

Convection-permitting ensembles: Challenges related to their design and use

Inger-Lise Frogner¹  | Andrew T. Singleton¹ | Morten Ø. Køltzow¹ | Ulf Andrae²

¹Development centre for weather forecasting, Norwegian Meteorological Institute, Oslo, Norway

²Research department, Swedish Meteorological and Hydrological Institute, Norrköping, Sweden

Correspondence

Inger-Lise Frogner, Norwegian Meteorological Institute, PO Box 43, Blindern N-0313, Oslo, Norway.

Email: i.l.frogner@met.no

Challenges related to the design and use of a convection-permitting ensemble (CPEPS) are discussed. In particular the scale-dependent predictability of precipitation and the use of a CPEPS as well as its potential added value over global ensemble prediction systems (EPS) are investigated. Forecasts of precipitation from the operational CPEPS in Finland, Norway and Sweden (MEPS) are used for the investigations. It is found that predictability for scales smaller than ~60 km is lost rapidly within the first 6 h of the forecast with the smallest predictable scale growing more slowly to ~100 km over the following 18–24 h. However, there is large case-to-case variability and the ensemble perturbations fail to become fully saturated, especially in winter, suggesting a weakness in the design of the ensemble. The added value of CPEPS over deterministic forecasts and coarser resolution EPSs is discussed with summary statistics and case-studies. It is shown that the added value varies between seasons and lead times. For precipitation there is an added value for both severe precipitation events and for precipitation/no precipitation decisions. The added value is higher in summer compared to winter and for shorter lead times compared to longer lead times.

KEYWORDS

added value, convection-permitting, CPEPS, ensemble prediction, EPS, MEPS, precipitation, scale dependency

1 | INTRODUCTION

The classical view of predictability is that the skill of a forecast, measured as the correlation between the forecast and an analysis or observations, immediately goes to zero at higher wave numbers (smaller scales), and with a loss of skill (growth of error) progressively penetrating to lower wave numbers with forecast range (e.g. Lorenz, 1969; Boer, 2003). The forecast time at which the predictability limit is reached is therefore shorter for forecasts of small-scale features than for large-scale features. The chaotic nature of the atmosphere imposes a finite limit on the predictability of atmospheric processes – this is known as the intrinsic predictability and can be thought of as the maximum predictability attainable with a perfect model with minimal initial

condition errors (Lorenz, 1969). However, current numerical weather prediction (NWP) models are imperfect and have uncertainties associated with both the initial conditions and modelling procedures. This further limits predictability and gives rise to the concept of practical predictability, that is, the predictability of a particular forecasting model. For the purposes of this article, the emphasis is on practical predictability rather than intrinsic predictability.

When predictability was investigated in the European Centre for Medium-range Weather Forecasts (ECMWF) model for temperature at 850 hPa (T850) (Simmons, 2003), three regimes of forecast skill were found. The largest scales (low spatial wave numbers), which represent climatologically forced structures for which errors grow slowly, have little impact on short-range forecasts. The expected, classical error

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2019 The Authors. *Quarterly Journal of the Royal Meteorological Society* published by John Wiley & Sons Ltd on behalf of the Royal Meteorological Society.

growth was found at intermediate scales where forecast skill goes to zero immediately at small scales and with loss of predictive skill penetrating to larger spatial scales with forecast time, and saturated by the 10-day forecast limit. Unexpected behaviour was found at small scales (high spatial wave numbers) where instead of the rapid error saturation and loss of skill, skill appeared to exist up to day 10. It can be argued that the latter is due to the atmosphere's interaction with topography, changes in surface roughness, thermodynamic forcing at coastlines, lakes, snow/ice margins and other static forcing mechanisms.

Jung and Leutbecher (2008) compared the predictability for deterministic forecasts (HRES) with that for probabilistic forecasts (IFSSENS) from ECMWF for 500 hPa geopotential (Z500). They also found that the skill for the sub-synoptic scale was much lower than for the synoptic scales. However, they showed that the skill was substantially higher for IFSSENS than for HRES. Such a result highlights the importance of a probabilistic approach for weather forecasts, especially at the small scales.

In recent years convection-permitting ensemble prediction systems (CPEPSs) with grid spacing typically between 1 and 4 km have been developed and are used by many Met Services (e.g. Gebhardt *et al.*, 2008; Bouttier *et al.*, 2012; Romine *et al.*, 2014; Schwartz *et al.*, 2015; Bouttier *et al.*, 2016; Frogner *et al.*, 2016; Hagelin *et al.*, 2017). Due to the limitations on the predictability at these scales, as discussed above, an assessment on the scale-dependent predictability of such a system, and to what extent it adds value over a global EPS, is important, in order to justify the cost of running a CPEPS. For convection-permitting models, it is more instructive to analyse predictability as a function of spatial scale for near-surface weather parameters than for Z500. In a more recent study by Surcel *et al.* (2015), predictability for precipitation was investigated in a 10-member ensemble with 4 km grid spacing. They found that, on average for the 22 cases under study, the smallest predictable scale grew as a function of lead time following a power law. In this article we also focus on precipitation and take a closer look at two challenges connected with CPEPS: scale-dependent predictability, where the method of Surcel *et al.* (2015) is used, and the use of CPEPS and to what extent it adds value over a global EPS. By choosing these two challenges, two very different perspectives of probability forecasts are discussed: (a) The study of scale-dependent predictability gives guidance as to which scales and forecast times can be trusted. Additionally, studying the way new developments and perturbation techniques alter the scale-dependent error growth, it can also be used as a tool in the development process. (b) The possible added value of CPEPS over global EPS gives guidance on when it is beneficial to use a CPEPS. For the developers it points to what situations (parameters, forecast lengths, seasons, etc.) to focus on when improving the system.

Both the predictability dependence on scale and the use and possible added value of CPEPS will be illustrated using



FIGURE 1 The operational domain for MEPS used for the data analysis and examples [Colour figure can be viewed at wileyonlinelibrary.com]

precipitation forecasts from MEPS (MetCoOp ensemble prediction system). MetCoOp is the operational cooperation between the national meteorological institutes of Finland, Norway and Sweden, and MEPS is the operational EPS of the cooperation and is described in section 2. Predictability studies with MEPS are in section 3 using the method of Surcel *et al.* (2015). In section 4 we discuss the possible added value of MEPS over IFSSENS and over its deterministic control. We also illustrate added value with two case-studies from MEPS. Section 5 consists of discussions and conclusions.

2 | THE METCOOP ENSEMBLE PREDICTION SYSTEM: MEPS

MEPS has served as an operational CPEPS for Finland, Norway and Sweden since November 2016. MEPS is based on the HIRLAM–ALADIN Research on Mesoscale Operational NWP in Euromed – Applications of Research to Operations at Mesoscale (HARMONIE-AROME) configuration of the ALADIN-HIRLAM system (Termonia *et al.*, 2018). The general aspects of the non-hydrostatic dynamics and physical parametrization are described by Termonia *et al.* (2018) and the details specific to the HARMONIE-AROME configuration in Bengtsson *et al.* (2017). The forecast model is run with 2.5 km grid spacing with 65 levels in the vertical and spans 900×960 grid points in the zonal and meridional directions respectively, see Figure 1. Note that there is no distinction between the control and perturbed members in terms of resolution, nor does MetCoOp run a higher-resolution forecast in addition to MEPS. The data assimilation system applied on the control member is based on three-dimensional variation (3D-Var) with a 3 h cycle using a combination

of conventional and non-conventional observations such as radar reflectivity, Global Navigation Satellite System (GNSS) total delay, satellite radiances and Advanced Scatterometer (ASCAT) winds (Müller *et al.*, 2017). The cut-off time for observations to enter the data assimilation is 1 h and 15 min. The surface initial condition is controlled by an optimal interpolation-based method for two-metre temperature ($T2m$), two-metre relative humidity ($RH2m$) and snow (Giard and Bazile, 2000). The ensemble runs with nine perturbed members, each of them running their own surface assimilation with a cycle of 6 h. Each member is perturbed in three different ways with respect to initial and boundary conditions. At first, perturbations are constructed in pairs by subtracting two ECMWF HRES forecasts 6 h apart but valid at the same time, similar to the “scaled lagged average forecasting” method (SLAF: Ebisuzaki and Kalnay, 1991; Hou *et al.*, 2001). Different perturbations are constructed for the different members by using increasingly long forecast lengths. The differences are scaled so that the perturbations for all members have the same magnitude. Secondly, the initial atmospheric state is perturbed for each member by adding perturbations calculated similarly to those for the boundaries to the control analysis. Finally, the initial surface uncertainty is modelled by randomly perturbing sea-surface temperature (SST), soil water and temperature, roughness and albedo following Bouttier *et al.* (2016). Each member runs to 54 h starting at 0000, 0600, 1200 and 1800 UTC and is delivered to the forecasters 2 h and 30 min after nominal time.

3 | SCALE-DEPENDENT PREDICTABILITY OF HOURLY PRECIPITATION ACCUMULATIONS IN MEPS

In this section, the predictability of precipitation for MEPS is explored. This is done in terms of the scale dependency of the predictability of patterns for hourly precipitation. Surcel *et al.* (2015) developed a method to assess the scale dependence of predictability from an ensemble forecast as “the loss of predictability of the model state.” It is based on the assumption that the forecast lead time at which the ensemble members become fully decorrelated for a particular spatial scale is the time at which there is no longer predictability at that spatial scale. They find that for hourly precipitation, averaged over 22 separate cases, the spatial scale at which predictability is lost follows a power law as a function of lead time; on average, the predictability of meso- γ scales (2–20 km) is lost after the first hour, and the predictability of meso- β scales (20–200 km) is lost at around 18 h lead time.

A common method to assess the spatial predictability and spatial skill of precipitation forecasts is the fractions skill score (FSS: Roberts and Lean, 2007). The method compares the spatial similarity between forecasted precipitation and observed precipitation (Roberts, 2008; Schwartz *et al.*,

2010; 2017) and/or the similarity between ensemble members (e.g. Dey *et al.*, 2014; 2016a; 2016b) for different thresholds. However, we choose to use the method of Surcel *et al.* (2015) to gauge the scale dependence of predictability for hourly precipitation forecasts from MEPS since this method aggregates all possible thresholds into a single measure. Sixty precipitation cases were chosen that occurred in Norway within the operational lifetime of MEPS (8 November 2016–28 February 2018, at the time of writing). To ensure a sufficient range of precipitation scales, and in particular theoretically less predictable small-scale heavy precipitation, events were chosen by ranking the 95th percentiles of observed hourly and daily precipitation accumulations and choosing the top 30 from each. Some of the top 30 hourly precipitation events occurred on the same day as each other, and some of the top 30 daily events had top 30 hourly events on the same day. This resulted in a total of 54 cases that can be classified as 18 1-hour precipitation events, 24 daily precipitation events and 12 cases that were both hourly and daily precipitation events.

For each event, the 0000 UTC operational forecast issued on that day is used in the predictability analysis. This is justified since in each case precipitation was well established in the model domain at initialisation time. Following Surcel *et al.* (2015), the scale-dependent variance of a precipitation field was obtained by computing the power spectrum from the discrete cosine transform (DCT: Denis *et al.*, 2002) of the field. To obtain the optimum contribution of each element of the DCT to the one-dimensional (1-d) wave numbers, the method of Ricard *et al.* (2013) was used to distribute power between the wave numbers. The power ratio of the ensemble was found for each spatial scale, λ , following Equation 1:

$$R(\lambda) = \frac{\sum_{i=1}^N P_{X_i}(\lambda)}{P_{X_T}(\lambda)}, \quad (1)$$

where $P_{X_i}(\lambda)$ is the variance of the field X for the i th ensemble member, $X_T = \sum_{i=1}^N X_i$, and N is the number of ensemble members. The largest spatial scale at which $R(\lambda) = 1$ is the scale at which the ensemble members become fully decorrelated as the covariance becomes zero.

The power ratio as a function of spatial scale for all 54 events is shown in Figure 2a for all lead times from 1 to 48 h. It is clear that the power ratio increases with lead time, but unlike Surcel *et al.* (2015) the power ratio does not reach a value of 1. This implies that the ensemble members do not become fully decorrelated over the forecast, suggesting that there is some inherent predictability for the smaller spatial scales. However, the horizontal part of the power ratio curves suggests that the level of decorrelation remains the same for all spatial scales shorter than ~ 100 km for lead times longer than ~ 10 h. There is also a small reduction in the power ratio for scales smaller than ~ 10 km. To get an impression of the scales at which the decorrelation does not reach the maximum, the power ratio curves were normalised to the 95th

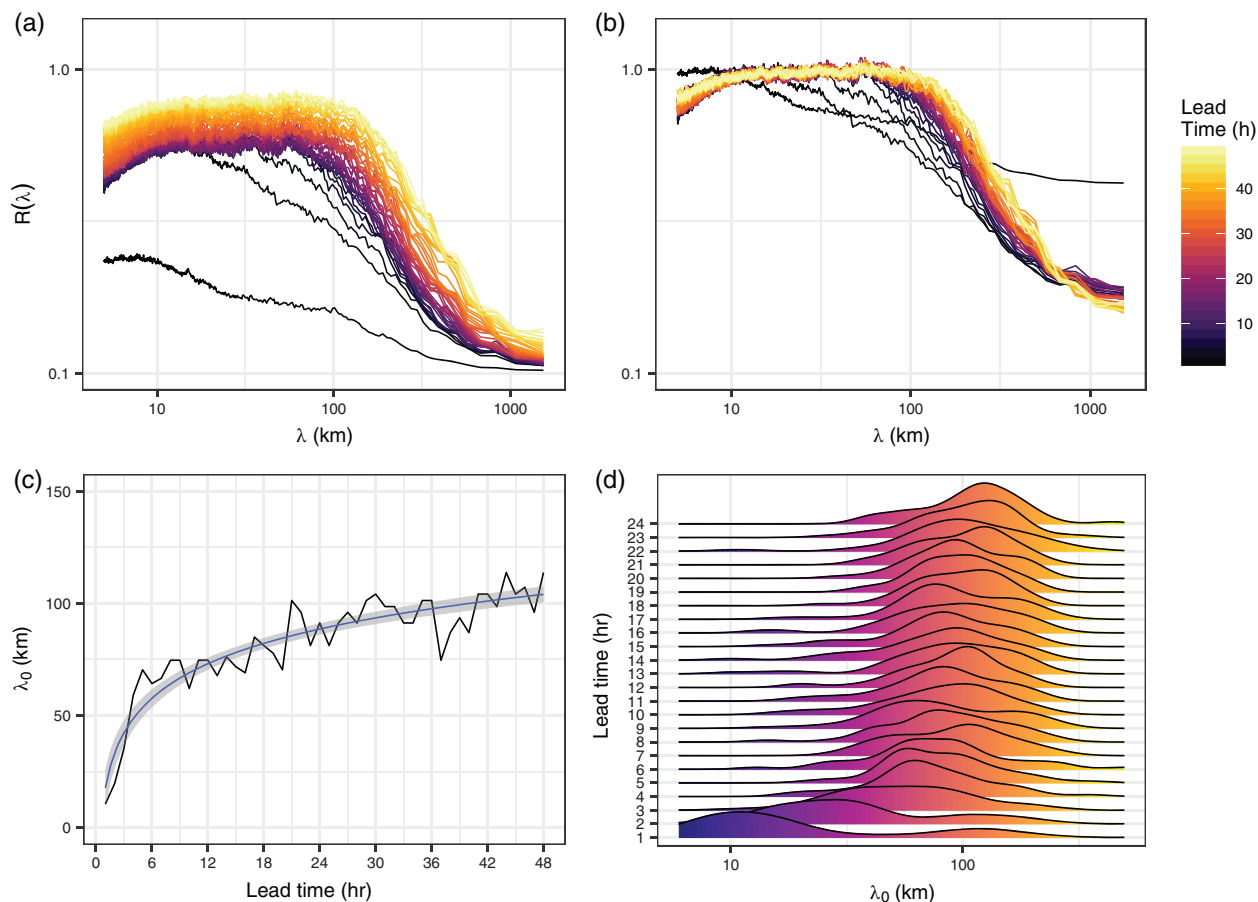


FIGURE 2 Scale-dependent predictability of hourly precipitation forecasts averaged over 54 MEPS cases. (a) The power ratio as a function of spatial scale (λ) for lead times from 1 (dark) to 48 h (light), (b) as (a) except the power ratio is normalised to the 95th percentile for each lead time, (c) the decorrelation length-scale as a function of lead time with the thick solid line showing a logarithmic fit with the 95% confidence interval of that fit shown in the grey shading, (d) the distribution (kernel fit) of decorrelation length-scales for each forecast for each lead time from 1 to 24 h – the shading is indicative of the length-scale with darker shades approximately representing meso- γ scales, becoming lighter for meso- β scales [Colour figure can be viewed at wileyonlinelibrary.com]

percentile¹ for each lead time (Figure 2b) and the largest spatial scales at which the normalised curves drop below 1 were identified (Figure 2c). These results suggest that there is a rapid loss of predictability in the first 5 h of the forecast for spatial scales below ~ 60 km and a slower loss of predictability for the larger scales for longer lead times. The loss of predictability with lead time has an approximate logarithmic relationship (Figure 2c), but there is, however, considerable variability between cases (Figure 2d).

The fact that the ensemble members do not become fully decorrelated suggests that the perturbations in MEPS may not be sufficient to completely model the uncertainty at scales below ~ 100 km. However, this lack of complete decorrelation may also be due to the combination of summer and winter events used in the analysis. Winter events are typically characterized by large-scale precipitation, which is inherently more predictable, interacting with coastal and orographic forcing mechanisms in the western and southern coastal regions of Norway. Such enhancement mechanisms mean that the small-scale structures are likely to be anchored to orographic

and coastal features in each ensemble member and thus the members do not become fully decorrelated. This would also go some way to explaining the drop in the power ratio for scales below ~ 10 km.

The scale-dependent predictability analysis was repeated for only the 18 1-hour precipitation events, most of which occurred in the summer, and only the 24 daily precipitation events, all of which occurred in the winter or autumn. The 1 h events were typically characterized by mesoscale systems coming from the south with strong embedded convection, while the daily events mostly consisted of synoptic-scale systems from the west that led to stationary precipitation with orographic enhancements over the western coastal regions of Norway. For the 1 h events, the power ratio is much closer to 1 for the smaller scales (Figure 3a) than for all 54 events (Figure 2a). Conversely, for the daily events (Figure 4a), the maximum power ratio is lower than that for all 54 events (Figure 2a). As was done for all 54 cases, the power ratios are normalised (Figures 3b and 4b) and the largest spatial scale at which the power ratio falls below 1 is identified for each lead time (Figures 3c and 4c). For the 1 h precipitation events, the spatial scales at which predictability is lost (Figure 3c) grows more slowly to ~ 50 km after 6 h compared with ~ 65 km after

¹The 95th percentile of each power ratio curve was set to the value of 1 and all other values are indexed relative to that point.

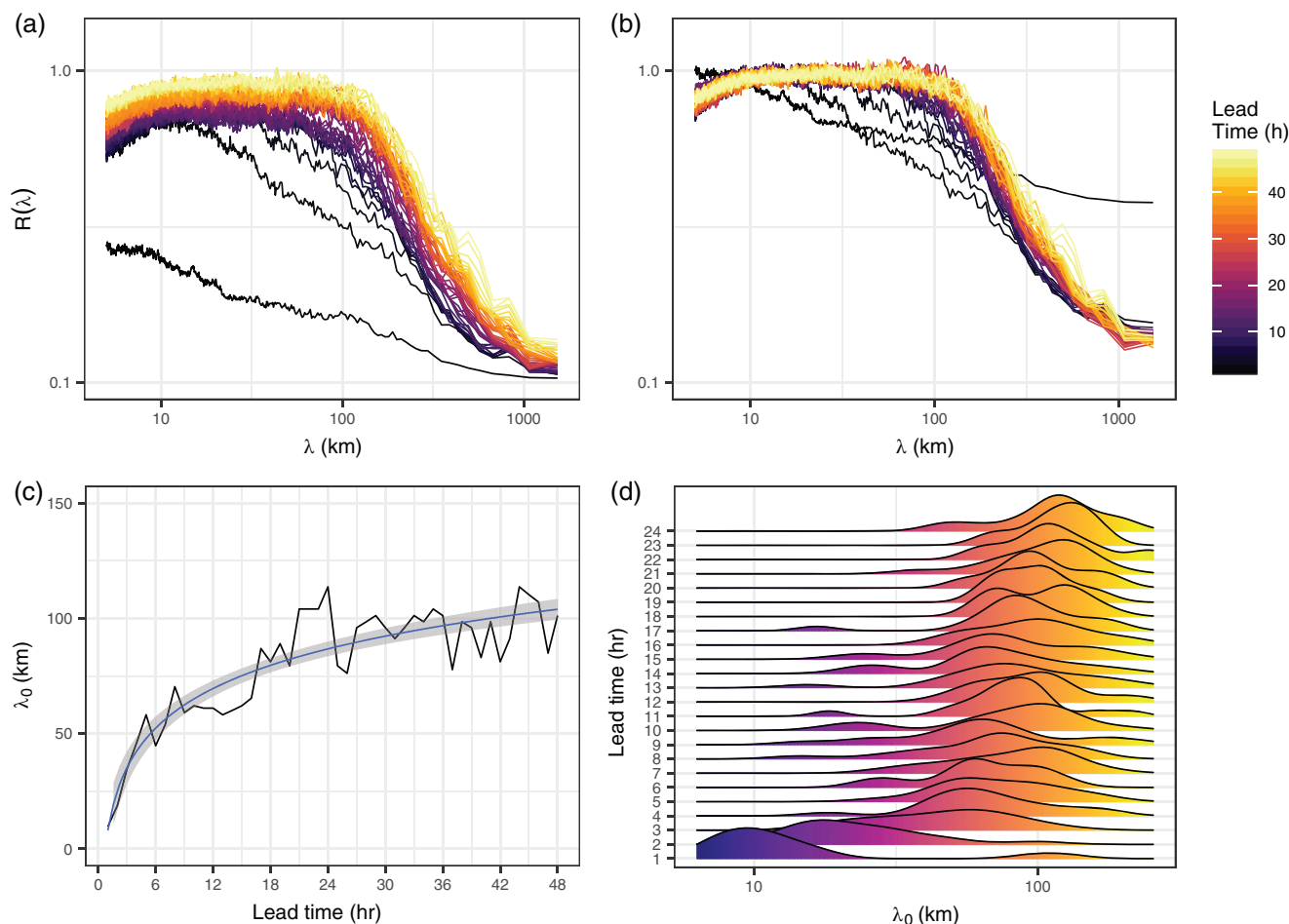


FIGURE 3 (a–d) As Figure 2 except for the 18 most extreme hourly precipitation cases [Colour figure can be viewed at wileyonlinelibrary.com]

6 h for the daily precipitation events (Figure 4c). Furthermore, the case-to-case variability is smaller for the 1 h precipitation events (Figure 3d, compare Figure 4d).

For MEPS, there is a clear degradation in the spatial scale of predictability in the first hours of the forecast, with scales smaller than ~ 65 km being unpredictable beyond 6 h lead time for daily precipitation events (Figure 4c). For 1 h precipitation events, spatial scales smaller than ~ 50 km have some predictability for up to 6 h lead time (Figure 3c). It should be noted that the method used here gives an indication of the spatial scales that have practical predictability for MEPS without giving any indication of the quality of the forecasts. Comparing these results with those of Surcel *et al.* (2015), the loss of predictability of the smaller scales in the first 5–6 h is similar for MEPS, but for longer lead times MEPS has predictability for scales around 100 km, compared with around 200 km in Surcel *et al.* (2015). This may be a result of MEPS being higher resolution than the experiments of Surcel *et al.* (2015) (2.5 km compared with 4 km). However, it should be noted that horizontal grid spacing is not the only difference – the forecasts considered in this article are also done using a different model over a different domain for different events compared to Surcel *et al.* (2015), so the results are not directly comparable. A comparison was also done with IFSSENS (not shown), but the lack of decorrelation between

the members was much more pronounced (power ratio < 0.5) in the first 24 h of the forecast, probably due to perturbations in IFSSENS being tuned to maximize error growth at longer lead times. Normalising the power ratios suggested a decorrelation length-scale of approximately 60 km for the first 24 h of the forecast, but the low power ratio makes it difficult to conclude that this is a true reflection of predictability and not simply a lack of spread in the ensemble.

4 | ADDED VALUE OF MEPS

4.1 | Added value compared to IFSSENS and deterministic forecasts

CPEPS should add value to a weather forecast compared with deterministic high-resolution runs (e.g. against its own control run) and against coarser resolution (global) EPSs. Partly, this is shown when comparing EPS systems using probabilistic verification scores. A traditional approach for evaluating a probabilistic system is to look at the spread, measured as the standard deviation of the ensemble members with respect to the ensemble mean, and skill, measured as the root-mean-square error (RMSE), of the system. Figure 5 shows a comparison of the spread and skill of IFSSENS and MEPS for a range of near-surface parameters. In both this and

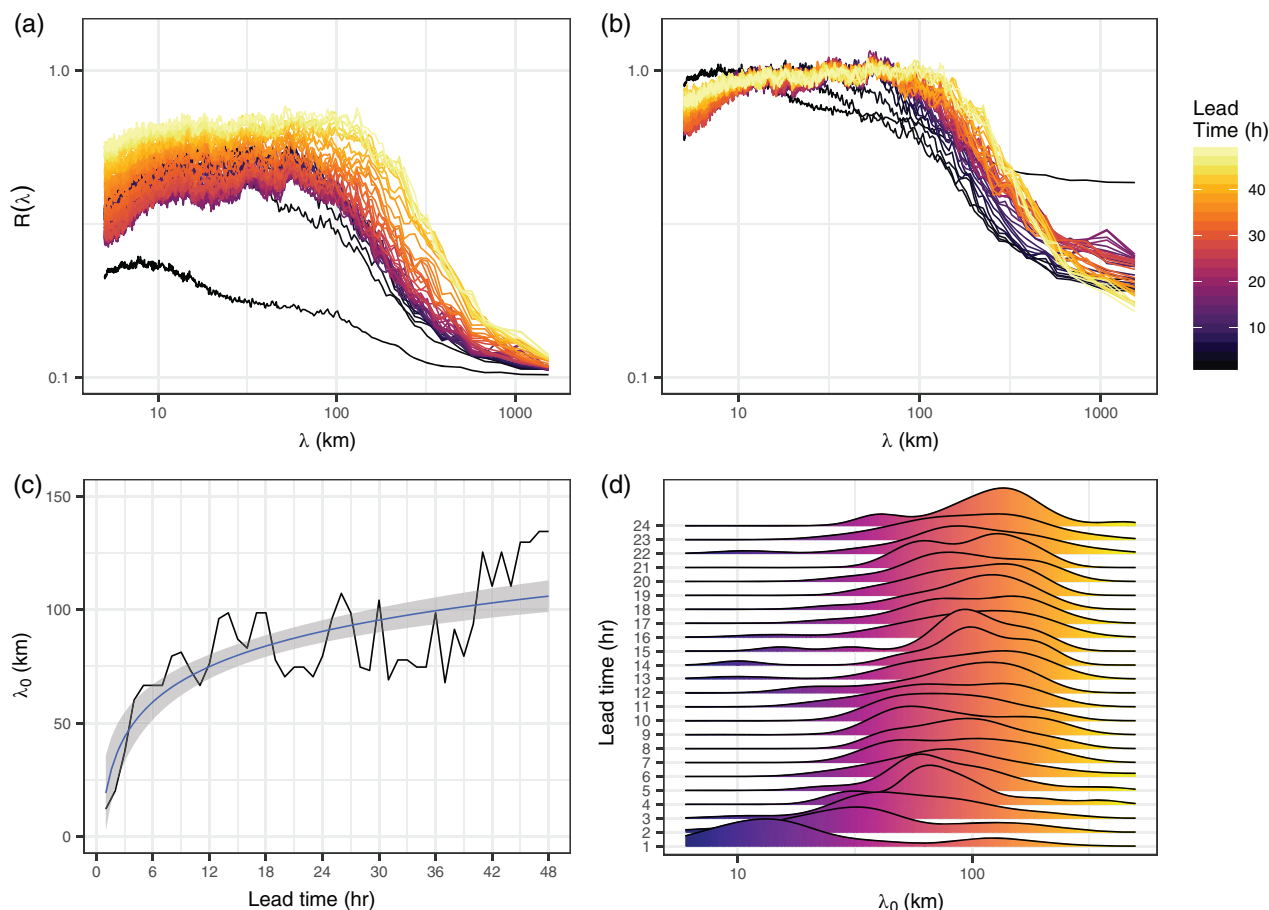


FIGURE 4 (a–d) As Figure 2 except for the 24 most extreme daily precipitation cases [Colour figure can be viewed at wileyonlinelibrary.com]

later figures, verification scores and probabilities are based on the full set of available members since the purpose is to compare currently available operational models. The total number of ensemble members is 51 for IFSENS, and 10 for MEPS. MEPS has higher spread and lower or equal RMSE compared to IFSENS, which gives a better spread–skill relationship.

Verification such as in Figure 5 is often what is presented to show the added value of regional systems. However, we want to investigate the added value of MEPS in more depth. A first criterion for added value is that there should be some differences between MEPS and the system it is supposed to add value to (hereafter referred to as the potential added value). In addition, this difference should bring MEPS forecasts closer to the observations, improve verification scores (Figure 5) and/or increase the value of the forecast for end-users (hereafter referred to as the added value). In the following, we investigate forecasts of precipitation/no-precipitation. For simplicity, probabilities are estimated by grid-point ensemble probability, hereafter EP (Schwartz and Sobash, 2017), based on the number of members with exceedance of the threshold in a given grid box. No further post-processing or calibration is applied if not mentioned. Further calibration, for example, neighbourhood methods, might therefore have an influence on the results presented below.

In Figure 6, probabilities for precipitation (defined as accumulated precipitation >0.2 mm in 12 h) based on IFSENS,

IFSENS control, MEPS and MEPS control during a one-year period are compared (upper panels). The probabilities from the deterministic control runs are binary (i.e. 0 or 100%). For a large portion of the time period, IFSENS and MEPS probabilities are similar to the probabilities obtained from their control runs. However, the potential added value increases from +18 h forecasts to +54 h forecast in accordance with a general increase in the uncertainty with lead time as reflected in a larger ensemble spread. The potential added value of MEPS versus MEPS control is larger than for IFSENS versus IFSENS control for the shorter lead times while it becomes smaller for +54 h. Comparing the probabilities from MEPS and IFSENS reveal that they often agree, but MEPS forecasts dry conditions more often than IFSENS. Convection-permitting models represent showers directly and will therefore have larger dry areas, whereas models with convection parametrization schemes represent the average effect over a wider area and will likely have fewer dry points. For +18 h forecasts MEPS forecasts lower probabilities than IFSENS $\sim 40\%$ of the time, higher probabilities $\sim 10\%$ of the time and the same probability $\sim 50\%$ of the time.

The middle panels of Figure 6 show time series of the daily potential added value. The smoothing of the time series hides a high day-to-day variability, most likely depending on the weather itself. Furthermore, the potential added value is largest between MEPS and IFSENS and between the two

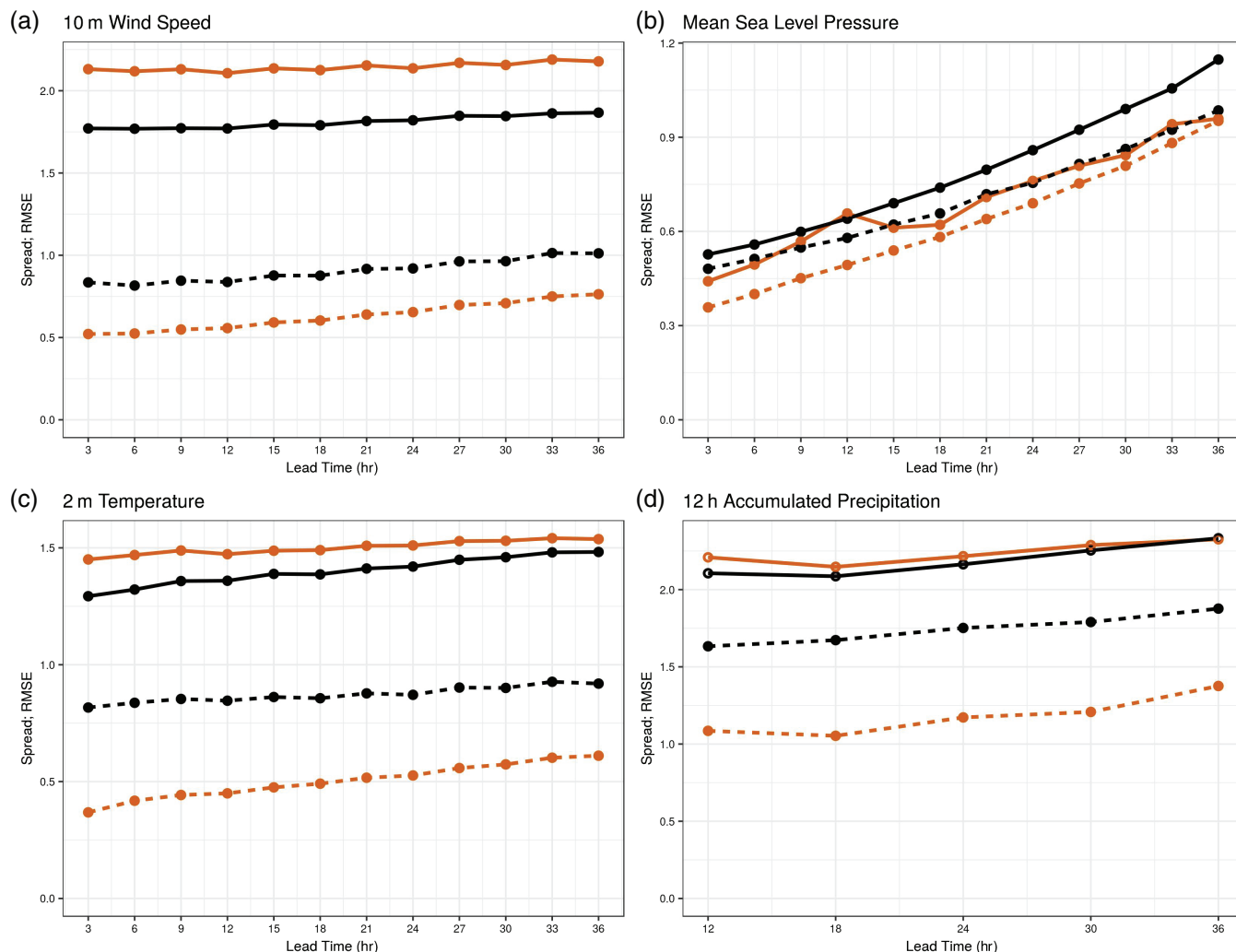


FIGURE 5 Spread (dashed) and skill (solid) for IFSENS (orange) and MEPS (black) for month in 2018. (a) S_{10m} , (b) MSLP, (c) T_{2m} and (d) 12 h accumulated precipitation. Closed circles indicate where differences between IFSENS and MEPS are statistically significant at the 95% confidence level. Statistical significance of the differences was calculated using a bootstrap method with 1,000 replicas sampled from the forecasts and observation data with replacement

control runs. This indicates that potential added value originates both from the resolution/model itself and from the ensemble compared with the control. In addition, there is a tendency for larger potential added value of MEPS and MEPS control in the warm season. This seems reasonable due to more small-scale precipitation. To investigate the added value further we use the Brier Skill Score (BSS). Instead of using for example a climatology as reference, we use the different model systems as a reference so the added value can easily be measured between model systems. In doing so we are far from discussing all aspects of the complex datasets, but BSS is related to decision making for (some) end-users that want to make binary decisions based on some probabilities from the forecast. In addition, BSS can be computed for both EPSs and deterministic runs.

The time series of BSS show that IFSENS and MEPS score better ($BSS > 0$) than their control runs, that is, the ensemble consistently adds value to its control run (lower panel Figure 6). Similarly, MEPS scores systematically better than IFSENS for the shortest lead times. However, for longer lead

times MEPS only adds value in the warm season. To a large extent, this seasonal and lead time dependency is also found when comparing the control runs only. It is therefore possible that this added value originates from the model system itself and maybe not necessarily be from the ensemble set-up. The score calculation was also repeated using 10 IFSENS members to account for the difference in ensemble members between the systems. However, this reduction in the number of members gave only a modest decrease in the skill of IFSENS in Figure 6 (not shown). In this comparison, the difference in ensemble size therefore contributes little to the differences seen between IFSENS and MEPS.

CPEPSs are specifically designed to target high precipitation events. To investigate this, BSS for high-precipitation events, defined by exceedance of the 99th percentile of the observed 12 h accumulated precipitation, are presented in Table 1. The numbers are given together with 95th percentile confidence intervals to indicate statistical significance. Similar averaged numbers with precipitation/no-precipitation are included for comparison and underline that the major features

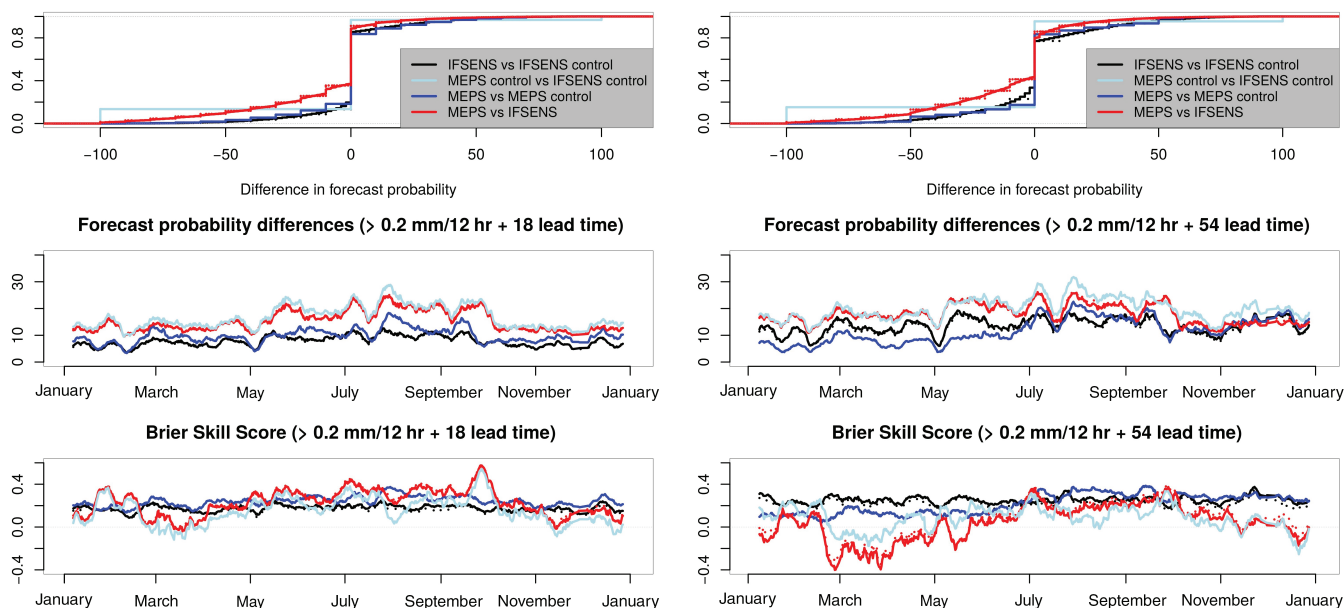


FIGURE 6 Examples of (potential) added value for +18 h (left) and +54 h (right) forecast for probability of precipitation > 0.2 mm in 12 h for all 170 Norwegian stations. First row show cumulative distribution of the difference in forecasted probabilities (potential added value) between IFSENS control and IFSENS ensemble (black), MEPS control and IFSENS control (light blue), MEPS control and MEPS ensemble (blue) and between MEPS ensemble and IFSENS ensemble (red). The second row show the same, but as a time series per day, while the third row show the Brier Skill Score of IFSENS ensemble with IFSENS control as reference forecast, MEPS ensemble with MEPS control as reference forecast, MEPS control with IFSENS control as reference forecast and MEPS ensemble with IFSENS ensemble as reference forecasts (real added value). A 20 days moving filter has been applied on the time series to remove some of the day to day variability

TABLE 1 Brier Skill Score (including 95th percentile confidence interval by bootstrapping) for exceedance of observed 99th percentile 12 h accumulated precipitation for cold (October–March precipitation > 17.8 mm/12 h) and warm (April–September precipitation > 21.2 mm/12 h) season

	Warm season + 18 h > 17.8 mm/12 h (0.2 mm/12 h)		Warm season + 54 h > 17.8 mm/12 h (0.2 mm/12 h)		Cold season + 18 h > 21.2 mm/12 h (0.2 mm/12 h)		Cold season + 54 h > 21.2 mm/12 h (0.2 mm/12 h)	
<i>MEPS control versus IFSENS control</i>	-0.04 ^{+0.05} _{-0.14}	(+0.26 ^{+0.27} _{+0.24})	-0.02 ^{+0.06} _{-0.11}	(+0.17 ^{+0.18} _{+0.16})	+0.13 ^{+0.20} _{+0.03}	(+0.12 ^{+0.13} _{+0.10})	+0.10 ^{+0.17} _{+0.02}	(+0.07 ^{+0.09} _{+0.06})
<i>IFSENS versus IFSENS control</i>	+0.27 ^{+0.30} _{+0.23}	(+0.18 ^{+0.19} _{+0.18})	+0.33 ^{+0.36} _{+0.29}	(+0.25 ^{+0.26} _{+0.25})	+0.25 ^{+0.29} _{+0.21}	(+0.17 ^{+0.18} _{+0.17})	+0.34 ^{+0.38} _{+0.31}	(+0.27 ^{+0.28} _{+0.26})
<i>MEPS versus MEPS control</i>	+0.36 ^{+0.40} _{+0.32}	(+0.28 ^{+0.29} _{+0.28})	+0.31 ^{+0.34} _{+0.27}	(+0.25 ^{+0.26} _{+0.24})	+0.25 ^{+0.29} _{+0.21}	(+0.22 ^{+0.23} _{+0.22})	+0.29 ^{+0.32} _{+0.25}	(+0.21 ^{+0.22} _{+0.20})
<i>MEPS versus IFSENS</i>	+0.10 ^{+0.15} _{+0.04}	(+0.35 ^{+0.35} _{+0.34})	-0.05 ^{+0.01} _{-0.12}	(+0.16 ^{+0.18} _{+0.15})	+0.12 ^{+0.19} _{+0.05}	(+0.17 ^{+0.18} _{+0.16})	+0.03 ^{+0.09} _{-0.04}	(+0.00 ^{+0.01} _{-0.01})

For comparison the numbers for precipitation/no-precipitation (>0.2 mm/12 h) are included in parentheses. Positive values indicated added value of the first system compared over the second system (i.e. MEPS add value to MEPS control). Approximately 57,000 cases (forecast – observation pairs) are used per season and forecast length.

discussed in Figure 6 are statistically significant. MEPS also adds value compared to IFSENS for high-precipitation thresholds, but only for the shorter lead times. However, the added value is less than for precipitation/no-precipitation. No added value is also seen when comparing the control runs from MEPS and IFSENS for high-precipitation events in summer. Most likely, the use of only raw grid-point probabilities penalizes MEPS (control) and highlights the need for more MEPS members and/or post-processing making use of neighbourhood information. Another common pattern is that the added value of both IFSENS and MEPS compared with their control runs are higher for high-precipitation thresholds than for precipitation/no-precipitation. Using 10 IFSENS members again modifies the numbers slightly, but does not change the conclusions (not shown).

In the following, we present two severe precipitation cases to further discuss the potential added value and challenges of

CPEPSs. The first case is a summer severe flash flood, and the second case is a large-scale precipitation weather event in the cold season. Based on the numbers in Table 1 it is believed that MEPS should add value in both cases.

4.2 | Flash flood 6 August 2016

On the morning of 6 August 2016 a weak low-pressure system was located between Norway and Denmark (Figure 7a). Warm air was present to the south and east, and cold air to the north and west. High humidity at low levels and cold air aloft reduced the vertical stability of the system and heavy convective precipitation developed early in the morning and moved slowly eastward. The most intense precipitation lasted for about 12 h. Estimated precipitation amounts based on radar reflectivity are shown in Figure 7b and give a good spatial picture of the precipitation, but underestimate the rain-gauge

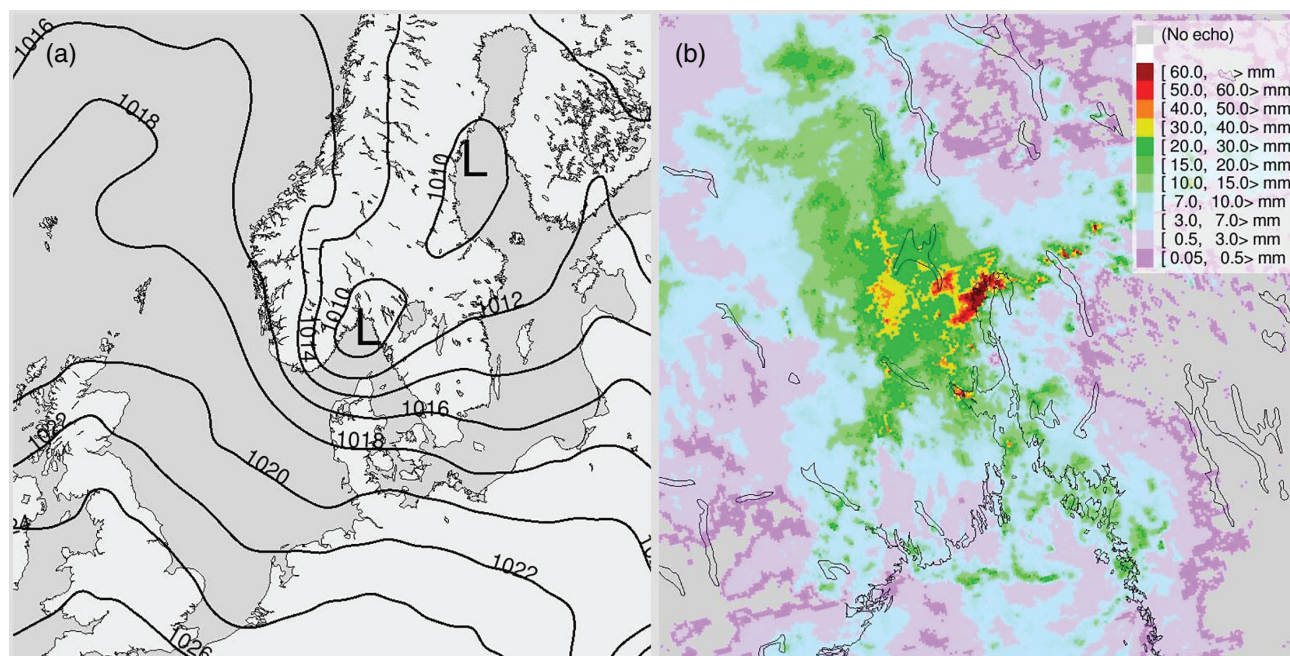


FIGURE 7 (a) Mean-sea-level pressure from IFS HRES (ECMWF) analysis valid at 0600 UTC 6 August 2016, and (b) estimated 12 h accumulated precipitation based on radar reflectivity in the region close to the low-pressure centre [Colour figure can be viewed at wileyonlinelibrary.com]

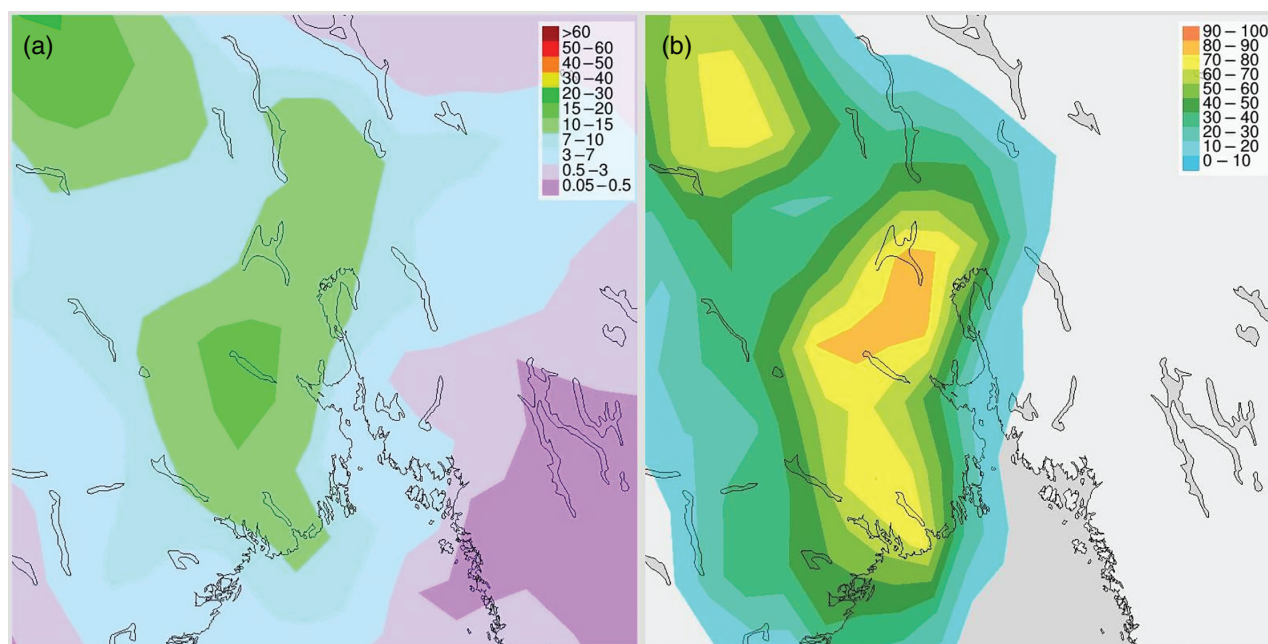


FIGURE 8 (a) Forecasts with +24 h lead time of 12 h accumulated precipitation from IFSENS control and (b) IFSENS grid-point ensemble probability >10 mm in 12 h [Colour figure can be viewed at wileyonlinelibrary.com]

observations. Several sites measured more than 30 mm in 1 h, more than 50 mm in 12 h and several private measurements suggested precipitation above 100 mm in 24 h.

In the following we compare IFSENS (all members), IFSENS control, MEPS (all members) and MEPS control forecasts for 12 h accumulated precipitation. In August 2016 MEPS had pre-operational status with four daily runs and nine ensemble members with lead times up to +36 h (control run up to +66 h). In IFSENS, the forecasts were consistent with time from approximately 4 days ahead, but with a more pronounced signal for the shorter lead times

(not shown). However, the forecasts underestimated the precipitation amounts. IFSENS control indicated 10–20 mm, probabilities for more than 20 mm were hardly seen (not shown), while the probability of more than 10 mm was high west of the low-pressure centre (Figure 8).

MEPS control had already forecasted the large-scale precipitation signature of the event at +66 h, and indicated more than 50 mm locally (100 mm in smaller regions) in all forecasts with lead times of +60 h and shorter (Figure 9). The challenge for the operational forecasters was to determine the maximum precipitation amount, in which region

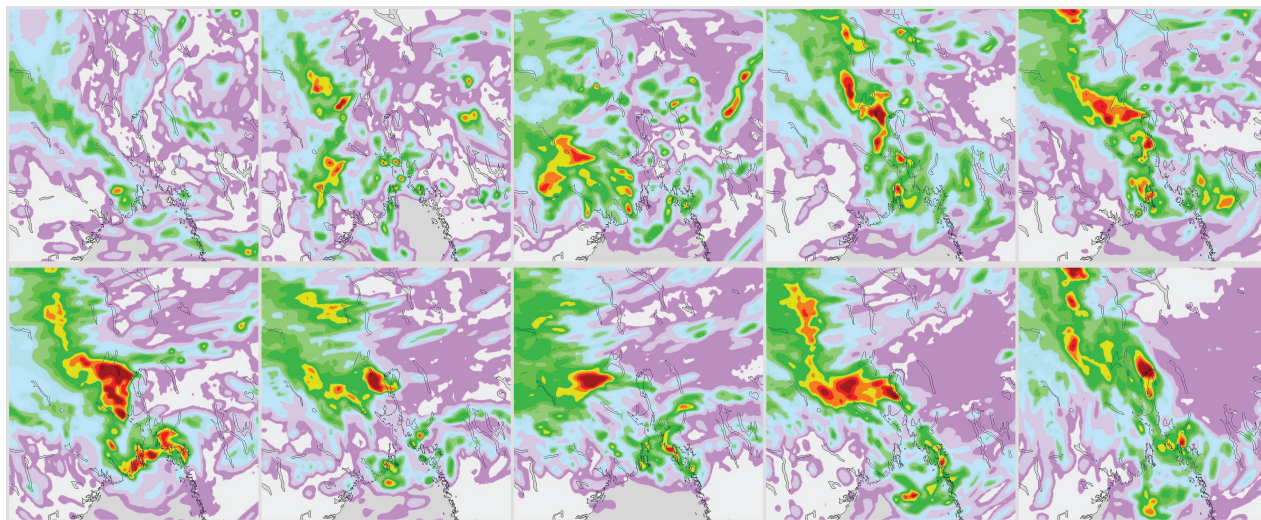


FIGURE 9 MEPS control 12 h accumulated precipitation, with lead times of +66, +60, +54, +48, +42, +36, +30, +24, +18 and +12 h starting with +66 h in upper left corner and ending with +12 h in lower right corner. Legend same as Figures 7b and 8a [Colour figure can be viewed at wileyonlinelibrary.com]

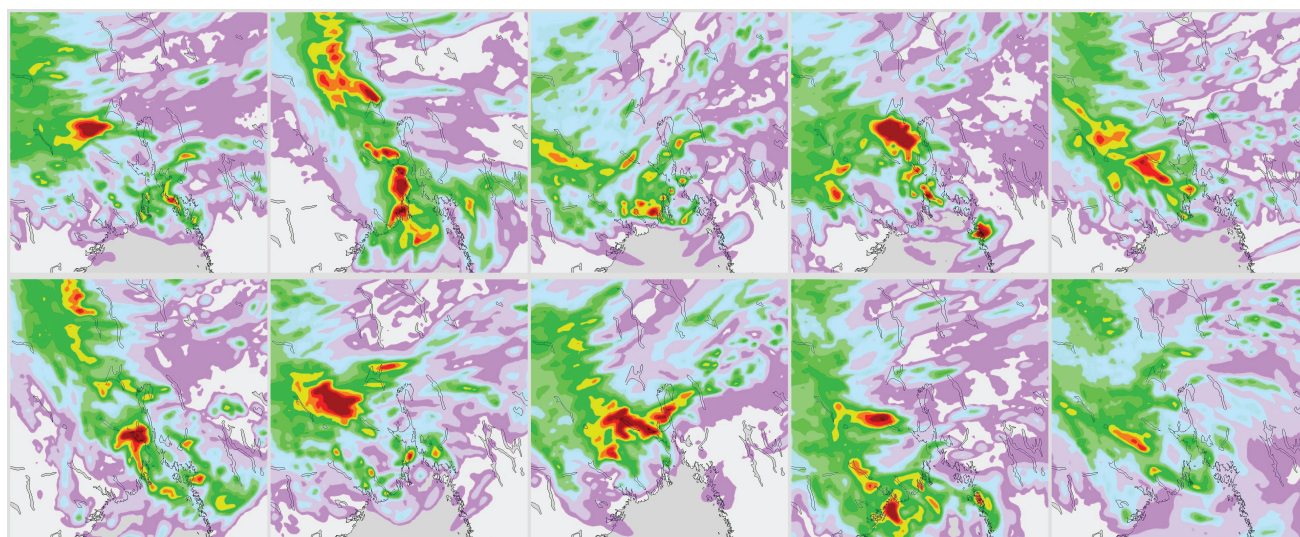


FIGURE 10 MEPS members 12 h accumulated precipitation, with lead time +24 h. Control run (upper left) and then member 1, 2, 3, 4 on upper row, and 5, 6, 7, 8, 9 on lower row. Legend same as Figures 7b and 8a [Colour figure can be viewed at wileyonlinelibrary.com]

and in what accumulation period. At +36 h ahead, when the full MEPS ensemble became available, all members indicated similar development as MEPS control, but with the local precipitation maxima at different locations (not shown). In Figure 10, MEPS forecasts with +24 h lead time for all members are shown. Locally, all members forecast more than 50 mm in smaller areas, but with a large spread in the location of the most intense precipitation. This additional information from all members makes the expectations of severe precipitation amounts between 50 and 100 mm more confident. However, it can be argued that this was already confirmed by the consistency between MEPS control updates and that a simpler time-lagged ensemble would have worked well in this case. In addition, the exact locations and total areas of severe precipitation are still challenging to forecast.

In section 3 it was shown that the spatial scales for which precipitation is unpredictable grow rapidly with forecast time.

Therefore, in order to extract more information, we compare the grid-point ensemble probability (EP, defined in the previous section) with a neighbourhood maximum probability approach (NMP). The latter is defined as the frequency of members where at least 2% of a neighbourhood area experiences an exceedance of some threshold (Schwartz and Sobash, 2017). The first probability gives the probability of exceedance of the threshold at a specific point (grid cell), while the latter gives the probability of exceedance somewhere in the vicinity of a point. These probabilities therefore serve different user requirements (i.e. exceedance in single location or somewhere in an area).

The appropriate choice of neighbourhood area varies with for example, precipitation type, topographic forcing, precipitation threshold, lead time and model configuration. Based on our experience with MEPS control we chose 15×15 grid cells or $37.5 \text{ km} \times 37.5 \text{ km}$. The choice of minimum 2% area in the calculations is slightly arbitrary, but we want

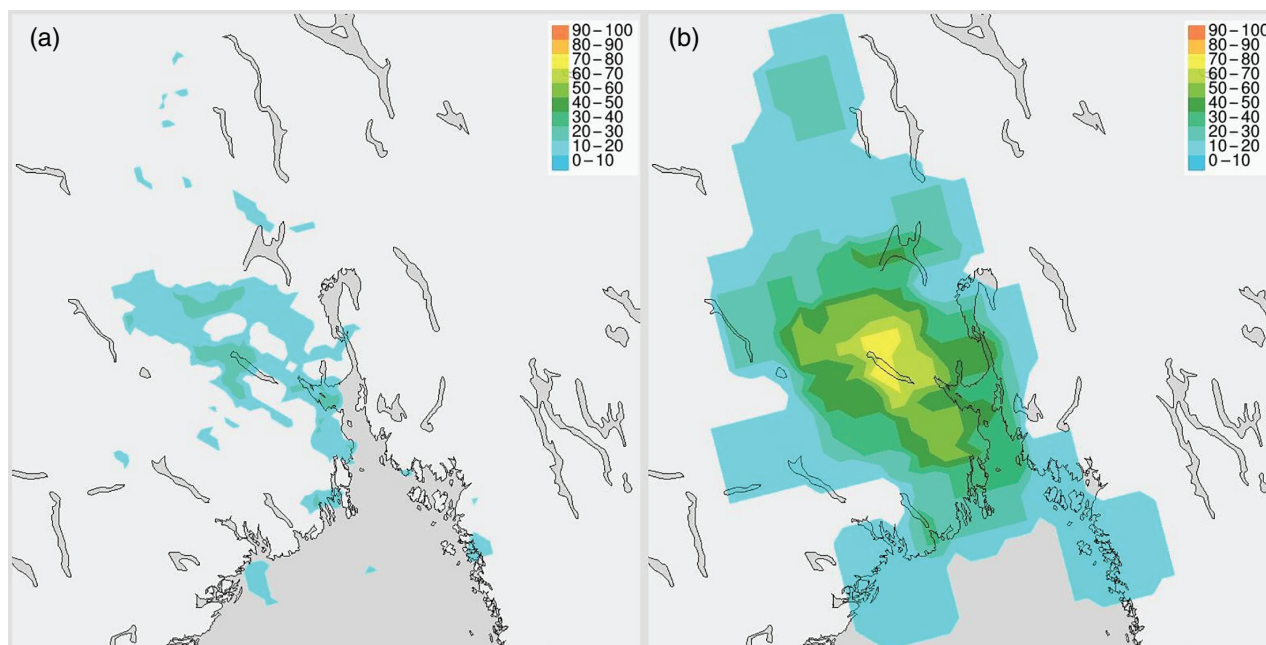


FIGURE 11 Probabilities based on MEPS for more than 50 mm in 12 h calculated by probability (a) in one grid cell, EP, and (b) within a neighbourhood area, NMP. Both forecasts with lead time 24 h [Colour figure can be viewed at wileyonlinelibrary.com]

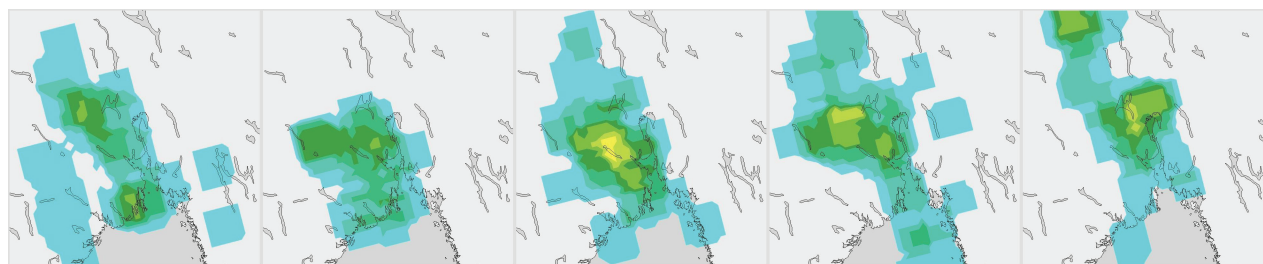


FIGURE 12 Same as Figure 11b, but for lead times from left to right: +36, +30, +24, +18 and +12 h. Legend same as Figure 11 [Colour figure can be viewed at wileyonlinelibrary.com]

to look at the extreme precipitation amounts, but exclude potential single “grid-point storms.” Again we focus on the exceedance of 50 mm in 12 h which is appropriate to this particular event. In general, the appropriate threshold depends on for example, user requirements, model and observed climatology.

Figure 11 shows the calculated probabilities by EP and NMP. The EP shows low probabilities for severe precipitation. However, these probabilities might hide important information in a warning perspective. Opposite to this the NMP approach shows much higher probabilities for severe precipitation. The difference arises from the high uncertainty in location since all members indicate severe precipitation (Figure 10).

The NMP signal of severe precipitation is consistent for all lead times (Figure 12). Notice that for the early forecasts (+36, +30 and +24 h) the maximum probability is slightly west of the low-pressure centre. In later forecasts (+18 and +12 h) the maximum probabilities move slightly eastward in accordance with what was observed (Figure 7b).

For this event, the control run of MEPS and its members made the operational forecasters more confident in the

expectations of severe precipitation amounts. However, it was still challenging to pinpoint areas at risk, but the forecasts with shortest lead times (+18/+12 h) highlighted the areas with the most intense observed precipitation. Other severe convective situations have resulted in similar experiences. To optimize the added value of CPEPS it might be useful to rethink how weather warnings are issued and distributed to end-users in these types of situation.

4.3 | Large-scale precipitation; extreme weather Birk 2017

On 22 and 23 December 2017, a deep low-pressure system developed northeast of Iceland and moved eastwards over north Norway, Sweden and Finland (Figure 13a). South of the system a southwesterly flow with warm and humid air masses resulted in large precipitation amounts with rain as high up as 1,500 m above sea level at the Norwegian west coast. Based on the expected precipitation amounts in combination with snow melt, an extreme weather warning, named Birk, was issued by MET-Norway. Observations show that accumulations of more than 100, 70 and 50 mm in 24 h were

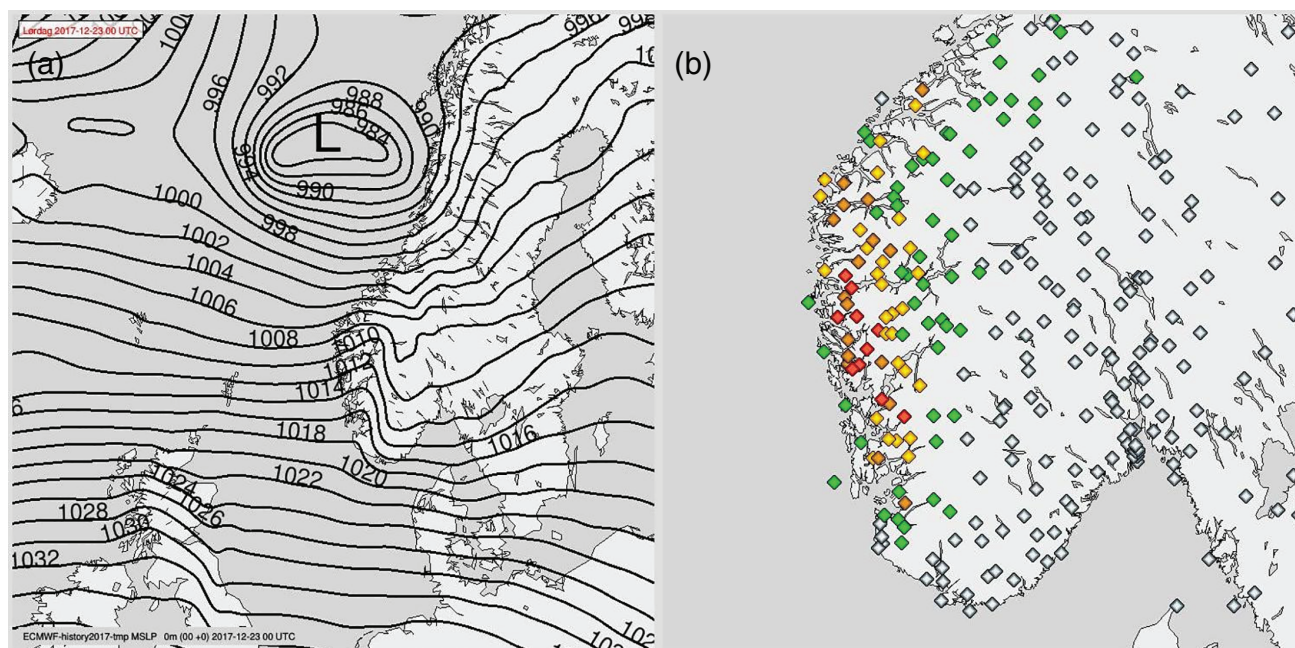


FIGURE 13 (a) Mean-sea-level pressure from IFS HRES (ECMWF) analysis valid at 0000 UTC 23 December 2017, and (b) observations of 24 h accumulated precipitation between 0600 UTC 22 December and 0600 UTC 23 December. Red, orange, yellow and green squares are more than 100, 70, 50 and 10 mm in 24 h, respectively

observed at 10, 27 and 56 stations, respectively, covering a large part of the Norwegian west coast (Figure 13b).

Both IFSSENS and MEPS indicated snow melting conditions in areas with severe precipitation (not shown). The potential added value of MEPS was therefore limited to forecasts of precipitation. The potential for anomalous precipitation (and warm air) was indicated several days ahead by the Extreme Forecast Index (EFI: Lalaurette, 2003) and Shift of Tails (SOT: Tsonevsky and Richardson, 2012) products from ECMWF. Despite being in the high end of its own climatology, IFSSENS underestimated the observed precipitation amounts with forecasts of more than 50 mm in 24 h (lead time +30 h), but most likely less than 70 mm (Figure 14a,b). However, the spatial coverage of the precipitation was well forecasted.

As in the convective case above, MEPS forecasted higher precipitation amounts than IFSSENS. Forecast updates for MEPS control for +66 h and shorter are very consistent, but both MEPS and IFSSENS forecasted higher precipitation amounts for shorter lead times (+54 and +30 h forecasts from MEPS are shown in Figure 15). However, compared to the observations in Figure 13b, MEPS control also underestimated the precipitation amounts.

All MEPS members were available for +54 h lead time and shorter. For this event we only show the EP probabilities. The NMP approach that was useful in the convective case will in this situation basically cover large areas with little spatial information.

MEPS produces high probabilities for severe precipitation (Figure 15) with increasing probabilities with decreasing lead times. Figure 15 also shows a rapid transition from probabilities of either 0 or 100% indicating little

spread among the MEPS members. A key question is if this behaviour indicates a highly predictable situation or an underestimation of the uncertainty by MEPS. Most likely it is a combination since IFSSENS shows a similar behaviour and the updates of the MEPS forecasts are quite consistent.

We should also note that the MEPS probabilities inherit the systematic error of the HARMONIE-AROME system, for example, precipitation release occurring too late over ocean/coastal areas with an onshore wind component resulting in precipitation maxima too far inland. Such model errors might be a larger challenge for CPEPSs than for global EPSs since systematic errors have a tendency to manifest in weather variables that are the target of CPEPSs and not necessarily in the large-scale variables that are more targeted by global EPSs.

For this large-scale precipitation event the finer-resolution model system produces higher precipitation maxima. However, it is more doubtful if the ensemble of fine-resolution members in itself adds significant value, but that might depend on the use and user needs. In Figure 16 we zoom in to the catchment scale to get more small-scale details. MEPS control gives a more realistic picture of the spatial variability than IFSSENS control. However, further investigation is needed to see how MEPS adds value (e.g. by forcing hydrology models) on the small-scale variability in such situations.

5 | DISCUSSION AND CONCLUSIONS

Challenges related to limited predictability and use of CPEPSs have been discussed in this article, illustrated by the operational CPEPS for Norway, Sweden and Finland,

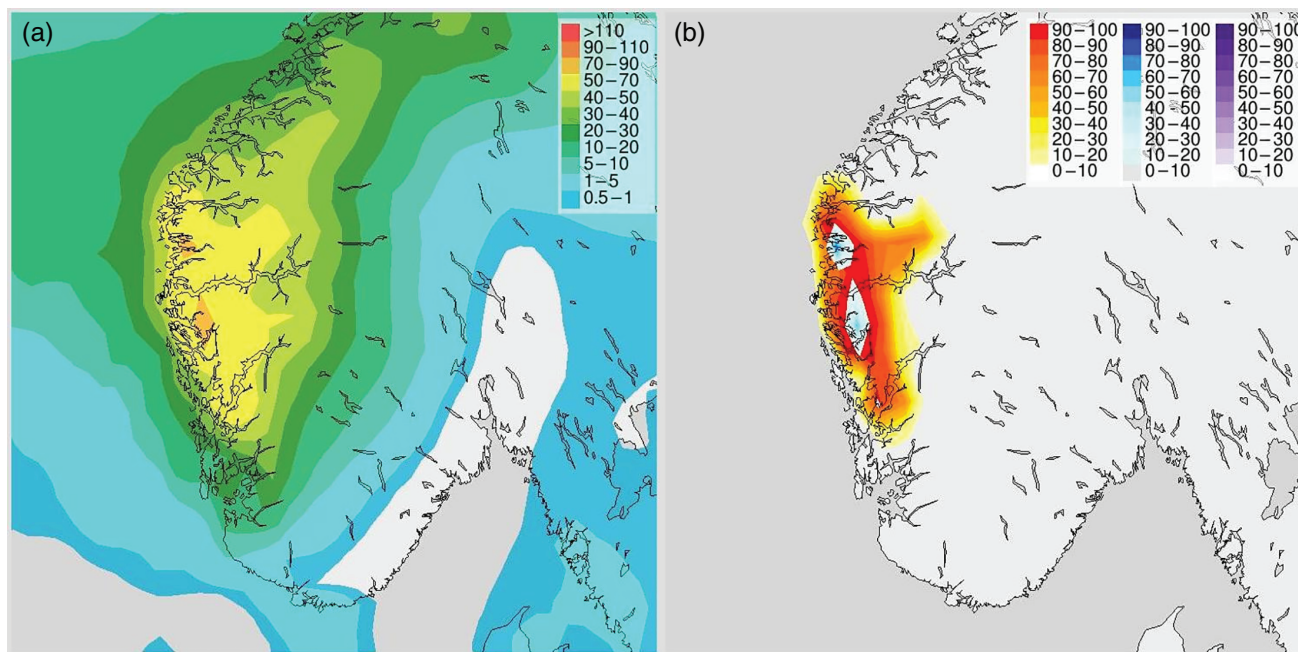


FIGURE 14 (a,b) IFSENS forecasts initialized at 0000 UTC 22 December 2017 and valid for the 24 h accumulation period from 0600 UTC 22 December–0600 UTC 23 December 2017. Panel (a) shows 24 h accumulated precipitation from control run, panel (b) shows probabilities of more than 50 mm in 24 h (yellow/red), more than 70 mm in 24 h (blue) and more than 100 mm in 24 h (purple). All probabilities calculated by fraction of members in a grid cell, EP

MEPS. The predictability of precipitation was investigated for hourly extremes and daily extremes as a function of spatial scale and lead time using a spectral decomposition approach in section 3. However, such an analysis does not say anything about the quality of a forecast compared with what was actually observed. In section 4, the added value of using ensemble forecasts opposed to deterministic forecasts as well as using a high-resolution ensemble compared with a low-resolution ensemble was investigated for precipitation and assessed in relation to observed precipitation. Both the average behaviour of the models over a one-year period, and the behaviour of the models for two very distinct extreme precipitation events were assessed. Ideally the unpredictable model scales should be identical to the unpredictable atmospheric scales, or in other words, the practical predictability should be the same as the intrinsic predictability. In practice this is not possible. Relatively large errors in the initial conditions, errors in lateral boundary conditions (LBCs) and model error will result in error growth that gives rise to limited practical predictability. The study presented in this article can be used in working towards getting the practical predictability closer to the intrinsic predictability by: (a) Designing better perturbations for the initial conditions, model and LBCs, as it was suggested by the scale-dependent predictability diagnosis in section 3 that there is a weakness in the design of MEPS. The scale-dependent predictability diagnosis highlights what scales are affected by the proposed perturbations and for what scales and forecast lengths predictability is improved or not. (b) Introducing new observation types, or analysis algorithms, and study to what extent it results in a decreased error of the initial conditions which in turn gives rise to higher predictability on smaller scales.

The analysis in section 3 was complicated by the fact that, on average, the ensemble members in MEPS do not become fully decorrelated for spatial scales below ~ 100 km, but instead reach a maximum level of decorrelation. This maximum level of decorrelation increases with lead time. For spatial scales smaller than ~ 10 km, there is a small drop in the power ratio implying a larger correlation between the members at these scales. This increased correlation can be explained by the influence of orographic forcing anchoring rainfall to local topographic features as discussed in Simmons (2003). For spatial scales between ~ 10 and ~ 100 km, the absence of full decorrelation is a little more difficult to explain. It is possible that there is some inherent predictability at these scales. Separating the events used in the analysis into 1-hour and daily precipitation events supports this explanation to some extent. The MEPS members for the daily events were shown to have a lower power ratio for spatial scales between ~ 10 and ~ 100 km, meaning that the ensemble members for the daily events are more correlated than those for the 1-hour events. The daily events were dominated by large-scale precipitation that one would expect to be more predictable than the 1-hour events that were more convective. The two case-studies in sections 4.2 and 4.3 clearly illustrate this behaviour. For the warm season convective event, small-scale precipitation features varied considerably in location giving low probabilities for extremes at grid-point scales, but high probabilities when larger neighbourhoods were considered. For the cold season large-scale precipitation case-study, probabilities for extremes shows a rapid transition from 0 to 100%, indicating little variability for many spatial scales. Furthermore, the analysis for the whole year in section 4.1 showed that, on average, MEPS adds more value over deterministic

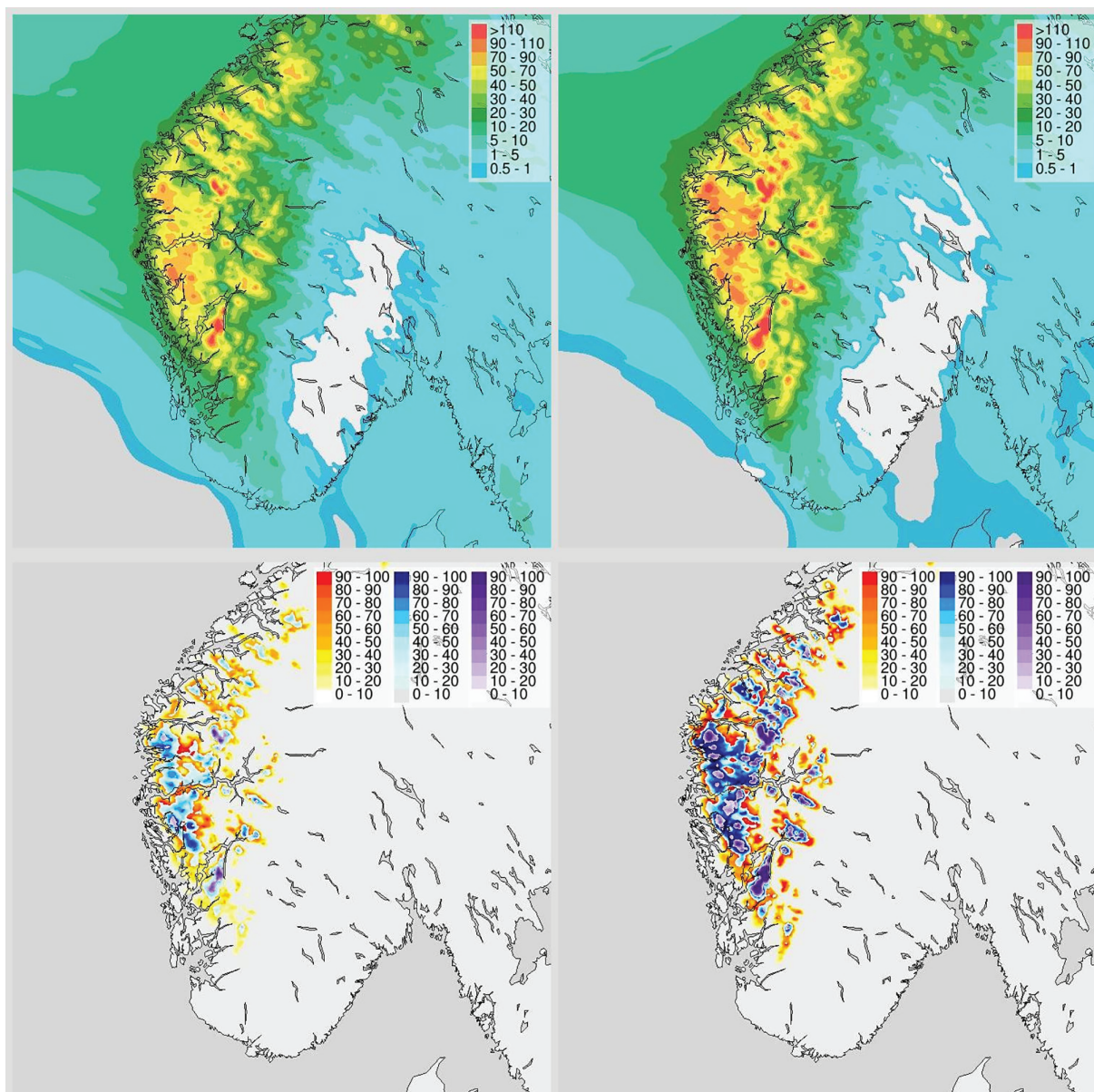


FIGURE 15 MEPS control 24 h accumulated precipitation (0600 UTC 22 December–0600 UTC 23 December 2017) for lead times +54 h (left) and +30 h (right) in the upper row. Probabilities from MEPS for more than 50 mm in 24 h (yellow/red colours), 70 mm in 24 h (blue colours) and 100 mm in 24 h (purple colours) for lead times +54 h (left) and +30 h (right) in lower row

forecasts for summer precipitation events than for winter. This is likely related to the same lack of variability in the ensemble during large-scale precipitation events.

The inability to reach full decorrelation suggests that the perturbations used in MEPS are not sufficient to fully model the uncertainty in the forecasts. A key difference between MEPS and the experiments of Surcel *et al.* (2015) is that MEPS only takes into account uncertainty in the initial conditions and lateral boundaries via the SLAF method, and at the surface by stochastic perturbations, whereas Surcel *et al.* (2015) also take model uncertainty into account by using different combinations of microphysics, planetary boundary-layer and radiation parametrization schemes. Further work would be needed to understand whether taking the model uncertainty into account is the missing ingredient for MEPS and, at the time of writing, research is ongoing

to incorporate model uncertainty into MEPS using methods such as Stochastically Perturbed Physics Tendencies (SPPT: Buizza *et al.*, 1999) and Stochastic Parameter Perturbations (SPP: Ollinaho *et al.*, 2017).

Despite the rapid loss of predictability for precipitation in MEPS as shown in section 3, we see that it has real added value over both its own control and over IFSSENS, except for the latter in winter for long lead times. The added value of MEPS is largest in summer, when predictability is lower. MEPS compared to MEPS control also has higher added value than IFSSENS compared to IFSSENS control in most cases, so both resolution/model and EPS contributes to the added value of MEPS over IFSSENS. From the winter scores for MEPS for +54 h (Figure 6) it seems that a better option would be to use IFSSENS for these lead times. This, and also the variability of the scale dependence of predictability from

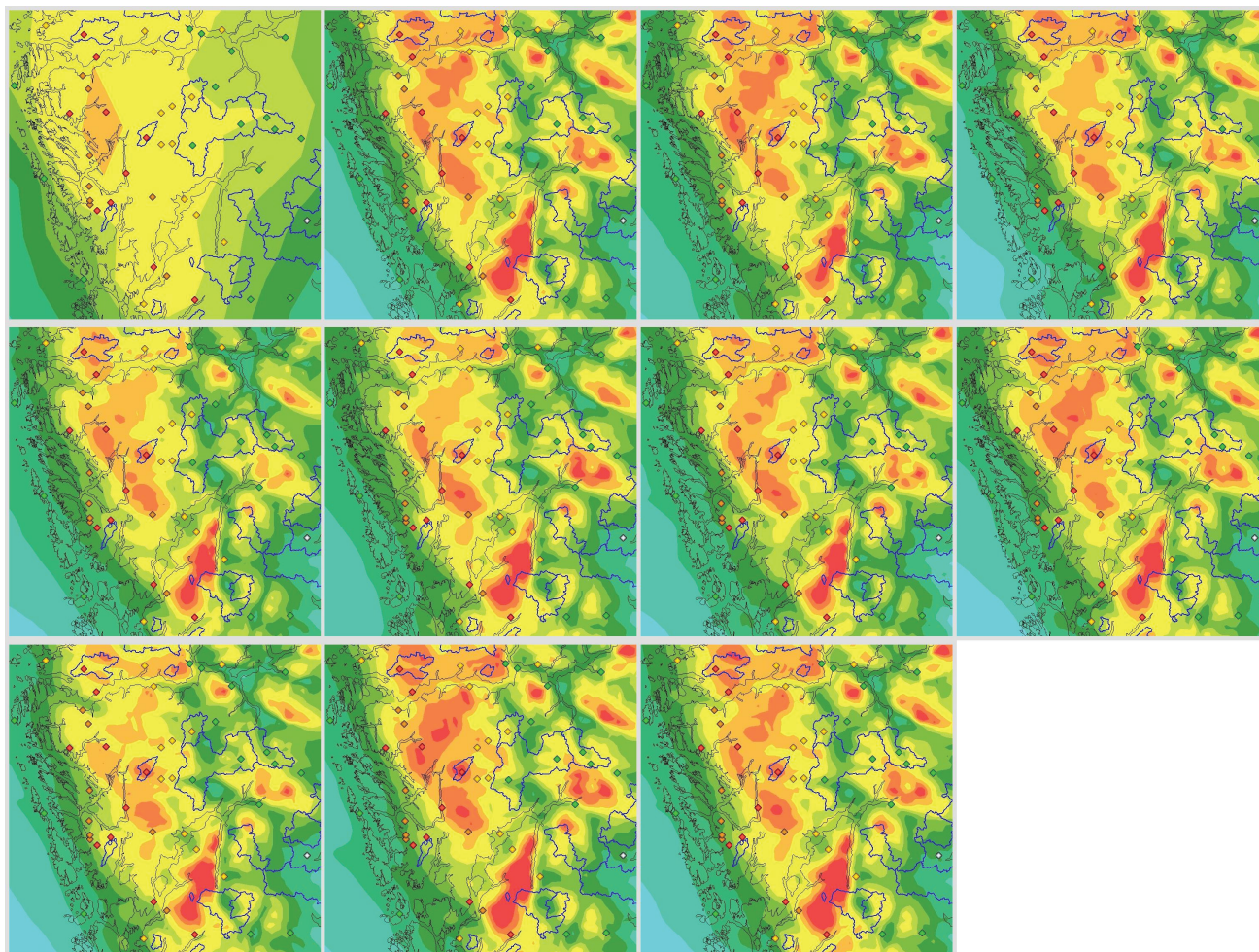


FIGURE 16 Regional-scale forecasts from IFSENS control (upper left) and MEPS members for 24 h accumulated precipitation. Blue colours indicate separate catchments. Same legend as in Figure 14a and observations from Figure 13b

case to case and from season to season, points to a complexity both for the design and use of CPEPS: its value depends on season, lead time and most likely also parameter. For designing a CPEPS it means that the impact of the perturbations should be assessed for different seasons. Furthermore, the lead time at which a global EPS becomes more useful than a CPEPS will vary depending on season and the forecast parameter of interest.

It was illustrated in section 4 that even though MEPS scores well overall and better than IFSENS, it is not straightforward to use from case to case. The two cases discussed in sections 4.2 and 4.3 are very different, both in terms of scale and in terms of the added value by MEPS. The summer flash flood event was a case where there was clear added value of MEPS, while for the large-scale winter case it did not really add much over its control. In predictable situations, this is what is to be expected. However, there are cases where small-scale uncertainty exists in large-scale situations and where MEPS has shown to add value when zooming in over, for example, a catchment area.

It was shown in Hagelin *et al.* (2017) and Raynaud and Bouttier (2017) that more members are more beneficial than higher resolution; however, one can easily imagine very

localized weather connected with complex topography that will benefit from higher resolution. High-resolution weather forecasts are more capable of forecasting small-scale features, but as discussed in the introduction, there is little skill at those scales and even at current CPEPS resolution predictability of small scales is rapidly lost as shown in section 3. An ensemble of infinite size would be needed to cover all possible distributions of the small-scale information. This is obviously not achievable, hence post-processing techniques should be an integrated part of any EPS. It was shown in section 4 that neighbourhood methods can be a valuable tool in assessing information from CPEPS. There are also more sophisticated neighbourhood methods that can be used. The so-called smart neighbourhood method uses high-resolution datasets such as topography and land–sea mask, and interpolates to the nearest neighbour with the same characteristics instead of the closest neighbour. Another example is downscaling using a reference (Seierstad *et al.*, 2016), to achieve higher resolution; one could, for example, run HARMONIE-AROME with, for example, 750 m grid spacing for a period and then downscale MEPS from 2.5 km to 750 m assuming that HARMONIE-AROME 750 m represents the reality.

To summarize the main findings in this study:

- 1 The inability of the MEPS members to become fully decorrelated suggests a weakness in the design of the ensemble, and further work is needed in order to understand the reasons for this and to incorporate more sources of uncertainty in all aspects of the system. The scale-dependent predictability can be used as a diagnostic in this work.
- 2 Despite the limited predictability of the smaller scales, MEPS shows added value over both its deterministic control and over IFSSENS, and the value of MEPS is largest in summer, when predictability is lower, than in winter. It is interesting to note that this is the case even though the number of members is far smaller in MEPS (9 + 1) than in IFSSENS (50 + 1).
- 3 Both the increased resolution/different model and EPS contribute to the added value of MEPS over IFSSENS.
- 4 The value of precipitation forecasts from MEPS depends on season, lead time and threshold.
- 5 With limited predictability of small scales, post-processing should be an integrated part of any system.

ACKNOWLEDGEMENTS

This article is based on a talk at the ECMWF annual seminar in 2017: Ensemble prediction: past, present and future. We wish to thank the organisers of this interesting seminar and for arranging the Special Supplement of the Q. J. R. Meteorol. Soc. on 25 years of ensemble prediction. We are grateful to the two anonymous reviewers for valuable comments and helping us getting the manuscript more focused.

ORCID

Inger-Lise Frogner  <https://orcid.org/0000-0002-3446-9873>

REFERENCES

- Bengtsson, L., Andrae, U., Aspelien, T., Batrak, Y., Calvo, J., de Rooy, W., Gleeson, E., Hansen-Sass, B., Homleid, M., Hortal, M., Ivarsson, K., Lenderink, G., Niemelä, S., Nielsen, K.P., Onville, J., Rontu, L., Samuelsson, P., Muñoz, D.S., Subias, A., Tjmm, S., Toll, V., Yang, X. and Költzow, M.Ø. (2017) The HARMONIE-AROME model configuration in the ALADIN-HIRLAM NWP system. *Monthly Weather Review*, 145, 1919–1935. <https://doi.org/10.1175/MWR-D-16-0417.1>.
- Boer, G.J. (2003) Predictability as a function of scale. *Atmosphere–Ocean*, 41(3), 203–215. <https://doi.org/10.3137/ao.410302>.
- Bouttier, F., Raynaud, L., Nuissier, O. and Ménétrier, B. (2016) Sensitivity of the AROME ensemble to initial and surface perturbations during HyMeX. *Quarterly Journal of the Royal Meteorological Society*, 142, 390–403. <https://doi.org/10.1002/qj.2622>.
- Bouttier, F., Vié, B., Nuissier, O. and Raynaud, L. (2012) Impact of stochastic physics in a convection-permitting ensemble. *Monthly Weather Review*, 140, 3706–3721. <https://doi.org/10.1175/mwr-d-12-00031.1>.
- Buizza, R., Miller, M. and Palmer, T.N. (1999) Stochastic representation of model uncertainties in the ECMWF Ensemble Prediction System. *Quarterly Journal of the Royal Meteorological Society*, 125, 2887–2908. <https://doi.org/10.1002/qj.49712556006>.
- Denis, B., Côté, J. and Laprise, R. (2002) Spectral decomposition of two-dimensional atmospheric fields on limited-area domains using the discrete cosine transform (DCT). *Monthly Weather Review*, 130, 1812–1829. [https://doi.org/10.1175/1520-0493\(2002\)130<1812:SDOTDA>2.0.CO;2](https://doi.org/10.1175/1520-0493(2002)130<1812:SDOTDA>2.0.CO;2).
- Dey, S.R., Leoncini, G., Roberts, N.M., Plant, R.S. and Migliorini, S. (2014) A spatial view of ensemble spread in convection-permitting ensembles. *Monthly Weather Review*, 142, 4091–4107. <https://doi.org/10.1175/MWR-D-14-00172.1>.
- Dey, S.R., Plant, R.S., Roberts, N.M. and Migliorini, S. (2016b) Assessing spatial precipitation uncertainties in a convective-scale ensemble. *Quarterly Journal of the Royal Meteorological Society*, 142, 2935–2948. <https://doi.org/10.1002/qj.2893>.
- Dey, S.R., Roberts, N.M., Plant, R.S. and Migliorini, S. (2016a) A new method for the characterization and verification of local spatial predictability for convective-scale ensembles. *Quarterly Journal of the Royal Meteorological Society*, 142, 1982–1996. <https://doi.org/10.1002/qj.2792>.
- Ebisuzaki, W. and Kalnay, E. (1991) *Ensemble experiments with a new lagged average forecasting scheme*. WMO Research Activities in Atmospheric and Oceanic Modelling. Report 15, 6.31–6.32. Geneva, Switzerland: WMO.
- Frogner, I.-L., Nipen, T., Singleton, A., Bremnes, J.B. and Vignes, O. (2016) Ensemble prediction with different spatial resolutions for the 2014 Sochi winter Olympic games: the effects of calibration and multimodel approaches. *Weather and Forecasting*, 31, 1833–1851. <https://doi.org/10.1175/WAF-D-16-0048.1>.
- Gebhardt, C., Theis, S.E., Krahe, P. and Renner, V. (2008) Experimental ensemble forecasts of precipitation based on a convection-resolving model. *Atmospheric Science Letters*, 9, 67–72. <https://doi.org/10.1002/asl.177>.
- Giard, D. and Bazile, E. (2000) Implementation of a new assimilation scheme for soil and surface variables in a global NWP model. *Monthly Weather Review*, 128, 997–1015.
- Hagelin, S., Son, J., Swinbank, R., McCabe, A., Roberts, N.M. and Tennant, W. (2017) The Met Office convective-scale ensemble, MOGREPS-UK. *Quarterly Journal of the Royal Meteorological Society*, 143, 2846–2861. <https://doi.org/10.1002/qj.3135>.
- Hou, D., Kalnay, E. and Droegemeier, K.K. (2001) Objective verification of the SAMEX'98 ensemble forecasts. *Monthly Weather Review*, 129, 73–91. [https://doi.org/10.1175/1520-0493\(2001\)129<0073:OVOTSE>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<0073:OVOTSE>2.0.CO;2).
- Jung, T. and Leutbecher, M. (2008) Scale-dependent verification of ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 134, 973–984. <https://doi.org/10.1002/qj.255>.
- Lalauette, F. (2003) Early detection of abnormal weather conditions using a probabilistic extreme forecast index. *Quarterly Journal of the Royal Meteorological Society*, 129, 3037–3057. <https://doi.org/10.1256/qj.02.152>.
- Lorenz, E.N. (1969) The predictability of a flow which possesses many scales of motion. *Tellus*, 21, 289–307. <https://doi.org/10.3402/tellusa.v21i3.10086>.
- Müller, M., Homleid, M., Ivarsson, K.-I., Költzow, M.A.Ø., Lindskog, M., Midtbø, K.H., Andrae, U., Aspelien, T., Berggren, L., Bjørge, D., Dahlgren, P., Kristiansen, J., Randriamampianina, R., Ridal, M. and Vignes, O. (2017) AROME-MetCoOp: a Nordic convective-scale operational weather prediction model. *Weather and Forecasting*, 32, 609–627. <https://doi.org/10.1175/WAF-D-16-0099.1>.
- Ollinaho, P., Lock, S.-J., Leutbecher, M., Bechtold, P., Beljaars, A.C.M., Bozzo, A., Forbes, R.M., Haiden, T., Hogan, R.J. and Sandu, I. (2017) Towards process-level representation of model uncertainties: stochastically perturbed parametrizations in the ECMWF ensemble. *Quarterly Journal of the Royal Meteorological Society*, 143, 408–422. <https://doi.org/10.1002/qj.2931>.
- Raynaud, L. and Bouttier, F. (2017) The impact of horizontal resolution and ensemble size for convective-scale probabilistic forecasts. *Quarterly Journal of the Royal Meteorological Society*, 143, 3037–3047. <https://doi.org/10.1002/qj.3159>.
- Ricard, D., Lac, C., Riette, S., Legrand, R. and Mary, A. (2013) Kinetic energy spectra characteristics of two convection-permitting limited-area models AROME and Meso-NH. *Quarterly Journal of the Royal Meteorological Society*, 139, 1327–1341. <https://doi.org/10.1002/qj.2025>.
- Roberts, N.M. (2008) Assessing the spatial and temporal variation in the skill of precipitation forecasts from an NWP model. *Meteorological Applications*, 15, 163–169. <https://doi.org/10.1002/met.57>.
- Roberts, N.M. and Lean, H.W. (2007) Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, 136, 78–97.
- Romine, G.S., Schwartz, C.S., Berner, J., Fossell, K.R., Snyder, C., Anderson, J.L. and Weisman, M.L. (2014) Representing forecast error in a

- convection-permitting ensemble system. *Monthly Weather Review*, 142, 4519–4541. <https://doi.org/10.1175/MWR-D-14-00100.1>.
- Schwartz, C.S., Kain, J.S., Weiss, S.J., Xue, M., Bright, D.R., Kong, F., Thomas, K.W., Levit, J.J., Coniglio, M.C. and Wandishin, M.S. (2010) Toward improved convection-allowing ensembles: model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Weather and Forecasting*, 25, 263–280. <https://doi.org/10.1175/2009WAF2222267.1>.
- Schwartz, C.S., Romine, G.S., Fossell, K.R., Sobash, R.A. and Weisman, M.L. (2017) Toward 1-km ensemble forecasts over large domains. *Monthly Weather Review*, 145, 2943–2969. <https://doi.org/10.1175/MWR-D-16-0410.1>.
- Schwartz, C.S., Romine, G.S., Sobash, R.A., Fossell, K.R. and Weisman, M.L. (2015) NCAR's experimental real-time convection-allowing ensemble prediction system. *Weather and Forecasting*, 30, 1645–1654. <https://doi.org/10.1175/WAF-D-15-0103.1>.
- Schwartz, C.S. and Sobash, R.A. (2017) Generating probabilistic forecasts from convection-allowing ensembles using neighborhood approaches: a review and recommendations. *Monthly Weather Review*, 145, 3397–3418. <https://doi.org/10.1175/MWR-D-16-0400.1>.
- Seierstad I, Kristiansen J and Nipen T. (2016) Better temperature forecasts along the Norwegian coast. *ECMWF Newsletter 148*, p. 6. <https://www.ecmwf.int/en/newsletter/148/news/better-temperature-forecasts-along-norwegian-coast>.
- Simmons, A. (2003) Observations, assimilation and the improvement of global weather prediction – some results from operational forecasting and ERA-40. *Seminar on Recent Developments in Data Assimilation for Atmosphere and Ocean, 8–12 September 2003, Conference Paper*. ECMWF.
- Surcel, M., Zawadzki, I. and Yau, M.K. (2015) A study on the scale dependence of the predictability of precipitation patterns. *Journal of the Atmospheric Sciences*, 72, 216–235. <https://doi.org/10.1175/JAS-D-14-0071.1>.
- Termonia, P., Fischer, C., Bazile, E., Bouyssel, F., Brožková, R., Bénard, P., Bochenek, B., Degrauwe, D., Derková, M., El Khatib, R., Hamdi, R., Mašek, J., Pottier, P., Pristov, N., Seity, Y., Smolíková, P., Španiel, O., Tudor, M., Wang, Y., Wittmann, C. and Joly, A. (2018) The ALADIN system and its canonical model configurations AROME CY41T1 and ALARO CY40T1. *Geoscientific Model Development*, 11, 257–281. <https://doi.org/10.5194/gmd-11-257-2018>.
- Tsonevsky, I. and Richardson, D. (2012) Application of the new EFI products to a case of early snowfall in central Europe. *ECMWF Newsletter 133*, p. 4.

How to cite this article: Frogner I-L, Singleton AT, Køltzow MØ, Andrae U. Convection-permitting ensembles: Challenges related to their design and use. *Q J R Meteorol Soc.* 2019;145 (Suppl. 1):90–106. <https://doi.org/10.1002/qj.3525>