BAMS Article

Adopting Citizen Observations in Operational Weather Prediction

Thomas N. Nipen, Ivar A. Seierstad, Cristian Lussana, Jørn Kristiansen, and Øystein Hov

ABSTRACT: Citizen weather stations are rapidly increasing in prevalence and are becoming an emerging source of weather information. These low-cost consumer-grade devices provide observations in real time and form parts of dense networks that capture high-resolution meteorological information. Despite these benefits, their adoption into operational weather prediction systems has been slow. However, MET Norway recently introduced observations from Netatmo's network of weather stations in the postprocessing of near-surface temperature forecasts for Scandinavia, Finland, and the Baltic countries. The observations are used to continually correct errors in the weather model output caused by unresolved features such as cold pools, inversions, urban heat islands, and an intricate coastline. Corrected forecasts are issued every hour. Integrating citizen observations into operational systems comes with a number of challenges. First, operational systems must be robust and therefore rely on strict quality control procedures to filter out unreliable measurements. Second, postprocessing methods must be selected and tuned to make use of the high-resolution data that at times can contain conflicting information. Central to resolving these challenges is the need to use the massive redundancy of citizen observations, with up to dozens of observations per square kilometer, and treating the data source as a network rather than a collection of individual stations. We present our experiences with introducing citizen observations into the operational production chain of automated public weather forecasts. Their inclusion shows a clear improvement to the accuracy of short-term temperature forecasts, especially in areas where existing professional stations are sparse.

AFFILIATIONS: Nipen, Seierstad, Lussana, Kristiansen, and Hov—Norwegian Meteorological Institute, Oslo, Norway

https://doi.org/10.1175/BAMS-D-18-0237.1

Corresponding author: Thomas N. Nipen, thomasn@met.no

In final form 19 September 2019

The prevalence of citizen weather stations has increased dramatically in recent years. There is now a well-established market for low-cost, off-the-shelf, easy-to-use devices that are owned and maintained by private individuals. Many devices are connected to the Internet, providing weather measurements in real time, which some manufacturers collect and store centrally and make available in various ways also in near–real time.

Weather forecast providers invest in research and development to provide more localized forecasts, thereby enhancing the usefulness and value of the forecasts. Whereas forecasts used to be available only for major cities, today many providers offer forecasts on the neighborhood scale. As forecasts no longer represent large areas, forecasts errors are more obvious and cannot be attributed to regional variation in the weather. To obtain satisfactory accuracy on this finescale is a research challenge.

The basis for localized forecasts comes from high-resolution numerical weather prediction (NWP) models. Many operational models have sufficient resolution to potentially resolve the variability found on the neighborhood scale. Still, these models often exhibit biases caused by unresolved topography, coastlines, and insufficient physics.

Citizen observations have long been argued to be a data source that potentially could benefit a wide variety of weather applications, including operational weather forecasting (Muller et al. 2015). Measurements from a dense network of citizen weather stations have been shown to improve weather forecasts by using them in the initialization of weather models (Madaus et al. 2014; Gasperoni et al. 2018). Mass and Madaus (2014) further argued that pressure measurements from smart phones, due to their high network density, could revolutionize weather forecasting by improving the initialization of weather models, which in turn would produce forecasts that more accurately resolve frontal and convective weather systems.

Weather observations are also important in the postprocessing of NWP output, where the aim is to reduce model biases, thereby ensuring that forecasts better match the locally observed weather. The Norwegian Meteorological Institute (MET Norway) uses a state-of-the-art NWP model at 2.5-km grid resolution (Müller et al. 2017; Frogner et al. 2019), and have traditionally adjusted the modeled temperatures on the fly using observations from roughly 200 professional weather stations nationally, owned and operated largely by MET Norway. Despite this, temperature forecast errors of 10°C are not unheard of in wintertime inversion conditions in areas where the observation network is sparse and adjustments are not possible. MET Norway and many other forecast providers would benefit from a denser network of stations to patch the holes that exist in the postprocessed field.

The demand for high-quality localized forecasts has risen with the use of and experience gained with our weather service Yr (www.yr.no), which has as many as 10 million unique users worldwide per week. Yr is developed and operated jointly by MET Norway and the Norwegian Broadcasting Corporation. Feedback from users has provided valuable support and direction for the R&D postprocessing efforts. In response to this, MET Norway introduced observations from Netatmo, a manufacturer of private weather stations, in the operational production of public temperature forecasts on Yr in March 2018 for locations in Norway, Denmark, Sweden, Finland, Estonia, Latvia, and Lithuania. Our users expect a nowcast that accurately reflects the current weather conditions and a short-range forecast that is up to date with the latest information integrated. The introduction of Netatmo observations has helped us improve both aspects, and the products are reissued every hour. We share lessons learned from integrating this unconventional data source into an operational weather forecast service that is relied upon by millions of users. This article focuses exclusively on improving temperature forecasts, but our vision is that unconventional observations will be integrated into a greater part of our automated forecast systems in the future.

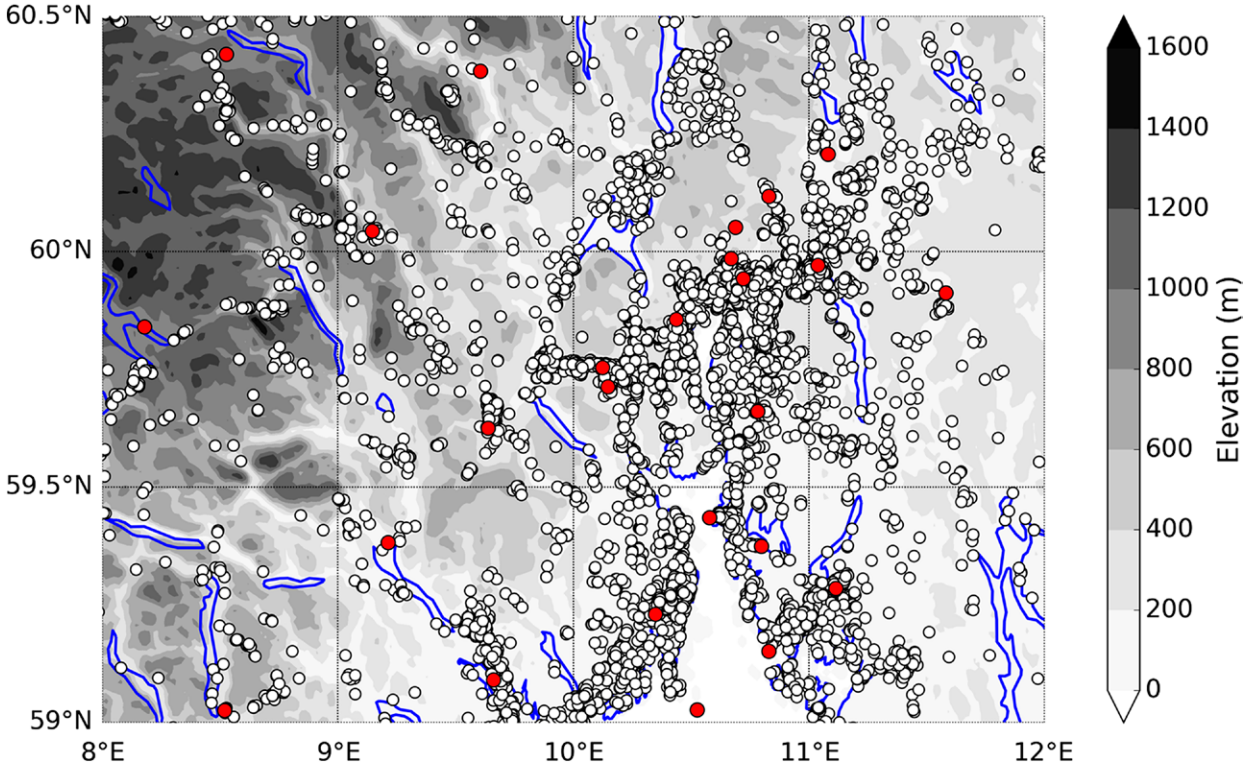

Fig. 1. Comparison of MET Norway's network of WMO-compliant stations (red circles) and Netatmo's network of citizen weather stations (white circles) in southeastern Norway.

Citizen observations

Citizen observations do not follow the rules and standards developed by the World Meteorological Organization (WMO) for weather observations (WMO 2014). For example, citizen stations can be located in backyards, on balconies, near walls, in direct sunlight, or even indoors. Poorly located instruments will not provide measurements that are representative of the nearby area, which is the purpose of the WMO standards for locating instruments. Additionally, citizen weather stations typically lack metadata about their location and how they are maintained. The lack of this information means that we cannot a priori determine the error characteristics of measurements from a particular station.

The strength of citizen observations is gained from their prevalence. Within Norway, the Netatmo network alone already outnumbers MET Norway's own network of WMO-compliant stations by a factor of around 50. A comparison of the networks for southeastern Norway (Fig. 1) highlights the significance of the increased coverage of the citizen network. Coverage is densest in cities along the coast; however, inland towns and mountainous areas with extensive cabin settlements are also well covered. We have focused our efforts on observations from Netatmo as their coverage in Norway is good and they have an application programming interface (API) that allows access to data in near–real time. We believe the methods we present here are applicable to other high-density networks with similar characteristics.

Although the uncertainty associated with individual citizen observations are higher than for professional stations, the massive redundancy of citizen stations gives the data source a unique advantage. Indeed, citizen networks have been shown to capture high-resolution weather phenomena, such as urban heat island effects in the Netherlands (Wolters and Brandsma 2012), London (Chapman et al. 2017), and Berlin (Meier et al. 2017). Figure 2a shows an early-morning, clear-sky wintertime temperature distribution in late March 2018 in Oslo, Norway, as measured by the Netatmo network. Several key meteorological features are represented by the observations, including an overall temperature inversion, valley cold pools, an urban heat island, and a warmer coastline. These features were not captured by

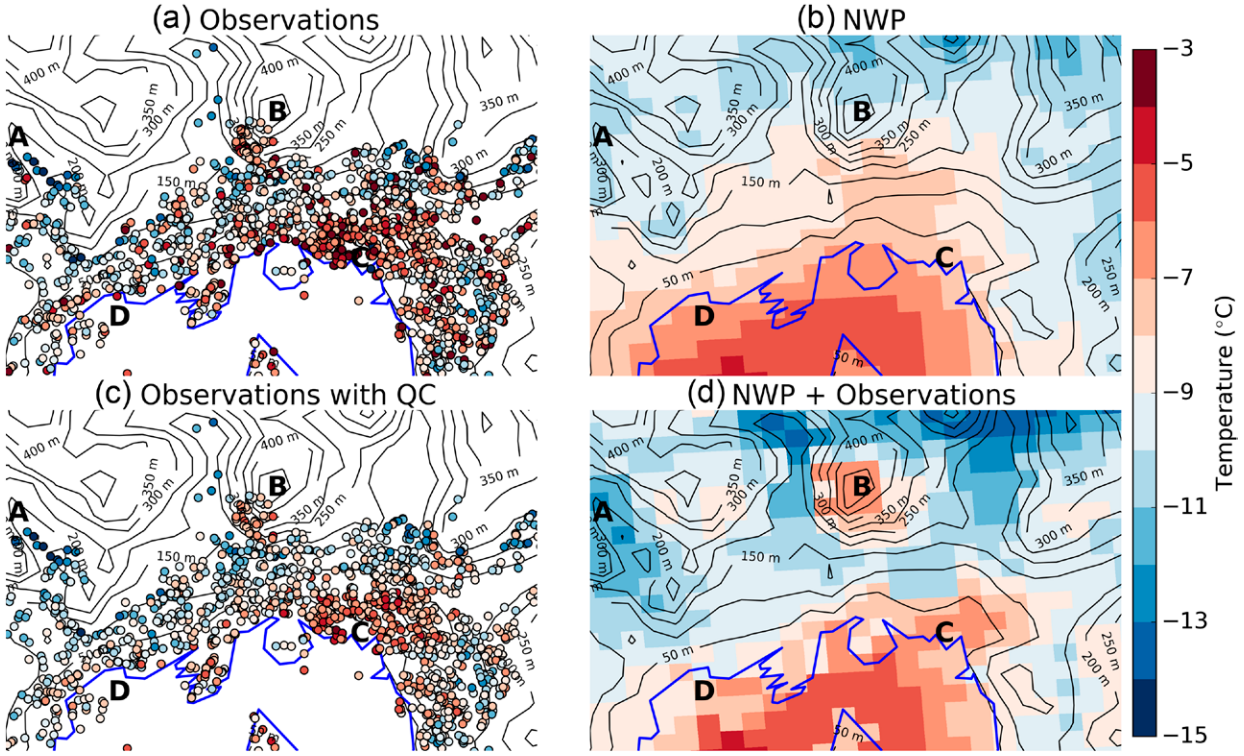

FIG. 2. Plots of 2-m temperatures over Oslo at 0500 UTC 28 Mar 2018 showing (a) all available observations from Netatmo, (b) model output from MEPS downscaled to 1 km, (c) observations from (a) after the QC procedures have been applied, and (d) postprocessed forecast incorporating model output from (b) and observations from (c). All panels use the color bar on the right. Black lines show elevation contour lines with a 50-m interval and blue lines are coastlines. The "A" marker identifies a valley with a cold pool, "B" a hilltop above an inversion, "C" an urban heat island, and "D" coastal effects.

MET Norway's operational NWP model at that particular time (Fig. 2b). In fact, the model predicts a standard temperature profile with lower temperatures at higher elevations. The two MET Norway WMO stations in the area (not shown) also do not adequately capture the meteorological features. In this case, the citizen network is the only available data source that potentially could represent the complex meteorological conditions.

Although the general spatial distribution of temperature makes physical sense, adjacent stations in Fig. 2a occasionally show contradictory temperatures that are unlikely to be caused by local effects. This is likely caused by poorly located stations.

Spatial quality control

Left untreated, such noisy data can be problematic for operational systems that require robust input. Using unreliable data that are not quality controlled to correct forecasts can lead to poor forecasts and reduced user confidence. Applying appropriate quality control (QC) methods is therefore an essential component when using these observations.

There is a wealth of QC methods available [see Fiebrich et al. (2010) for a review]. When stations are properly located, regularly maintained, and calibrated, they are less likely to make measurements that are not representative of the surrounding area. In this case, QC methods can focus on other sources of errors, such as faults in the sensor. Standard tests include plausibility checks, rate-of-change checks, and checks for measurements that are constant in time, indicating a mechanical failure (Zahumensky 2010).

With citizen networks, such time series methods can fail to identify systematic errors that are caused by ill-chosen locations. For example, a station placed too close to a building can exhibit diurnally varying biases that are difficult to detect by analyzing the time series alone. As metadata is rarely available, a black list of stations known to be poorly located cannot be

created. Human QC, where a QC expert manually evaluates observations using auxiliary data sources, cannot be used due to the large number of stations, and because operational NWP requires quality-controlled observations to be available with little delay. The QC methods selected must therefore be fully automatic.

Citizen networks need a different approach. We have focused on using methods that exploit the spatial properties of the observing network instead of temporal properties of individual stations. The massive redundancy of stations means that in many cases we have independent nearby observations available to help confirm or reject a particular observation. From a statistical point of view, neighboring observations are better predictors of performance than past observations for this type of observing network. Such spatial methods are also used for professional stations (Fiebrich et al. 2010); however, they are even more suitable for dense networks. The methods used are available in the open-source software package TITAN (www.github.com/metno/titan).

The first spatial test we perform is the "buddy check." The buddy check compares an observed value against other (buddy) observations that are within a 3-km radius and have elevations within 30 m. The observation is removed if its deviation from the average is more than twice the standard deviation of the observations in the neighborhood. This removes unrealistic finescale variability in the observations.

Next, the spatial consistency test (SCT; Lussana et al. 2010) is applied. This test is done by first fitting a vertical profile to observations and their station elevations in 100 km × 100 km regions using the approach of Frei (2014). The profile is then adjusted locally within the region based on the station density. Unlike the buddy check, the SCT bases the threshold for removing an observation on the observation density. It therefore performs a more restrictive test for data-dense than for data-sparse regions. A similar concept has been used, for example, by Dee et al. (2001). For a given observation, the SCT computes an expected value and error variance based on the other observations. If the ratio of squared deviation to the cross-validation error variance is greater than 4, the observation is removed. As most error sources for citizen observations contribute to a warm bias (e.g., direct sunlight or proximity to walls), we allow a less strict threshold (ratio of 8) for cold deviations from the expected. These settings were chosen based on trial and error. After an observation is removed, the process is repeated until no observations are removed. For further details about the SCT, the reader is referred to Lussana et al. (2010).

The spatial approach implies that lone stations cannot be trusted since there is no independent information to validate them. We therefore use an "isolation test" that removes stations that do not have at least five other stations within a 15-km radius and a 200-m elevation difference. The settings for the radius and elevation difference are conservative and are effectively a compromise between having a robust system and increasing the area where postprocessing can be applied.

After applying these three tests, the meteorological features in the Oslo example become much clearer (Fig. 2c). The tests are performed independently each hour, which causes some stations to be removed at different times of the day or in different months of the year. This is especially true for stations that receive direct sunlight at specific sun directions. In the example, the set of stations removed in the morning (Fig. 3a) is different than the set removed in the afternoon (Fig. 3b). This time independence allows observations from poorly located stations to contribute at times when their location does not lead to unrepresentative measurements.

On average, the tests remove 21% of observations. This does not mean that 21% of the observations are unreliable, but is a consequence of the conservative thresholds we have set in the spatial tests. Good observations, and consequently realistic finescale information, are occasionally removed. A greater number of observations are removed in daytime (Fig. 4) due to more observations being exposed to the sun. The spatial consistency test removes the

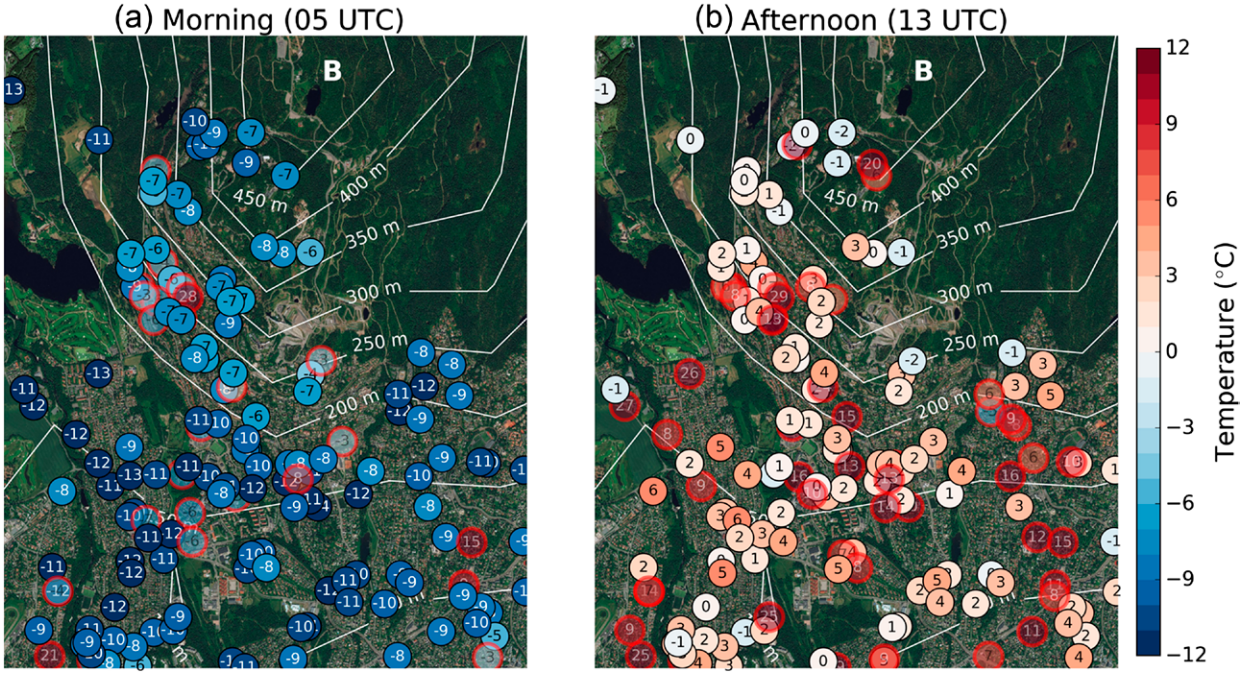

Fig. 3. The result of the QC procedure for observations near marker "B" in Fig. 2 for (a) 0500 and (b) 1300 UTC 28 Mar 2018. Observations with a red border have been flagged by QC. White lines show elevation contour lines with a 50-m interval. The "B" marker is the same as in Fig. 2. Background image source: ESRI World Imagery.

most (16.3% on average), followed by the isolation test (3.6% on average), and finally the buddy check (1.5% on average). The time-independent checks allow us to retain a higher fraction of measurements than, for example, Meier et al. (2017), where 53% of measurements were removed.

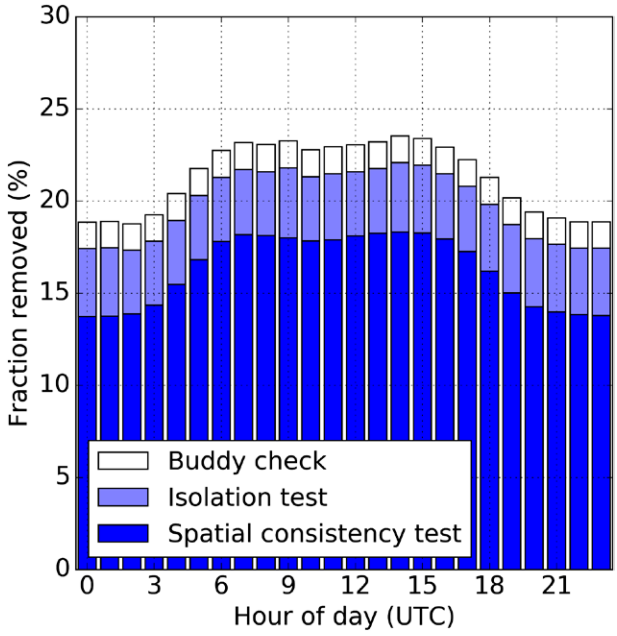

Fig. 4. Fraction of observations removed by three different QC methods as a function of time of day.

Creating a gridded truth

The remaining observations are now reasonably accurate measurements of the current temperature. These are then used to create a gridded truth (often called an analysis) on a 1 km × 1 km grid. The gridded truth gives the best estimate of the current conditions and is an important product on our forecast platform Yr. Again, since the interface allows lookup on the neighborhood scale, a high-resolution gridded product is needed. In addition to the public forecasts, specialized downstream users such as hydropower companies and flood prediction agencies use the gridded truth in their own models. The gridded truth also forms the basis for correcting our short-range weather forecasts, as will be discussed in the next section.

We construct a gridded truth by combining NWP model output and available observations, where each data source is weighted by its certainty. The product is updated every hour with the latest observations and the most recent NWP run. The NWP model used is the MetCoOp ensemble prediction system (MEPS; Frogner et al. 2019), which is run every 6 h and takes 2.5 h to complete. For a given point in time, this means that the most recent NWP run will be between 3 and 8 h old. The lead time that corresponds to the current time is then extracted

and is used together with the most recent observations, which are generally less than 10 min old. NWP output has inherent forecast error, whereas observations have uncertainty related to the station's representativity in the 1 km × 1 km square. However, the uncertainty of observations is generally much lower than for the NWP output. In areas with a high density of observations, the resulting analysis is therefore nearly a pure observation product whereas in observation-sparse areas, the analysis is close to the raw NWP output.

To combine the two data sources, we use a technique frequently used in data assimilation. A notable difference is that we are interpolating the observations to create a gridded truth as opposed to initializing a model. In particular, we use the analysis step of the local ensemble transform Kalman filter (LETKF; Hunt et al. 2007). In this approach, the NWP field is treated as a background, which is subsequently adjusted by the observations. The difference between the background and the observations (innovation) is interpolated in space. The adjustment of a particular grid point is based on the innovations at observation points in the vicinity and is based on the correlation between the grid point and the observation points. This correlation is modeled as a function that decays with horizontal and vertical distance, but is also based on the correlation structure provided by the 10-member ensemble.

The ensemble correlation structure ensures that innovation at one observing station will not be spread across meteorological features such as fronts, if the ensemble determines that the innovations are not correlated across the feature. The SCT in the quality control provides estimates of each station's representativity, which is used as a weight in the interpolation here. To get a deterministic forecast, we use the member from the gridded truth ensemble that corresponds to the control member of the NWP ensemble. The method is implemented in MET Norway's open-source gridded postprocessing tool gridpp (www.github.com/metno/gridpp). Further details of the method are presented in the appendix.

The application of the method to the complex situation in the Oslo area is shown in Fig. 2d. The gridded truth reconstructs the features captured by the observations. The temperature in the valleys is generally reduced and the temperature on top of hills increased. The inversion conditions at B are spread onto the other hills in the area. In observation-dense regions, the gridded truth not only corrects for model biases but also introduces features at spatial scales that are unresolved by the NWP model.

Adjusting short-term forecasts

The next step is to correct the short-range forecasts, which currently extend 60 h into the future. Users expect a seamless transition from the current conditions to the short-range forecast and consequently there can be no unrealistic jumps in temperature. The short-range forecasts must therefore be updated simultaneously with the current conditions at every hour and postprocessed on the same 1 km × 1 km grid.

The need to postprocess gridded NWP forecasts is a challenge faced by most forecast providers. A simple framework is to postprocess each grid point independently by using the gridded truth as an answer key. In this way, the complex problem of postprocessing gridded fields is split into two operations. First, the spatial problem of determining how biases vary in space is handled by the gridded truth, and second, the temporal problem of how biases vary across forecast lead times is handled grid point by grid point.

To tackle the temporal problem, we have developed a simple approach that recognizes the fact that NWP model bias has a real-time and diurnal component. The model bias the next few hours is correlated with the bias at the current time B_0. The bias at a particular time of day is also correlated with the bias the model had at the same time of day yesterday B_t. The corrected forecast $\hat{F}_t$ is a weighted combination of the raw forecasts F_t and the two biases:

$$\hat{F}_t = F_t + a_t B_0 + b_t B_t, \qquad (1)$$

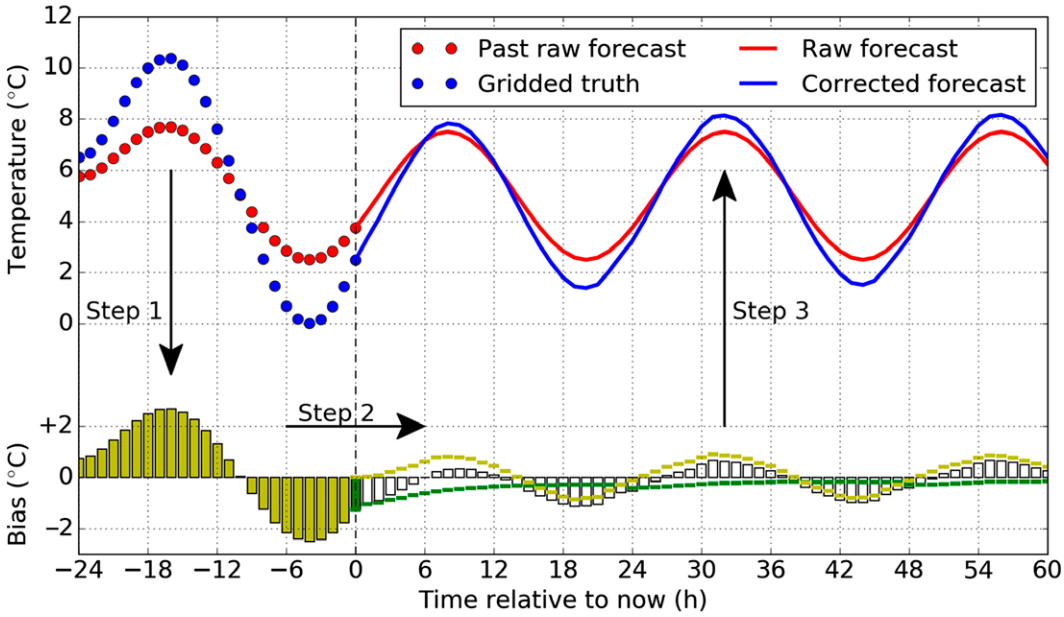

Fig. 5. Schematic diagram showing the steps to postprocess the forecasts. The first step computes past forecast biases against the gridded truth (yellow and green bars). The second step computes a forecast correction (white bars) based on these past biases. The forecast correction for a particular lead time is the weighted sum of the real-time bias (green bar) and the diurnally varying bias (yellow bars) that corresponds to the same time of day. The contribution of these two biases after weighting has been applied is shown by green and yellow lines, respectively. In step 3, this correction is applied to the raw forecast.

where a_t and b_t are weights that vary with lead time. The weights can be obtained from bivariate least squares estimation using historical NWP forecasts and corresponding observations at WMO-compliant stations. This single set of weights is then used to correct all grid points, and weights are fixed across all seasons.

The procedure for correcting a forecast is illustrated in Fig. 5. At the present time, a set of past NWP model output and gridded truths are available for the last 24 h (circles). The model exhibits a cold bias in daytime and a warm bias at night. Additionally, the model has a warm bias at the present time. These 25 biases are used to correct the raw forecasts (red line). The present bias contributes to the correction mostly for the first few hours and gradually decreases with lead time. The diurnal component in this example increases the temperature in daytime and decreases at night. The result is a forecast time series that matches the gridded truth in real time, incorporates the diurnally varying biases, and varies smoothly from lead time to lead time.

The method makes a compromise between removing long-term biases caused by systematic errors in the model, and adapting quickly to changing biases caused by weather regime changes. There is a wide variety of methods to postprocess NWP output using observations, such as Kalman filtering (Homleid 1995), model output statistics (MOS; Glahn and Lowry 1972), and analog methods (Delle Monache et al. 2011). In fact, the bias removal used here is similar in principle to the Localized Aviation MOS Program (LAMP; Ghirardelli and Glahn 2010) used by the U.S. National Weather Service to produce point forecasts of aviation parameters, where recent observations are also used to perform real-time correction of NWP output. Our point here is not to present a perfect postprocessing method, but rather to illustrate how the gridded-truth framework allows a high-density network of citizen observations to be integrated into the postprocessing of short-range forecasts. Within this framework, most existing point forecast methods can be used. An alternative viable approach is Gridded MOS (GMOS; Glahn et al. 2009), where forecasts at observation points are first postprocessed, and then interpolated out onto the grid.

Impact on accuracy

It is important to establish the impact of the citizen network on the quality of the gridded truths and forecasts. Operationally, we include observations from both Netatmo and our own network of WMO stations to get the best possible result. However, for the impact evaluation, we have rerun our forecast system without including WMO stations. The Netatmo-based analyses and forecasts are subsequently verified against the nonassimilated WMO observations. The system was rerun for the time period from July 2017 to July 2018 and the weights in Eq. (1) were trained using data from November 2016 to June 2017. Analyses and forecasts were produced for each hour, resulting in 24 sets of analyses and 60-h forecasts every day. For this evaluation, we have chosen all WMO stations in Norway that have at least 5 Netatmo stations within a 5-km radius and within 200-m elevation. This results in 93 stations.

We first evaluate the gridded truth using the mean absolute error (MAE), a commonly used metric in forecast verification. This is shown in Figs. 6a and 6b. On average, the inclusion of citizen observations reduces the MAE by 33% (a reduction from 1.24° to 0.84°C). The improvement is larger during winter, likely due to the prevalence of inversions that the NWP model often fail to capture. The figures also show the result of creating the gridded truth using observations that have not been quality controlled. These analyses are only marginally better than the NWP forecast, and are worse in daytime and during the summer. QC is thus essential for getting added value from citizen observations.

Many of our end users are not sensitive to small forecast errors. We therefore also evaluate the system by focusing on large errors, which we will call analysis/forecast busts and are defined as errors greater than 3°C (Figs. 6c,d). The citizen network reduces the frequency of analysis busts by 68% (a reduction from 8.1% to 2.6%), which is an even larger impact than for the MAE.

We also investigated the bias, defined as the average difference between the analysis and the observation (Figs. 6e,f). The uncontrolled Netatmo observations have a clear bias, which is in line with results found by Chapman et al. (2017), and is likely caused by some fraction of stations placed in direct sunlight. Even with QC, a small warm bias of 0.5°C is still present in the analyses. The bias is stronger in daytime and during the summer. The remaining bias is likely caused by observations that are not in direct sunlight, but still exhibit a warm bias that is too small for the QC to remove. Despite the bias, the overall quality of the forecasts are still greatly improved. The bias likely affects the MAE more than the analysis bust metric, which can explain why the latter has a greater improvement.

The improvements to the gridded truth carry over to the forecasts (Fig. 7). The improvement is most noticeable in the first 6 h due to the strong correlation with the real-time bias, but a relatively uniform improvement still exists beyond 12 h and is caused by local systematic biases that are lead-time independent and that the diurnally varying component is able to correct for. In fact, the improvements for both scores are statistically significant at the 0.05 level for all lead times, as determined by 10,000 bootstrap samples where temporal correlation has been accounted for by first aggregating the initialization times into 2-day blocks.

Effect of station density

We have seen significant improvements in areas where the Netatmo density is high. How does the improvement diminish with decreasing station density? We computed the two scores again, this time using all 254 WMO-standard-compliant stations in Norway, and computed the average improvement for different network densities (Fig. 8). The network density at a particular point is computed by summing up the number of citizen stations that influence the point. Each observation is weighted by its distance to the grid point. In this way a single number can be computed that represents the total influence of the network. A sum of 4 could either represent a grid point where four citizen stations are exactly collocated with the grid

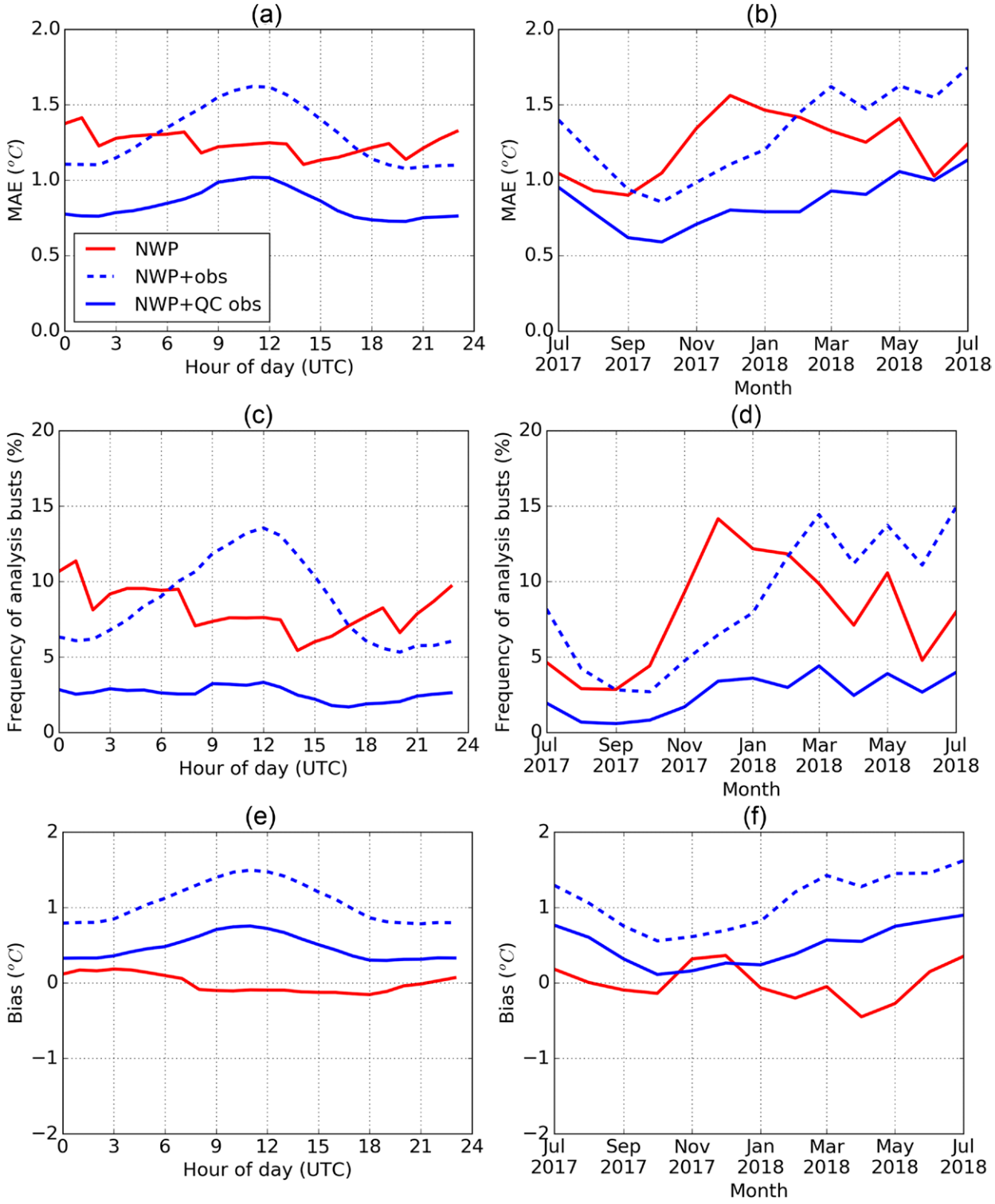

FIG. 6. Verification scores for the Netatmo-based gridded truths verified against nonassimilated observations from 93 WMO-compliant stations in Norway for which the Netatmo network is sufficiently dense. (a),(b) MAE. (c),(d) Frequency of analysis busts (errors greater than 3°C). (e),(f) bias (gridded truth minus observed) for scores as a function of (left) time of day and (right) month of year. The panels show the raw model output (solid red), the merged product using unchecked Netatmo observations (dashed blue), and the merged product using quality-controlled Netatmo observations (solid blue).

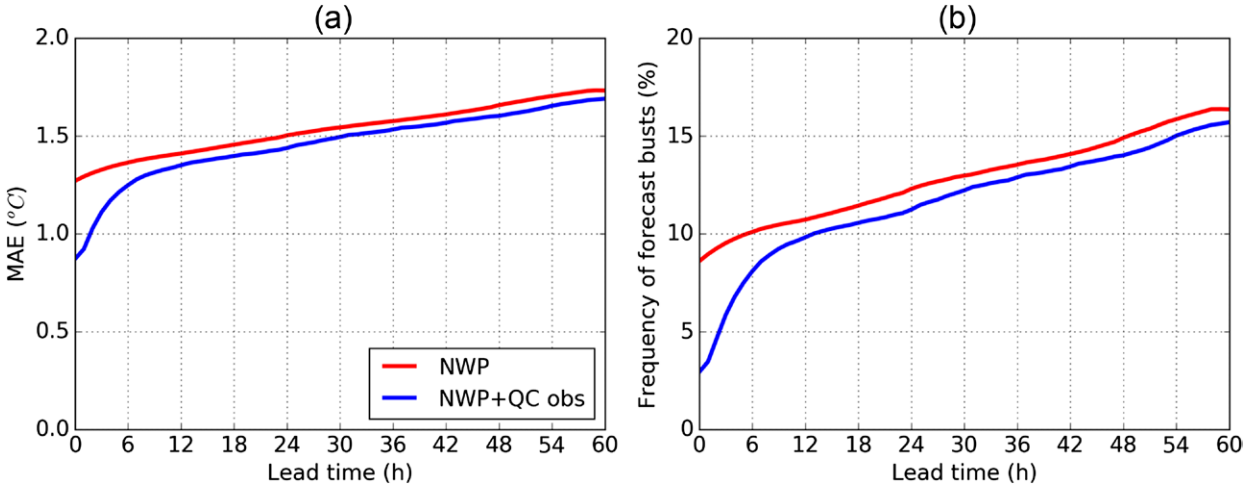

FIG. 7. As in Fig. 6, but for the forecasts as a function of forecast lead time.

point, six stations that are located 6 km away, or eight stations that are located 12 km away.

As expected, higher network densities result in greater analysis improvements. Improvements to the frequency of analysis busts and MAE, however, saturate when more than four citizen stations are available. This can be caused by systematic biases in the network, such as the warm bias noted earlier, that do not reduce as the number of stations increase. A second reason is that in some cases, the verifying WMO stations are located at airports outside cities, whereas the citizen stations that influence the point are located in urban areas.

By using the relationship between density and improvement in Fig. 8, we can construct a map of estimated analysis improvements (Fig. 9) by using the local density. The map shows that the citizen network provides significant analysis improvements for large areas of southern Norway. As the network grows, remote areas will start to fill with stations, which will also allow many of the currently isolated stations to have an effect. Currently, the areas without improvements are mainly mountainous regions with few settlements.

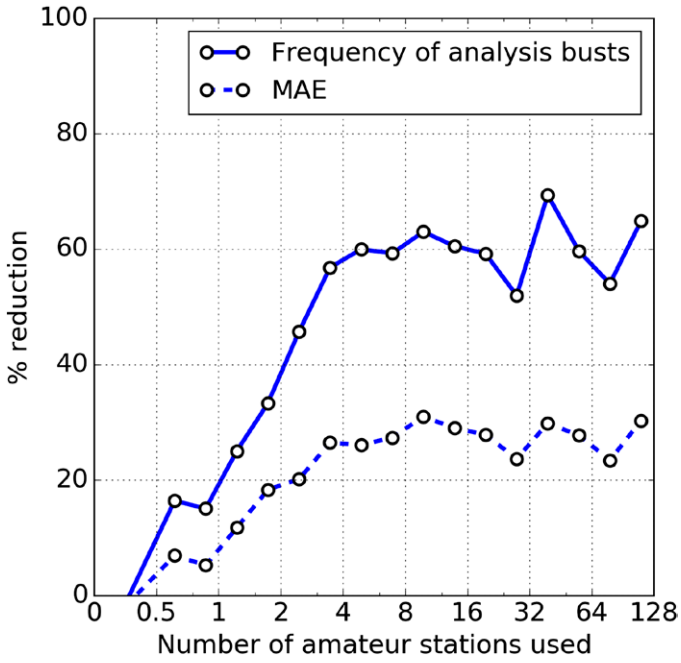

FIG. 8. Analysis improvement as a function of station density. Station density is given as the number of stations available, each reduced in weight by their localization. Analysis improvement is shown as the percentage reduction compared against the raw NWP for two verification metrics defined in Fig. 6.

Conclusions

In this study we have presented a possible pathway for integrating a vast network of citizen observations into operations at a national weather forecast service. To summarize, here are some key points.

First, QC is essential to get value from the network. Without QC, the skill of the postprocessed temperature forecasts, compared to the NWP forecasts, is worse in daytime and during summer. With QC, the citizen network has a large positive impact. MAE for the current conditions is reduced by 33% and the frequency of analysis busts are reduced by a factor 3.

Second, citizen networks should be treated as a spatial data source rather than individual time series. The massive redundancy of stations means that there are many independent nearby stations that can be used to confirm or reject a given observation. The spatial approach allows us to retain a higher fraction of observations as different stations may be rejected at different times of the day. In this way only 21% of the observations are rejected. Also, as there is no temporal component in the checks, observations from moving sensors (e.g., on board cars or bicycles) can easily be integrated into this framework in the future.

And third, the postprocessing of the forecasts needs to take into account observation uncertainty. Even after QC, there is uncertainty associated with all observations due to both measurement errors and lack of representativeness. To account for this uncertainty, we employ methods often used in data assimilation for NWP models. A gridded truth is created where the data sources are weighted by their certainty. This also greatly simplifies the task of correcting the short-term forecasts. The corrections can be applied grid point by grid point without dealing with how the corrections should vary in space. Another benefit of this approach is that changes to the network of stations are handled without any changes to the operational system.

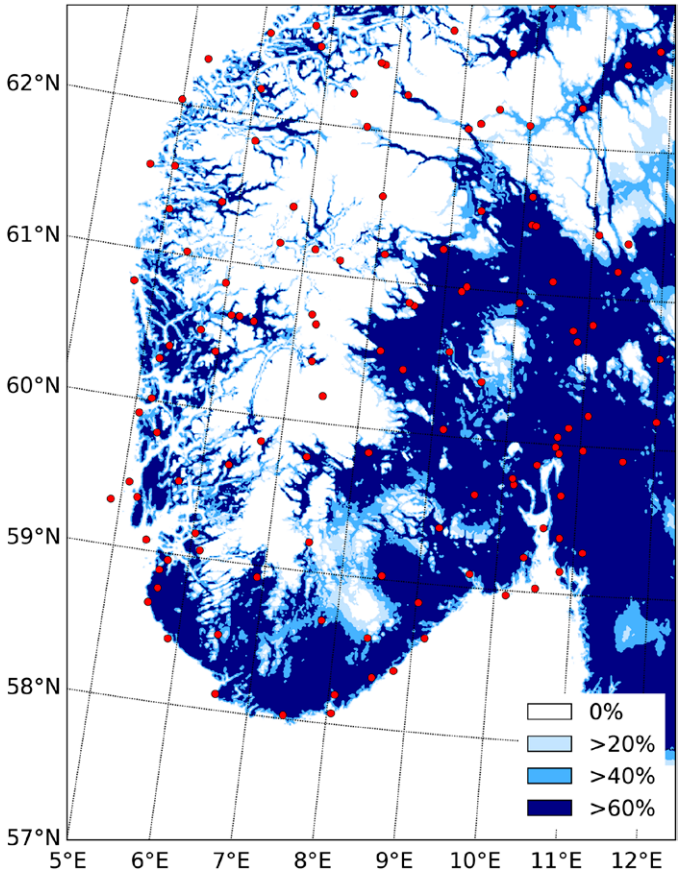

Fig. 9. Estimate of the improvement of the gridded truth for different areas in southern Norway. The score is the reduction in the fraction of analyses with errors larger than 3°C. This is based on the information in Fig. 8 and the actual station density. The red circles are locations of MET Norway's network of WMO-compliant stations. Blue lines are coastlines and lakes.

A prerequisite for integrating citizen observations into operational weather forecasts is access to the data in near–real time. In the case of Netatmo, the station owner can choose to share their weather observations with Netatmo, which is then anonymized and made available through an API. There is also a growing number of networks that collect citizen observations independent of manufacturer, such as the Citizen Weather Observer Program (CWOP; www.wxqa.com) and mPING (Elmore et al. 2014) in the United States, and the Weather Observation Website (WOW; http://wow.metoffice.gov.uk) in the United Kingdom. The funding model for citizen stations differs from that of conventional stations in that private citizens take on the ownership of the station and its maintenance. The use of networks with distributed ownership in operational systems relies on centralized data access, and we therefore strongly encourage instrument manufactures and network operators, both public and private, to provide APIs where the observations from the entire network can be accessed in near–real time.

The number of citizen observations are growing rapidly, and we believe this data source will play a key role in helping to reach the goal of providing high-quality forecasts on a local scale. This will increase the value of the forecasts for a range of applications in diverse societal sectors like energy production and distribution, transportation, agriculture, health, and water management.

Acknowledgments. This research was partially supported by RADPRO (Radar for Improving Precipitation Forecast and Hydropower Energy Production), an innovative industry project funded by the Research Council of Norway (NFR) and partnering hydropower industries.

Appendix: Method for creating a gridded truth

This appendix describes the method used to create a gridded truth. The notation is based on both Ide et al. (1997) and Sakov and Bertino (2011). We denote the total number of grid points in the forecast domain by m, the number members in the NWP ensemble by k, and the number of observations by p. Uppercase boldface sans serif symbols are used for matrices, lowercase boldface serif symbols for vectors, and italic symbols for scalars. For an arbitrary matrix $\mathbf{X}$, $\mathbf{X}_i$ means the ith column; $\mathbf{X}_{i,:}$ the ith row; and $\mathbf{X}_{ij}$, the element at the ith row and jth column. For an arbitrary vector $\mathbf{x}$, $\mathbf{x}_i$ denotes the ith element.

The background ensemble members are combined by columns into the $m \times k$ matrix $\mathbf{X}^b$. The ensemble mean is the m vector $\mathbf{x}^b = 1/k\mathbf{X}^b\mathbf{1}$, where $\mathbf{1}$ is the m vector with all elements equal to 1. The $m \times k$ perturbation matrix $\mathbf{A}^b$ has columns $\mathbf{A}^b_i = \mathbf{X}^b_i - \mathbf{x}^b$. A similar notation applies to the gridded truth (analysis), only with the superscript a instead of b. The in situ observations are stored in the p vector $\mathbf{y}^o$.

We implemented a grid point by grid point analysis scheme based on the analysis step of LETKF (Hunt et al. 2007). An $\mathbf{R}$ localization technique (Greybush et al. 2011) has been applied to adjust for spurious long-distance correlations, where the $\mathbf{R}$ elements are multiplied by a distance-dependent function. For a generic grid point i, the local domain includes only the nearest q observations. As in the paper by Sakov and Bertino (2011), an upper accent i has been used to denote the local version of a variable. The local observation q vector is $\overset{i}{\mathbf{y}}{}^o$ and its $q \times q$ error covariance matrix is $\overset{i}{\mathbf{R}}$. The step-by-step interpolation scheme is

$$\overset{i}{\mathbf{\Gamma}}{}^a = \left[\Delta^{-1}(k-1)\mathbf{I} + \left(\overset{i}{\mathbf{H}}\mathbf{A}^b\right)^{\mathrm{T}} \overset{i}{\mathbf{R}}{}^{-1}\left(\overset{i}{\mathbf{H}}\mathbf{A}^b\right)\right]^{-1}, \tag{A1}$$

$$\overset{i}{\mathbf{K}}_{i,:} = \mathbf{A}^b_{i,:}\overset{i}{\mathbf{\Gamma}}{}^a\left(\overset{i}{\mathbf{H}}\mathbf{A}^b\right)^{\mathrm{T}} \overset{i}{\mathbf{R}}{}^{-1}, \tag{A2}$$

$$\mathbf{x}^a_i = \mathbf{x}^b_i + \overset{i}{\mathbf{K}}_{i,:}\left(\overset{i}{\mathbf{y}}{}^o - \overset{i}{\mathbf{H}}\mathbf{x}^b\right), \tag{A3}$$

$$\overset{i}{\mathbf{A}}{}^a_{i,:} = \mathbf{A}^b_{i,:}\left[(k-1)\overset{i}{\mathbf{\Gamma}}{}^a\right]^{1/2}, \tag{A4}$$

$$\mathbf{X}^a_{i,:} = \mathbf{x}^a_i + \overset{i}{\mathbf{A}}{}^a_{i,:}. \tag{A5}$$

Here $\overset{i}{\mathbf{H}}$ is the linear observation operator mapping m vectors into q vectors (when applied to a matrix is intended to be applied separately to each of its columns) and consists of a nearest-neighbor interpolation with an adjustment for elevation differences between stations and the nearest grid points using a constant near-surface lapse rate of −6.5°C km^{-1}. The term $\overset{i}{\mathbf{K}}$ represents the ith row of the local Kalman gain and its expression is derived directly from the LETKF theory; and Δ is an ensemble scaling factor that accounts for underdispersion of the ensemble and which we, through trial and error, have set to a value of 2.

Observation errors have been assumed to be independent of one another. Therefore, the $q \times q$ matrix $\overset{i}{\mathbf{R}}{}^{-1}$ is diagonal with error variances σ_o^2 set to 0.25°C^2. Because $\overset{i}{\mathbf{R}}{}^{-1}$ appears in Eqs. (A1) and (A2), its element $\overset{i}{\mathbf{R}}{}^{-1}_{jj}$ can be written as

$$\overset{i}{\mathbf{R}}{}^{-1}_{jj} = \frac{\rho\left(\mathbf{r}_i, \mathbf{r}_j\right)}{\sigma_o^2}. \tag{A6}$$

The localization factor $\rho(\mathbf{r}_i, \mathbf{r}_j)$ is a function of the geographical parameters at the two locations $\mathbf{r}_i$ and $\mathbf{r}_j$, which accounts for the horizontal distance between the two points $d(\mathbf{r}_i, \mathbf{r}_j)$,

their elevation difference $z(\mathbf{r}_i, \mathbf{r}_j)$, and the difference between their land-area fractions $w(\mathbf{r}_i, \mathbf{r}_j)$. $\rho(\mathbf{r}_i, \mathbf{r}_j)$ is modeled as

$$\rho\left(\mathbf{r}_i, \mathbf{r}_j\right) = \exp\left\{-\frac{1}{2}\left[\frac{d\left(\mathbf{r}_i, \mathbf{r}_j\right)}{D^h}\right]^2 - \frac{1}{2}\left[\frac{z\left(\mathbf{r}_i, \mathbf{r}_j\right)}{D^z}\right]^2\right\}\left[1-\left(1-w_{\min}\right)\left|w\left(\mathbf{r}_i, \mathbf{r}_j\right)\right|\right], \tag{A7}$$

where D^h = 10,000 and D^z = 100 m are reference length scales used to introduce different covariance suppression rates along the horizontal and vertical directions, and $w_{\min}$ = 0.5 sets the minimum value for the factor related to land area fraction when $w(\mathbf{r}_i, \mathbf{r}_j)$ is at a maximum (i.e., equals 1).

In Eq. (A5) $\mathbf{X}^a_{i,:}$ is the vector of ensemble members representing the gridded truth for the ith grid point.

References

Chapman, L., C. Bell, and S. Bell, 2017: Can the crowdsourcing data paradigm take atmospheric science to a new level? A case study of the urban heat island of London quantified using Netatmo weather stations. *Int. J. Climatol.*, **37**, 3597–3605, https://doi.org/10.1002/joc.4940.

Dee, D. P., L. Rukhovets, R. Todling, A. M. da Silva, and J. W. Larson, 2001: An adaptive buddy check for observational quality control. *Quart. J. Roy. Meteor. Soc.*, **127**, 2451–2471, https://doi.org/10.1002/qj.49712757714.

Delle Monache, L., T. Nipen, Y. Liu, G. Roux, and R. Stull, 2011: Kalman filter and analog schemes to postprocess numerical weather predictions. *Mon. Wea. Rev.*, **139**, 3554–3570, https://doi.org/10.1175/2011MWR3653.1.

Elmore, K. L., Z. L. Flamig, V. Lakshmanan, B. T. Kaney, V. Farmer, H. D. Reeves, and L. P. Rothfusz, 2014: MPING: Crowd-sourcing weather reports for research. *Bull. Amer. Meteor. Soc.*, **95**, 1335–1342, https://doi.org/10.1175/BAMS-D-13-00014.1.

Fiebrich, C. A., C. R. Morgan, A. G. McCombs, P. K. Hall, and R. A. McPherson, 2010: Quality assurance procedures for mesoscale meteorological data. *J. Atmos. Oceanic Technol.*, **27**, 1565–1582, https://doi.org/10.1175/2010JTECHA1433.1.

Frei, C., 2014: Interpolation of temperature in a mountainous region using nonlinear profiles and non-Euclidean distances. *Int. J. Climatol.*, **34**, 1585–1605, https://doi.org/10.1002/joc.3786.

Frogner, I.-L., A. T. Singleton, M. Ø. Køltzow, and U. Andrae, 2019: Convection-permitting ensembles: Challenges related to their design and use. *Quart. J. Roy. Meteor. Soc.*, **145**, 90–160, https://doi.org/10.1002/QJ.3525.

Gasperoni, N. A., X. Wang, K. A. Brewster, and F. H. Carr, 2018: Assessing impacts of the high-frequency assimilation of surface observations for the forecast of convection initiation on 3 April 2014 within the Dalla–Fort Worth test bed. *Mon. Wea. Rev.*, **146**, 3845–3872, https://doi.org/10.1175/MWR-D-18-0177.1.

Ghirardelli, J. E., and B. Glahn, 2010: The meteorological development laboratory's aviation weather prediction system. *Wea. Forecasting*, **25**, 1027–1051, https://doi.org/10.1175/2010WAF2222312.1.

Glahn, B., K. Gilbert, R. Cosgrove, D. P. Ruth, and K. Sheets, 2009: The gridding of MOS. *Wea. Forecasting*, **24**, 520–529, https://doi.org/10.1175/2008WAF2007080.1.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203–1211, https://doi.org/10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2.

Greybush, S. J., E. Kalnay, T. Miyoshi, K. Ide, and B. R. Hunt, 2011: Balance and ensemble Kalman filter localization techniques. *Mon. Wea. Rev.*, **139**, 511–522, https://doi.org/10.1175/2010MWR3328.1.

Homleid, M., 1995: Diurnal corrections of short-term surface temperature forecasts using the Kalman filter. *Wea. Forecasting*, **10**, 689–707, https://doi.org/10.1175/1520-0434(1995)010<0689:DCOSTS>2.0.CO;2.

Hunt, B. R., E. J. Kostelich, and I. Szunyogh, 2007: Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D*, **230**, 112–126, https://doi.org/10.1016/j.physd.2006.11.008.

Ide, K., P. Courtier, M. Ghil, and A. Lorenc, 1997: Unified notation for data assimilation: operational, sequential and variational. *J. Meteor. Soc. Japan*, **75**, 181–189, https://doi.org/10.2151/jmsj1965.75.1B_181.

Lussana, C., F. Uboldi, and M. R. Salvati, 2010: A spatial consistency test for surface observations from mesoscale meteorological networks. *Quart. J. Roy. Meteor. Soc.*, **136**, 1075–1088, https://doi.org/10.1002/qj.622.

Madaus, L. E., G. J. Hakim, and C. F. Mass, 2014: Utility of dense pressure observations for improving mesoscale analyses and forecasts. *Mon. Wea. Rev.*, **142**, 2398–2413, https://doi.org/10.1175/MWR-D-13-00269.1.

Mass, C. F., and L. E. Madaus, 2014: Surface pressure observations from smartphones: A potential revolution for high-resolution weather prediction? *Bull. Amer. Meteor. Soc.*, **95**, 1343–1349, https://doi.org/10.1175/BAMS-D-13-00188.1.

Meier, F., D. Fenner, T. Grassmann, M. Otto, and S. D., 2017: Crowdsourcing air temperature from citizen weather stations for urban climate research. *Urban Climate*, **19**, 170–191, https://doi.org/10.1016/j.uclim.2017.01.006.

Muller, C., L. Chapman, S. Johnston, C. Kidd, S. Illingworth, G. Foody, A. Overeem, and R. Leigh, 2015: Crowdsourcing for climate and atmospheric sciences: Current status and future potential. *Int. J. Climatol.*, **35**, 3185–3203, https://doi.org/10.1002/joc.4210.

Müller, M., and Coauthors, 2017: AROME-MetCoOp: A Nordic convective-scale operational weather prediction model. *Wea. Forecasting*, **32**, 609–627, https://doi.org/10.1175/WAF-D-16-0099.1.

Sakov, P., and L. Bertino, 2011: Relation between two common localisation methods for the EnKF. *Computat. Geosci.*, **15**, 225–237, https://doi.org/10.1007/s10596-010-9202-6.

WMO, 2014: Guide to meteorological instruments and methods of observation. WMO-8, World Meteorological Organization, 1167 pp., www.wmo.int/pages/prog/www/IMOP/CIMO-Guide.html.

Wolters, D., and T. Brandsma, 2012: Estimating the urban heat island in residential areas in the Netherlands using observations by weather amateurs. *J. Appl. Meteor. Climatol.*, **51**, 711–721, https://doi.org/10.1175/JAMC-D-11-0135.1.

Zahumensky, I., 2010: World guidelines on quality control procedures for data from automatic weather stations. Guide to the Global Observing System, WMO-488, 198–207, https://library.wmo.int/doc_num.php?explnum_id=4236.