nearly all geomembers predict the precipitation at forecast hour 60. Also, the NAM24 accumulates most of its precipitation around this time. However, the observation shows no precipitation past forecast hour 40. Only the 6-, 12-, and 18-h lead ensemble forecasts capture the main precipitation events in forecast hours 24–40 and hence have, at the end of the forecast period, nearly the correct amount of accumulated precipitation *for the correct reason*. Most of the other geomembers, and the NAM and NAM24 forecasts, accumulated a nearly correct amount of precipitation, but at a time when no precipitation was observed.

5. Conclusions

While Figs. 5 and 6 may at first indicate that there is little value in selecting geomembers, it is clear from Fig. 7 that the ensemble with $CI > 20$ has improved skill. Figure 3 shows that there is a clear dependence of skill on the selection of geomembers. The selection criterion γ influences the overall skill of the ensemble. A γ of one and strict selection of geomembers with $CI > 10$ can reduce the CRPS from around 8 to 6.25 for the 60-h forecast. Choosing a GE3 versus as GE0 can also reduce the CRPS from around 7 to 6.25.

The added value of the ensemble approach has been demonstrated with a series of skill measures, when compared to deterministic forecasts with the same initial date. The focus was on accumulated precipitation at the end of a 60-h forecast. While the mesoscale forecast used to initialize the ensemble had little skill at the gridpoint comparison, the ensemble, combined with the higher

FIG. 9. Ensemble prediction over the forecast period 0–24 and 24–60 h. Transparent lines represent individual ensemble geomembers for (a) GE0 and (b) GE3. The observation is indicated by the blue line and the total accumulated precipitation over forecast hours 24–60 of NAM24, rerun24, and ensemble with $CI_{\min} = 20$ and $\gamma = 1$ and 2, respectively, are indicated by the circles. The shaded areas indicate the selection area with increasing γ around the observation.

resolution, increased the forecast skill. When compared to the 24-h delayed forecasts, the increase of skill through the use of the ensemble is reduced. This reduction means that it is critical to select the geomembers early, to capitalize on the improved forecast early.

The experiments with reduced selection periods (cf. Fig. 10) have shown that even periods as short as 6 h can produce forecasts closer to the observations.

A total of 24 hours can make a difference when it comes to the mobilization of emergency response, and sensibilization and potential evacuation of civilians. Also, the elimination of false alarms, which the ensemble has performed well, even in the most challenging

comparison with the rerun24, is arguably a crucial aspect of the methodology presented here. Predicting inundations and flash floods where no such event materialize later on erodes public confidence and trust. While these outliers have some chance of becoming a reality, civil protection has to take these seriously. However, communicating such extreme events, extreme in both their intensity and low probability, to the general public will do more damage than good when it eventually comes to a real catastrophe, a highly likely very extreme event with a public reluctant to take the warning seriously.

Current work focuses on the identification of suitable predictors other than precipitation, to eliminate the

FIG. 10. As in Fig. 9, but also showing the evolution of the ensemble mean when selection is performed under different lead times (cyan).

failure in cases without precipitation. Also using variables with lower spatial variability in the selection criterion, would improve the method, because it would be less likely to reject otherwise well-performing geomembers. Furthermore, the use of personal weather stations and remotely sensed observations by satellite and/or radar, which would complement the one official observation in the study area, is investigated. This data will then enable the authors to evaluate the forecast over a larger region.

It is important to note that the analysis with different lead times has highlighted a potential problem with the methodology. First, it was hypothesized that geomembers are selected based on their performance, arguing that a good performance early on will be more likely to be followed with a good performance later in the forecast cycle. To emphasize this logic, earlier it was stated that a poorly performing geomember would for no physical reason return to become a well-performing geomember. This statement, in particular when using precipitation as guidance, cannot be understood as being black and white. Precipitation fields often have strong precipitation in small areas, surrounded by moderate precipitation in the neighboring areas. They have also been described as spatially patchy (Tapiador et al. 2019). Missing such a “hot-spot” or in reverse, forecasting it at the observation location while in reality, it occurred a few hundred meters away from the observation, can then quickly lead to false discrimination against geomembers that, in the overall picture, performed well.

The potential impacts of the presented method range from aiding the agricultural sector, by providing more accurate forecasts regarding quantity and locality of precipitation to assisting emergency relief operations. From a research perspective, the method has shown its ability in eliminating severe outliers, which at times manifest themselves in the mesoscale forecasts. These outliers can too readily cause unnecessary concern, and the communication of such erroneous predictions can erode the support of the public and political decision-makers.

Acknowledgments. The authors greatly appreciate constructive comments by the anonymous reviewers, which have improved the manuscript markedly, the encouragement by the editor, and the support of Dr. Jonás D. De Basabe, CICESE, for providing access to the cluster lamb.

REFERENCES

- Abdalla, S., L. Isaksen, P. Janssen, and N. Wedi, 2013: Effective spectral resolution of ECMWF atmospheric forecast models. *ECMWF Newsletter*, No. 137, ECMWF, Reading, United Kingdom, 19–22, <https://doi.org/10.21957/rue4o7ac>, <https://www.ecmwf.int/en/elibrary/17358-effective-spectral-resolution-ecmwf-atmospheric-forecast-models>.
- Ancell, B. C., 2016: Improving high-impact forecasts through sensitivity-based ensemble subsets: Demonstration and initial tests. *Wea. Forecasting*, **31**, 1019–1036, <https://doi.org/10.1175/WAF-D-15-0121.1>.
- Berger, V. W., and Y. Zhou, 2005: Binomial distribution: Estimating and testing parameters. *Encyclopedia of Statistics in Behavioral Science*, B. Everitt and D. Howell, Eds., Wiley, <https://doi.org/10.1002/0470013192.bsa051>.
- Black, T., H. M. H. Juang, and M. Iredell, 2009: The NOAA Environmental Modeling System at NCEP. *23rd Conf. on Weather Analysis and Forecasting/19th Conf. on Numerical Weather Prediction*, Omaha, NE, Amer. Meteor. Soc., 2A.6, https://ams.confex.com/ams/23WAF19NWP/techprogram/paper_154223.htm.
- Blaylock, B. K., J. D. Horel, and S. T. Liston, 2017: Cloud archiving and data mining of high-resolution rapid refresh forecast model output. *Comput. Geosci.*, **109**, 43–50, <https://doi.org/10.1016/j.cageo.2017.08.005>.
- Candille, G., 2009: The multiensemble approach: The NAEFS example. *Mon. Wea. Rev.*, **137**, 1655–1665, <https://doi.org/10.1175/2008MWR2682.1>.
- Clark, A. J., and Coauthors, 2012: An overview of the 2010 Hazardous Weather Testbed Experimental Forecast Program Spring Experiment. *Bull. Amer. Meteor. Soc.*, **93**, 55–74, <https://doi.org/10.1175/BAMS-D-11-00040.1>.
- , and Coauthors, 2018: The Community Leveraged Unified Ensemble (CLUE) in the 2016 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. *Bull. Amer. Meteor. Soc.*, **99**, 1433–1448, <https://doi.org/10.1175/BAMS-D-16-0309.1>.
- Done, J., C. A. Davis, and M. Weisman, 2004: The next generation of NWP: Explicit forecasts of convection using the Weather Research and Forecasting (WRF) model. *Atmos. Sci. Lett.*, **5**, 110–117, <https://doi.org/10.1002/asl.72>.
- Dong, L., and F. Zhang, 2016: OBEST: An observation-based ensemble subsetting technique for tropical cyclone track prediction. *Wea. Forecasting*, **31**, 57–70, <https://doi.org/10.1175/WAF-D-15-0056.1>.
- Du, J., and M. S. Tracton, 2001: Implementation of a real-time short-range ensemble forecasting system at NCEP: An update. Preprints, *Ninth Conf. on Mesoscale Processes*, Fort Lauderdale, FL, Amer. Meteor. Soc., 355–356.
- Ebert, E. E., 2001: Ability of a poor man’s ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480, [https://doi.org/10.1175/1520-0493\(2001\)129<2461:AOAPMS>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2).
- Fosser, G., E. Kendon, S. Chan, A. Lock, N. Roberts, and M. Bush, 2019: Optimal configuration and resolution for the first convection-permitting ensemble of climate projections over the United Kingdom. *Int. J. Climatol.*, <https://doi.org/10.1002/joc.6415>, in press.
- Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.*, **102**, 359–378, <https://doi.org/10.1198/016214506000001437>.
- Gowan, T. M., W. J. Steenburgh, and C. S. Schwartz, 2018a: Validation of mountain precipitation forecasts from the convection-permitting NCAR ensemble and operational forecast systems over the western United States. *Wea. Forecasting*, **33**, 739–765, <https://doi.org/10.1175/WAF-D-17-0144.1>.
- , —, and —, 2018b: Validation of mountain precipitation forecasts from the convection-permitting NCAR ensemble

Kingdom, 19–22, <https://doi.org/10.21957/rue4o7ac>, <https://www.ecmwf.int/en/elibrary/17358-effective-spectral-resolution-ecmwf-atmospheric-forecast-models>.

- and operational forecast systems over the western United States. *Wea. Forecasting*, **33**, 739–765, <https://doi.org/10.1175/WAF-D-17-0144.1>.
- Jones, P. W., 1999: First- and second-order conservative remapping schemes for grids in spherical coordinates. *Mon. Wea. Rev.*, **127**, 2204–2210, [https://doi.org/10.1175/1520-0493\(1999\)127<2204:FASOCR>2.0.CO;2](https://doi.org/10.1175/1520-0493(1999)127<2204:FASOCR>2.0.CO;2).
- Kain, J. S., S. J. Weiss, J. J. Levit, M. E. Baldwin, and D. R. Bright, 2006: Examination of convection-allowing configurations of the WRF model for the prediction of severe convective weather: The SPC/NSSL spring program 2004. *Wea. Forecasting*, **21**, 167–181, <https://doi.org/10.1175/WAF906.1>.
- , and Coauthors, 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Wea. Forecasting*, **23**, 931–952, <https://doi.org/10.1175/WAF2007106.1>.
- Klaver, R., R. Haarsma, P. L. Vidale, and W. Hazeleger, 2020: Effective resolution in high resolution global atmospheric models for climate studies. *Atmos. Sci. Lett.*, <https://doi.org/10.1002/asl.952>, in press.
- Köppen, W., and R. Geiger, Eds., 1936: Das geographische System der Klimate. *Handbuch der Klimatologie*, Vol. 1, Borntraeger, 1–44.
- Kottke, M., J. Grieser, C. Beck, B. Rudolf, and F. Rubel, 2006: World map of the Köppen-Geiger climate classification updated. *Meteor. Z.*, **15**, 259–263, <https://doi.org/10.1127/0941-2948/2006/0130>.
- Mass, C. F., D. Ovens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Meteor. Soc.*, **83**, 407–430, [https://doi.org/10.1175/1520-0477\(2002\)083<0407:DIHRPM>2.3.CO;2](https://doi.org/10.1175/1520-0477(2002)083<0407:DIHRPM>2.3.CO;2).
- NOAA/NCEP/NWS, 2015a: NCEP GFS 0.25 degree global forecast grids historical archive. National Center for Atmospheric Research, Computational and Information Systems Laboratory, accessed August 2019, <https://doi.org/10.5065/D65D8PWK>.
- , 2015b: NCEP North American Mesoscale (NAM) 12 km analysis. National Center for Atmospheric Research, Computational and Information Systems Laboratory, accessed August 2019, <https://doi.org/10.5065/G4RC-1N91>.
- Palmer, T. N., 2002: The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quart. J. Roy. Meteor. Soc.*, **128**, 747–774, <https://doi.org/10.1256/0035900021643593>.
- Parrish, D. F., and J. C. Derber, 1992: The National Meteorological Center's spectral statistical-interpolation analysis system. *Mon. Wea. Rev.*, **120**, 1747–1763, [https://doi.org/10.1175/1520-0493\(1992\)120<1747:TNMCSS>2.0.CO;2](https://doi.org/10.1175/1520-0493(1992)120<1747:TNMCSS>2.0.CO;2).
- Pavia, E. G., F. Graef, and R. Fuentes-Franco, 2016: Recent ENSO–PDO precipitation relationships in the Mediterranean California border region. *Atmos. Sci. Lett.*, **17**, 280–285, <https://doi.org/10.1002/asl.656>.
- Peckham, S. E., T. G. Smirnova, S. G. Benjamin, J. M. Brown, and J. S. Kenyon, 2016: Implementation of a digital filter initialization in the WRF Model and its application in the Rapid Refresh. *Mon. Wea. Rev.*, **144**, 99–106, <https://doi.org/10.1175/MWR-D-15-0219.1>.
- Perconti, P., Ed., 2017: History of the U.S. Army research Laboratory. ARL, Government Printing Office, 75 pp., <https://www.arl.army.mil/wp-content/uploads/2019/10/History-of-the-U.S.-Army-Research-Laboratory.pdf>.
- Qi, L., H. Yu, and P. Chen, 2013: Selective ensemble-mean technique for tropical cyclone track forecast by using ensemble prediction systems. *Quart. J. Roy. Meteor. Soc.*, **140**, 805–813, <https://doi.org/10.1002/qj.2196>.
- Robert, D., R. Brian, S. Jeffrey, K. David, and C. Huaqing, 2014: A WRF-based mixed variational and nudging assimilation scheme for US army convection-scale nowcasting. *The World Weather Open Science Conf.*, Las Cruces, NM, WMO, 1–18, https://www.wmo.int/pages/prog/arep/wwrp/new/wwosc/Presentations_wwosc_17082014.html.
- Rubel, F., and M. Kottke, 2010: Observed and projected climate shifts 1901–2100 depicted by world maps of the Köppen-Geiger climate classification. *Meteor. Z.*, **19**, 135–141, <https://doi.org/10.1127/0941-2948/2010/0430>.
- Saha, S., and Coauthors, 2011: NCEP climate forecast system version 2 (CFSv2) 6-hourly products. National Center for Atmospheric Research, Computational and Information Systems Laboratory, accessed August 2019, <https://doi.org/10.5065/D61C1TXF>.
- Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting*, **5**, 570–575, [https://doi.org/10.1175/1520-0434\(1990\)005<0570:TCSIAA>2.0.CO;2](https://doi.org/10.1175/1520-0434(1990)005<0570:TCSIAA>2.0.CO;2).
- Schwartz, C. S., and R. A. Sobash, 2017: Generating probabilistic forecasts from convection-allowing ensembles using neighborhood approaches: A review and recommendations. *Mon. Wea. Rev.*, **145**, 3397–3418, <https://doi.org/10.1175/MWR-D-16-0400.1>.
- , and Coauthors, 2009: Next-day convection-allowing WRF Model guidance: A second look at 2-km versus 4-km grid spacing. *Mon. Wea. Rev.*, **137**, 3351–3372, <https://doi.org/10.1175/2009MWR2924.1>.
- Skamarock, W. C., 2004: Evaluating mesoscale NWP models using kinetic energy spectra. *Mon. Wea. Rev.*, **132**, 3019–3032, <https://doi.org/10.1175/MWR2830.1>.
- , and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp., <http://doi.org/10.5065/D68S4MVH>.
- Tapiador, F. J., R. Roca, A. D. Genio, B. Dewitte, W. Petersen, and F. Zhang, 2019: Is precipitation a good metric for model performance? *Bull. Amer. Meteor. Soc.*, **100**, 223–233, <https://doi.org/10.1175/BAMS-D-17-0218.1>.
- Theis, S. E., A. Hense, and U. Damrath, 2005: Probabilistic precipitation forecasts from a deterministic model: A pragmatic approach. *Meteor. Appl.*, **12**, 257–268, <https://doi.org/10.1017/S1350482705001763>.
- Thielen, J., K. Bogner, F. Pappenberger, M. Kalas, M. del Medico, and A. de Roo, 2009: Monthly-, medium-, and short-range flood warning: Testing the limits of predictability. *Meteor. Appl.*, **16**, 77–90, <https://doi.org/10.1002/met.140>.
- Tolman, H., 2014: Descriptions of the major modeling systems operated at NOAA/NWS/NCEP. NOAA Tech. Rep., 70 pp.
- Weidle, F., Y. Wang, and G. Smet, 2016: On the impact of the choice of global ensemble in forcing a regional ensemble system. *Wea. Forecasting*, **31**, 515–530, <https://doi.org/10.1175/WAF-D-15-0102.1>.
- Weisman, M. L., C. Davis, W. Wang, K. W. Manning, and J. B. Klemp, 2008: Experiences with 0–36-h explicit convective forecasts with the WRF-ARW Model. *Wea. Forecasting*, **23**, 407–437, <https://doi.org/10.1175/2007WAF2007005.1>.
- Yan, H., and W. A. Gallus, 2016: An evaluation of QPF from the WRF, NAM, and GFS models using multiple verification methods over a small domain. *Wea. Forecasting*, **31**, 1363–1379, <https://doi.org/10.1175/WAF-D-16-0020.1>.