Meteorological Applications

RMetS Royal Meteorological Society

RESEARCH ARTICLE OPEN ACCESS

Impact of the Assimilation of Surface Observations on Limited-Area Forecasts Over Complex Terrain

Giorgio Doglioni[1,2] | Stefano Serafin[3] | Martin Weissmann[3] | Gianluca Ferrari[4] | Dino Zardi[1,2]

[1]Department of Civil Environmental and Mechanical Engineering (DICAM), University of Trento, Trento, Italy | [2]Center Agriculture, Food, Environment (C3A), University of Trento, Trento, Italy | [3]Department of Meteorology and Geophysics, University of Vienna, Vienna, Austria | [4]Hypermeteo SRL, Padua, Italy

Correspondence: Giorgio Doglioni (giorgio.doglioni@unitn.it)

Received: 23 December 2024 | **Revised:** 19 August 2025 | **Accepted:** 15 September 2025

Funding: This work was supported by Hypermeteo S.r.l., Università degli Studi di Trento, European Union-NextGenerationEU, PNRR-PRIN2022, Grant code 2022NEWP4J, CUP E53D23004450006.

Keywords: complex terrain | data assimilation | European Alps | surface observations | WRF model

ABSTRACT

The article presents results from a computationally low-cost regional numerical weather prediction chain based on the Weather Research and Forecasting (WRF) model and its data assimilation (DA) suite WRFDA. Experiments with 24-h forecasts were performed twice daily (at 00 and 12 UTC) over a domain encompassing the European Alps and their surroundings with a 3.5 km grid spacing. The assimilation of surface observations with the 3D-Var algorithm improves near-surface temperature and humidity forecasts compared to control runs without assimilation. The forecast skill for near-surface variables is evaluated using independent surface observations. In the first six forecast hours, it is generally better in the assimilation experiments than in the control ones, with a mean error reduction of 0.26 K for temperature and 0.13 g kg^{-1} for specific humidity in the 00 UTC runs, and of 0.12 K for temperature and 0.18 g kg^{-1} for specific humidity in the 12 UTC runs. The assimilation reduces the standard deviation of the errors by a factor between 7% and 10% both for temperature and specific humidity. Verification with radiosonde measurements shows that assimilating surface observations increases the mean error in temperature and humidity forecasts within the planetary boundary layer (PBL), relative to the control. We show that the vertical structure of the adjustments to the model state resulting from DA (the analysis increments) is such that model biases are reduced near the surface but amplified higher up in the PBL. Finally, the assimilation of surface observations has a different impact on surface temperature forecasts in mountainous regions compared to adjacent plains. The error reduction is substantially higher in the plains than in the mountains, which likely depends on the inappropriate spreading of information along terrain-following model levels by the static covariances in 3D-Var. The relative accuracy of surface temperature forecasts in these two regions has a diurnal variability, with larger mean errors in the mountains during the day and in the plains at night.

1 | Introduction

Timely and accurate weather forecasts are crucial for many social and economic activities. Numerical weather prediction (NWP) models have become essential tools in this regard, with steady improvements over the past decades from advances in computational power, numerical modeling, and data assimilation (DA) algorithms. The starting point in generating NWP forecasts is the initialization of the numerical model runs, a process usually performed cyclically every 3– 6 h and relying on DA. DA is the algorithmic method that integrates model data from a previous run, the so-called "first guess," with recent

atmospheric observations to produce an updated estimate of the initial atmospheric state. This estimate, that is, the "analysis," serves as the starting point for the model run. The DA process and the model run require substantial computer time (i.e., typically several hours). This implies that the most accurate early hours of a forecast often become available only after their validity time has passed. This issue is especially critical for large domains and high-resolution models with high computational demands.

To meet the need for timely, convection-permitting short-term forecasts, rapid update cycle (RUC) systems (Benjamin et al. 2004) assimilate near-real-time observations into a limited area model (LAM) every 1–3 h. While effective, RUC requires dense observations, advanced algorithms, and substantial computational resources, making it mainly feasible for national weather services.

This work aims to improve the quality and timeliness of LAM forecasts over the European Alps (3.5 km grid) without significantly increasing their computational cost. We develop a Weather Research and Forecasting (WRF)-based system that can be implemented by private companies or local weather services with limited resources, focusing on forecasts of surface temperature and humidity.

We derive initial condition (IC) and boundary condition (BC) for our deterministic limited-area simulations with the WRF model (WRF, Skamarock et al. 2019) from the operational ICON-EU prediction system (ICOsahedral-Nonhydrostatic model, Zängl et al. 2015), and use 3D-Var (Lorenc 1986) to assimilate a set of additional surface observations that are not used in ICON-EU analyses. At this stage, we do not aim at building a stand-alone forecasting system operating in cycling mode. Rather, we rely on the ICON-EU assimilation system to ingest all observations available until 6 h before initialization. Our assimilation step adds near-real-time surface data at initialization to improve WRF forecasts relative to pure dynamical downscaling.

This approach leverages a dense surface observation network available in the Alpine region (Giazzi et al. 2022). Indeed, for most operationally run weather prediction models, only a small fraction of all the available surface measurements is routinely assimilated. In this work, we assimilate a fraction of the potentially available additional surface observations and determine their impact on short-term forecast skill. As we will demonstrate, the assimilation of surface observations, while conceptually straightforward, poses some challenges that must be carefully addressed.

Some previous studies have already examined the benefits and challenges associated with assimilating surface observations from a variety of diverse observational networks, operated either by institutional bodies or by private citizens. For instance, Sgoff et al. (2022) tested the assimilation of crowdsourced temperature observations in the ICON model using the Kilometer-scale ENsemble Data Assimilation (KENDA; Schraff et al. 2016) system and reported that they improved the forecast if an appropriate bias correction scheme was used to reduce systematic deviations between the model and the observations. In particular, the assimilation of bias-corrected crowdsourced temperature and humidity observations led to a root mean square error (RMSE) reduction of up to 15% for 2-m dewpoint temperature and up to 5% for 2-m relative humidity compared to control forecasts (i.e., a parallel forecast whose IC is an analysis computed without the additional set of observations). This benefit was visible in the forecasts of up to 12 h in lead time. Demortier et al. (2024) assimilated crowdsourced surface pressure observations in the AROME model with the 3DEnVar algorithm and found improvements in surface pressure forecasts lasting up to 12 h. The benefit reported by Demortier et al. (2024) following the assimilation of bias-corrected, quality-checked, and thinned surface pressure observations with respect to a control forecast is a 24% reduction in RMSE in surface pressure at the analysis time, decaying to about 2% at 9-h lead time. In a subsequent work, Demortier et al. (2025) tested the independent assimilation of crowdsourced surface temperature and relative humidity observations, reporting significant improvements in the forecasts for the same variables up to 3-h lead time with the 3DEnVar algorithm. Instead, forecast degradation was observed when using 3D-Var. The works cited above have shown the potential added value from the assimilation of nonconventional surface observations in the French and German NWP operative systems. Here, we adopt a similar approach but using a different NWP system, investigating the impact on the forecasts of the assimilation of nonconventional surface observations in the previously described WRF-based low-cost operational system.

Several studies have examined the impact of assimilating surface observations from conventional and unconventional networks on WRF forecasts. Pu et al. (2013) compared the assimilation of surface observations in an observing system simulation experiment (OSSE) setup using the WRF model and the ensemble Kalman filter (EnKF) implemented in Data Assimilation Research Testbed (DART; Anderson et al. 2009) to a corresponding setup with the 3D-Var algorithm from WRFDA. They underlined the difficulties in the assimilation of surface observations in complex terrain with 3D-Var, even in an idealized setup. In another study by Ha and Snyder (2014), the EnKF from DART was used to assimilate surface observations over the contiguous United States, with beneficial effects on short-term forecasts of surface wind components, temperature, and dewpoint temperature. While the above studies underlined the efficacy of the EnKF algorithm, positive results using 3D-Var are documented for precipitation forecasts in Ha and Lee (2012) and Chen et al. (2020), where it is shown that the assimilation of GTS and non-GTS surface observations led to improved precipitation forecasts in the Korean peninsula. The impact of the assimilation of surface observations with the 3D-Var algorithm on forecasts of surface variables was not thoroughly investigated in the above works. Here, we focus on that, albeit considering a different domain and setup.

The target area of the present study is the mountainous terrain of the European Alps and the flat, regular terrain of the adjacent Po Valley. Previously, Maggioni et al. (2023) investigated the effect of the assimilation of surface observations on a precipitation event in this area, obtaining overall positive results. While the target area is roughly the same as Maggioni et al. (2023), we have a different aim, which is assessing the

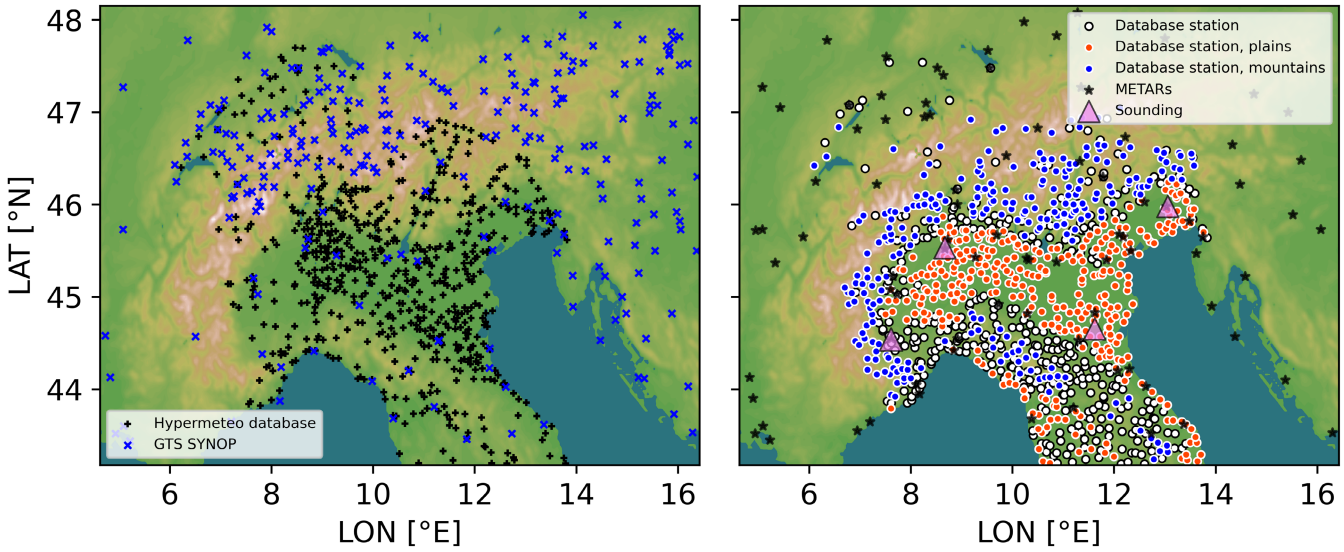

FIGURE 1 | (a) Experimental domain and terrain elevation. Dots and stars indicate stations whose observations were assimilated. (b) Map of the verifying observations. Black stars correspond to METAR stations and lilac triangles to soundings. Stations in Hypermeteo S.r.l. database are represented by dots: Red for plain, blue for mountain, and white for stations that do not fall into either category.

impact of the assimilation of surface observations on temperature and humidity forecasts.

Orographic variability adds complexity to our task, as surface observations already face representativeness issues in flat terrain and even more so in mountainous regions. Previous research has highlighted these issues, showing that surface observation assimilation must be approached cautiously in such regions (Ancell et al. 2014; Pu et al. 2013; Hacker et al. 2018).

Rotach et al. (2022) compared forecasts of surface variables by the European Center for Medium-Range Weather Forecasts, Integrated Forecasting System model (ECMWF IFS) with observations made over the European Alps (elevation greater than 1000 m) and the surrounding plains (elevation less than 500 m), reporting substantially higher model skill over the plains. We repeat a similar analysis but with two main differences: first, we consider a high-resolution LAM, and second, we use a network of surface observations that provides a denser coverage across the mountainous regions within the Italian borders and the adjacent plains.

To summarize, we can state the following specific aims of this work:

- Evaluate the impact of assimilating GTS and non-GTS surface observations using the 3D-Var algorithm on forecast verification scores, quantified here as the mean and the standard deviation (STDV) of differences between forecasts and independent observations, with a focus on near-surface temperature and humidity forecasts.
- Evaluate the spatial impact of the assimilation of surface observations on the forecasts, focusing on differences in forecast scores between mountains and adjacent plains.
- Identify the shortcomings of the evaluated system and propose solutions.

The remainder of this paper is organized as follows: Section 2 describes the components of the NWP chain, its design, the numerical experiments, and the verification methods. Section 3 presents the results of the numerical experiments. Section 4 discusses the results and summarizes the conclusions.

2 | Methods and Data

2.1 | WRF-ARW Model

The advanced research WRF (ARW) model (Skamarock and Klemp 2008) is a widely used community-based mesoscale NWP model designed to support operational weather forecasting and atmospheric research. It features a flexible, fully compressible, non-hydrostatic solver. WRF-ARW also provides various physical parameterization options to simulate processes related to cumulus convection, microphysics, planetary boundary layer (PBL), radiation, land-surface exchange, and gravity-wave drag. In this work, version 4.5 of the WRF-ARW model is used. Further details on the model's configuration and capabilities are provided in Skamarock et al. (2019).

The present study focuses on the European Alps and the surrounding regions (Figure 1a), in a domain with a horizontal grid spacing of 3.5 km and 38 hybrid terrain-following vertical levels, with the first half-model level located respectively at 15 m above ground level. This resolution is in the convection-permitting range and eliminates the need for a convection parameterization scheme. IC and BC for the WRF simulations were obtained from the ICON-EU system operated by the German Weather Service (DWD), using the WRF preprocessing system (WPS). The grid spacing of the input data is 7 km. The BCs were updated every 3 h. The following parameterization schemes were used in the experiments: the Thompson graupel scheme for microphysics (Thompson et al. 2008), the rapid radiative transfer model for general circulation models applications (RRTMG, Iacono et al. 2008) for the parameterization

of longwave and shortwave radiation components, and the Yonsei State University (YSU; Hong et al. 2006) as the PBL scheme. The land surface scheme is Noah (Tewari et al. 2004).

2.2 | WRFDA

The WRFDA system (Barker et al. 2004, 2012; Liu et al. 2020; Huang et al. 2009; Sun et al. 2020) was built upon the NCAR/Penn State University Mesoscale Model Version 5 (MM5) 3D-Var system. The 3D-Var algorithm provides the analysis x_a, which is the best estimate of a system's state obtained by optimally combining prior model information (i.e., the background) and observations, upon minimization of the following cost function:

$$J(x) = \frac{1}{2}\left(x - x_b\right)^T B^{-1}\left(x - x_b\right) + \frac{1}{2}(y - H(x))^T R^{-1}(y - H(x)) \tag{1}$$

where x_b is the first guess, or background, consisting of the model state from a previous forecast; y is the observation vector; and R and B are the observation and background error covariance matrices, respectively. H is the observation operator that computes the equivalent observation from the model's state. The difference $y - H\left(x_b\right)$ is referred to as the *innovation* vector, while the *analysis increment* $x_a - x_b$ is the difference between the analysis and the background.

The control variables for the minimization have been chosen following Xu (2019) and Sun et al. (2016) and are longitudinal and meridional wind components (u, v), temperature (T), surface pressure (p_s) and pseudo-relative humidity (RH_s); these correspond to the option CV7 in WRFDA.

The background error covariance matrix B is generated via the NMC method (Parrish and Derber 1992) using a climatology of forecast differences. The B used in the experiments was calculated using the NMC method implemented in WRFDA (GEN_BE routine), considering 1 month of 6-h forecast differences between 18- and 12-h forecasts valid at 00, 06, 12, and 18 UTC for the period between February 22 to March 23, 2024. The resulting B provides a climatological estimate of the background error, meaning it is not flow-dependent. The chosen B model is isentropic, horizontally homogeneous, and univariate. For a detailed description of the implementation of the NMC method in WRFDA, the reader is referred to Barker et al. (2004) and Sun et al. (2016). The B model obtained with the NMC method implemented in the GEN_BE routine was used in the assimilation system without further modification, which means that no tuning of the parameters of the B model (i.e., background error variances and horizontal background error correlation length scales) was performed for the assimilation experiments. These model parameters have been diagnosed using pseudo-observation tests (PSOTs) and are reported in Table 1. The procedure and a visual representation of the spatial impact of the assimilation of a single temperature increment are reported in Appendix A.

TABLE 1 | Parameters of the error covariance models for the 00 UTC and 12 UTC assimilation experiments.

Variable	σ_o	σ_b	Corr. Len. (km)
T	2 K	0.67 K	31
RH	10%	5.6%	22
p	200 Pa	29 Pa	310
u	1.1 $m\,s^{-1}$	1.4 $m\,s^{-1}$	20
v	1.1 $m\,s^{-1}$	1.5 $m\,s^{-1}$	19

Note: σ_o and σ_b are, respectively, the observation and background error standard deviations and Corr. Len. is the background error correlation length scale.

The observation operator chosen for the experiments calculates the model equivalent of the observations by selecting the neighboring grid point whose height most closely matches the observation height, thereby reducing the discrepancies due to elevation differences. All observations whose actual elevation differs from the model surface elevation by more than 400 m are discarded. This value is relatively large because we are also interested in including stations in the domain's mountainous regions, where, in general, the differences between observation elevation and model surface elevation are larger due to enhanced terrain variability that cannot be fully captured at the model's resolution. During the assimilation, a correction is applied to the observations to account for the elevation difference between observations and their equivalent in model space. For temperature observations, the correction is based on the method by Ruggiero et al. (1996), while for pressure observations, it is based on hydrostatic balance. Wind observations are corrected using similarity laws, while relative humidity observations are not corrected.

Further details on WRFDA's configuration and capabilities can be found in Barker et al. (2004), Barker et al. (2012), Liu et al. (2020), Huang et al. (2009), and Sun et al. (2020).

2.3 | Observational Datasets and Observation Preprocessing

Three different observational datasets were used in this work to compute the analyses and/or to verify forecasts.

- *GTS*: The global telecommunications system collects and distributes standard-coded global observations from several platforms, including SYNOP and METAR. Here, we used SYNOP observations in the assimilation experiments and METAR reports to verify forecasts. These observations were obtained from the NCAR Research Data Archive (RDA) at NCEP (2008).
- *Hypermeteo S.r.l. Database stations*: The Hypermeteo S.r.l. hosts and maintains a dataset of surface observations covering the Italian territory and surrounding regions. These observations are collected from a variety of sources, including both private and public surface weather stations, and are available hourly in near-real time (i.e., with less than 1-h delay). The observed variables are temperature, relative humidity, surface pressure, and wind. Database observations were used both in the assimilation experiments and for forecast verification.

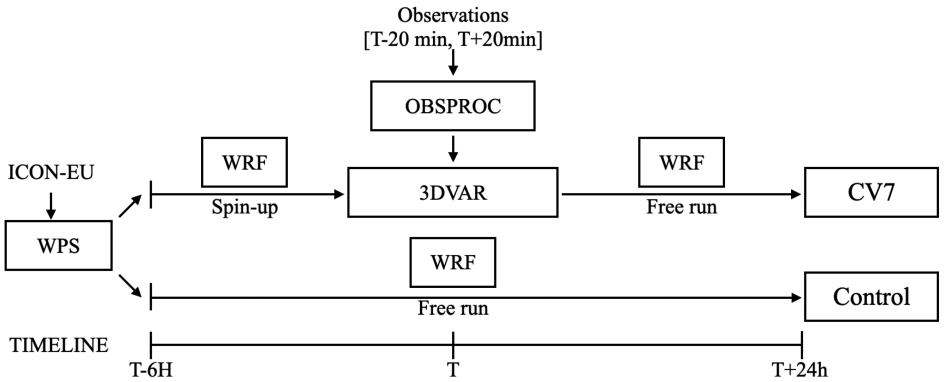

FIGURE 2 | Workflow for the forecasts initialized at time *T*. The experiments with or without the assimilation step are labeled, respectively, as "CV7" (upper branch) and "control" (lower branch).

Further information about the sources of the observations included in the Hypermeteo S.r.l. database is included in the Data Availability Statement.

- *IGRA radiosondes*: The Integrated Global Radiosonde Archive stores and distributes radiosonde observations worldwide (Durre et al. 2018). In this study, radiosonde observations were only used for forecast verification.

The WRFDA system preprocesses the observations for the assimilation and verification of the forecasts. The preprocessing of observations removes duplicates, flags data outside the domain, assigns the observation error, and prepares data for assimilation by WRFDA. The STDV of the observation errors is assigned according to the tabulated values of WRFDA, which are reported in Table 1.

Then, another level of preprocessing is carried out at the moment of the assimilation, the innovation check. In this study, observations are not assimilated if the absolute value of their innovations is more than five times the observation error. During the preprocessing, observations are thinned to ensure a minimum distance of 7 km between stations.

A total of 830 observational points were selected for DA: 631 from the Hypermeteo S.r.l. database and 199 GTS SYNOP stations. This selection was made in two steps. First, the stations providing observations of at least temperature, humidity, and pressure were selected. Then, considering preliminary DA experiments, a subset of these stations, which passed the internal quality checks by WRFDA and the observational thinning, was identified. This subset was then selected for the assimilation experiments. These two steps are intended to provide run-to-run consistency in the assimilated stations and to exclude from the assimilation stations those would be systematically rejected due to their height difference from the model's terrain. Their spatial distribution, presented in Figure 1a, reveals a higher density of stations in the plains than in the mountains.

The non-assimilated stations in the Hypermeteo S.r.l. database were used for verification. They were divided into two subsets: one with stations in the mountains and one in plain regions. Dividing the set of observations in this way can give us insight into the effect of the assimilation of surface observations in mountains and plains, and evaluate the skill difference in temperature forecasting in regions with different orographic complexity. The partitioning of the observations in the two subsets was based on their altitude above sea level and the variability of the model's orography in their close surroundings, quantified here by the STDV of the surface elevation of the four closest model points:

- *Mountain stations*: All stations at an elevation of at least 400 m above sea level and for which the STDV of the neighboring model points is larger than 100 m.
- *Plain stations*: All stations below 300 m above the sea level and for which the STDV of the neighboring model points is lower than 30 m.

Following this classification, the subset of mountain stations and plain stations used in the verification consisted of 250 and 300 observational points, respectively; the spatial distribution of the stations within the two subsets is shown in Figure 1b.

2.4 | Experimental Setup

The experiments were conducted over 1 month, from April 10 to May 10, 2024, with 24-h forecasts initialized twice daily at 00 and 12 UTC without cycling. The workflow of the developed setup is sketched in Figure 2. The DA process was initiated after a 6-h model spin-up period. For instance, a WRF simulation is initialized from the 18 UTC output of the ICON-EU model and provides a background field for the 00 UTC assimilation. To assess the impact of the assimilation of surface observations, control runs were conducted in parallel without the DA step.

The experiments are not cycled, meaning that the background used in successive assimilation steps does not originate from the previous model integration but from the ICON-EU parent model following the 6-h spin-up.

2.5 | Forecast Verification

The accuracy of the model runs was verified against independent observational datasets, including radiosondes, non-assimilated observations from the Hypermeteo S.r.l. database, and METAR reports from the GTS. The spatial distribution of these datasets in the domain is shown in Figure 1b.

The verification of the forecasts, based on the statistical analysis of the observation minus forecast (OMF) values, was carried out hourly.

The same filtering process used by WRFDA to screen the innovations before assimilation was also applied to the OMF populations to remove potential outliers. The filtering process was carried out independently for the two experiments (CV7 and control). Then, the two filtered OMF populations were made consistent with each other: If an observation's OMF value was rejected for one experiment, it was also excluded from the other. The same filtering procedure was carried out for the verification against METAR stations and non-assimilated observations from the database of Hypermeteo S.r.l.

Following the filtering, the output from the verification of every run consisted of a time series of OMF values for every verifying station and every variable. These time series of OMF values were then used to calculate domain-wide verification metrics such as the mean error (ME) and the STDV.

$$\mathrm{ME}(v,lt) = -\frac{1}{N_s N_r}\sum_{s=1}^{N_s}\sum_{r=1}^{N_r}\mathrm{OMF}(r,v,s,lt) \tag{2}$$

$$\mathrm{STDV}(v,lt) = \frac{1}{N_s}\sum_{s=1}^{N_s}\sqrt{\frac{1}{N_r}\sum_{r=1}^{N_r}\left[\mathrm{OMF}(r,v,s,lt) - \frac{1}{N_r}\sum_{r=1}^{N_r}\mathrm{OMF}(r,v,s,lt)\right]^2} \tag{3}$$

where the indices stand for run (r), lead time (lt), station (s), and variable (v). The number of runs and verification stations are N_r and N_s, respectively. The ME indicates the average bias for the OMF population at a specific lead time and for a particular variable, quantifying the average component of the forecasting error. The minus sign in Equation (2) was added to ease the interpretation of the results so that, for example, a positive temperature ME arises from consistently warmer forecasts than observations. The STDV measures the pattern difference in the OMF population, representing the random component of the forecast error.

We performed hypothesis testing to evaluate the statistical significance of differences in domain-wide verification metrics between the CV7 and control experiments. Our null hypothesis was that there is no difference between the values of verification metrics of the control and CV7 experiments. We computed confidence intervals for the differences between verification metrics using block bootstrapping (Hamill 1999), where the dataset was resampled in blocks of 3 days and considered the 2.5 and 97.5 percentiles of the bootstrapped distribution. The null hypothesis was rejected at the 95% confidence level when the confidence interval did not include the zero value. In such cases, the skills of the control and CV7 experiments were deemed significantly different.

The root mean square (RMS) of OMF differences is used to quantify the station-specific impact of the assimilation step, accounting for its contribution to both random and systematic components of the forecast error. The RMS difference ($\Delta\mathrm{RMS}_{N_{lt}}$) over a range of N_{lt} hours of lead time was calculated as follows:

$$\mathrm{RMS}(v,s,lt) = \sqrt{\frac{1}{N_r}\sum_{r=1}^{N_r}\left[\mathrm{OMF}(r,v,s,lt)\right]^2} \tag{4}$$

$$\Delta\mathrm{RMS}_{N_{lt}}(v,s) = 100\times\frac{\frac{1}{N_{lt}}\sum_{lt=1}^{N_{lt}}\left(\mathrm{RMS}_{\mathrm{CV7}} - \mathrm{RMS}_{\mathrm{Control}}\right)}{\frac{1}{N_{lt}}\sum_{lt=1}^{N_{lt}}\mathrm{RMS}_{\mathrm{CV7}}} \tag{5}$$

3 | Results

3.1 | Characterization of Assimilated and Verifying Surface Observations

As we are using surface observations from various conventional and nonconventional sources, a comparative evaluation of their quality is conducted, considering the distributions of the observation minus background (OMB; innovations) values and their RMS. These are represented in Figure 3.

In general, for most variables and in both forecast runs, the systematic deviations between the background and different observational datasets (judged by the median of the boxplots in Figure 3) differ only marginally. The same applies to the variability of OMB departures.

Few notable differences and features of the OMB distributions are listed below.

- *Temperature*: RMS values are slightly larger at 00 UTC than at 12 UTC across datasets. At 12 UTC, Hypermeteo S.r.l. database observations show a slightly positive median, while GTS observations exhibit a negative or near-zero median, indicating a different bias relative to the background (Figure 3a,b).

  In terrain-separated subsets, at 12 UTC, OMB values in plains (DBVP) show less spread and a smaller positive bias compared to OMB values in the mountains (DBVM).

- *Specific humidity*: OMB distributions are narrower at 00 UTC than at 12 UTC. All datasets show positive medians, slightly larger at 12 UTC. Hypermeteo S.r.l. database observations exhibit greater variability in OMB than conventional observations (Figure 3c,d).

- *Pressure*: METAR OMBs have the narrowest distribution (both at 00 and 12 UTC), while SYNOP OMBs have the largest variability among assimilated data. DBA OMB values are more narrowly distributed than SYNOP, but more than METAR. DBV RMS values are significantly larger than DBA. This difference could be due to the observation preselection procedure; however, it was not investigated further in this work since DBV pressure observations have not been used for assimilation or verification in the current experiments (Figure 3e,f).

Overall, the distributions of OMB values from conventional stations and the stations in the Hypermeteo S.r.l database used in this work, either for assimilation or verification, present some limited differences. Still, the overall quality (intended here as the fit between the observations and the background)

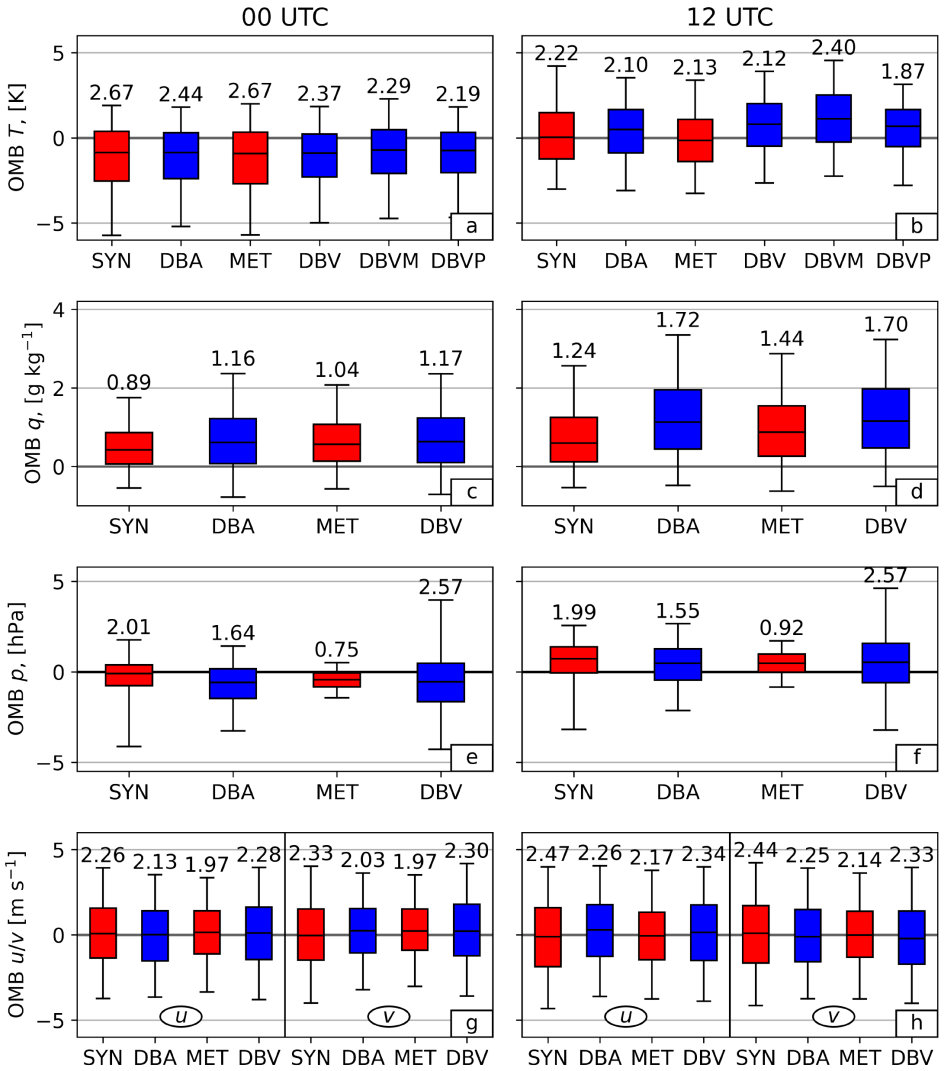

FIGURE 3 | Boxplots showing the distribution of OMB (observation minus background) values for temperature (T), specific humidity (q), pressure (p), and wind components (u/v) displayed, respectively, in panels a, b; c, d; e, f; and g, h. Panels a, c, e, and g correspond to the 00 UTC experiments, while Panels b, d, f, and h represent the 12 UTC experiments. Blue boxes indicate observations from the Hypermeteo S.r.l. database, and red boxes represent SYNOP (SYN) and METAR (MET) observations. Within the Hypermeteo S.r.l. dataset, DBA refers to assimilated stations, DBV to verifying observations, and for temperature, DBVM and DBVP denote verifying observations located in mountainous and plain regions, respectively. Each box spans the interquartile range (25th–75th percentiles), with whiskers extending to the 5th and 95th percentiles. The population RMS is annotated at the top of each upper whisker.

is comparable among the populations. This also indicates that the preselection procedure implemented in the definition of the assimilated and non-assimilated observations does not generate significantly different populations apart from the case of pressure.

3.2 | DA Diagnostics

3.2.1 | Observation Space Diagnostics

The number of observations assimilated in the experiments, shown in Figure 4, exhibits some fluctuations, with some days experiencing more severe variations than others. These fluctuations are likely due to issues in the collection and processing of data on those specific days. The percentage of available observations passing the quality check is high, suggesting that only a few outliers exist in the observational dataset. The number of assimilated stations is higher in the 12 UTC experiments compared to the 00 UTC experiments. However, the difference is slight, with both cases averaging around 3000 surface observations of pressure p, temperature T, specific humidity q, and wind speed components u and v.

A diagnostic check is presented in Table 2, based on the averages and STDVs of the populations of OMB and observation minus analysis (OMA; analysis residuals) values obtained from the assimilation steps over the month of CV7 experiments. The primary objective of DA is to minimize random discrepancies between observations and the model state. This improvement is reflected in the STDVs of the OMB and OMA values shown in the first columns of Table 2. The STDVs

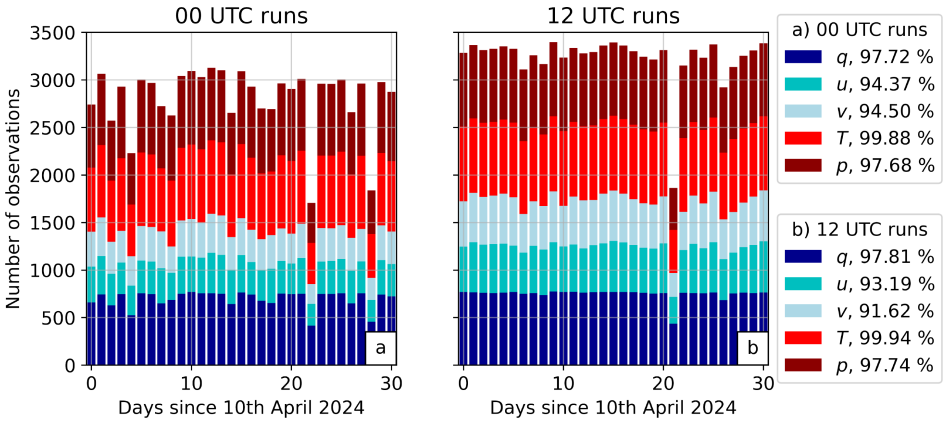

FIGURE 4 | Assimilated observations in 00 UTC (Panel a) and 12 UTC (Panel b) runs. The average fraction of observations passing the quality check is reported in the legends on the right.

TABLE 2 | Mean and standard deviations of the innovations (OMB) and analysis residuals (OMA) for T, u, v, q, and p over the 00 and 12 UTC assimilation experiments.

		Standard deviation		Mean	
		00 UTC	12 UTC	00 UTC	12 UTC
T (K)	OMB	1.78	1.64	−1.19	0.36
	OMA	1.21	1.03	−0.41	0.09
u ($m s^{-1}$)	OMB	1.81	2.07	0.02	0.12
	OMA	0.91	1.11	0.004	−0.001
v ($m s^{-1}$)	OMB	1.76	2.03	0.18	−0.01
	OMA	0.87	1.04	0.01	−0.002
q ($g kg^{-1}$)	OMB	0.76	0.90	0.61	1.13
	OMA	0.54	0.64	0.34	0.54
p (Pa)	OMB	51.25	60.84	−65.65	37.53
	OMA	40.48	50.67	−7.94	4.57

of OMB values consistently exceeds those of OMA values, demonstrating that the implemented DA step accomplished its main goal.

Ideally, the average of the OMB and OMA values should be zero. A nonzero average of OMB values indicates that the innovations are biased, which could arise from either a bias in the observations or the model. Biased innovations, when assimilated, inevitably lead to systematic adjustments to the model state, or equivalently, biased analysis increments. In the same way, a nonzero average of OMA values indicates that either the analysis or the observations are biased. As shown in the last columns of Table 2, there are indeed biases in the innovations, especially for T, q, and p. Biases are a known complex issue in the assimilation of surface observations (Sgoff et al. 2022; Ancell et al. 2011) that cannot be addressed by the standard quality checks implemented in WRFDA. The differences between the means of OMB and the means of OMA for T, p, and q show that, besides, DA is effective at reducing bias.

3.2.2 | Average Analysis Increments

Figures 5 and 6 show the average analysis increments for temperature (Panels a and b) and humidity (Panels c and d) in the model state space. Visualizing these average analysis increments shows the systematic impact of DA of surface observations on the model state, helping to interpret the effect of the assimilation on the forecast scores in the next section. A nonzero average analysis increment is yet another symptom of bias in the model or the observations.

For temperature, the assimilation leads to a widespread cooling effect of the lowermost atmospheric levels during the night (Figure 5a,b) and a heating effect during the day (Figure 6a,b), which means that the assimilation of observations tends to increase the amplitude of the diurnal cycle of T. In both the 00 and 12 UTC experiments, there is a noticeable increase in moisture in the same regions (Figures 5c,d and 6c,d).

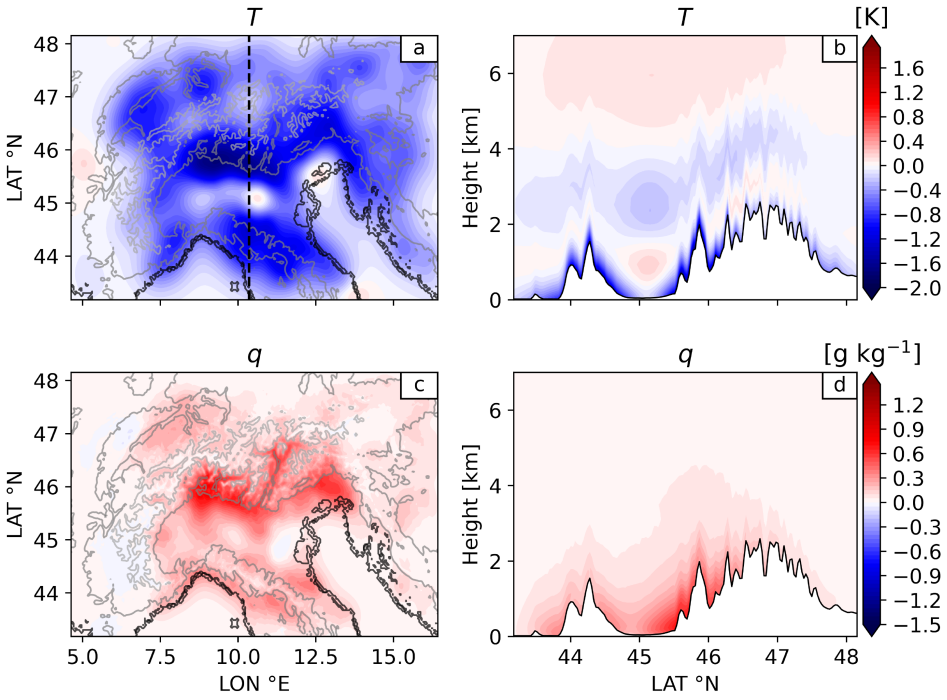

FIGURE 5 | Average analysis increments at the first model level from the 00 UTC experiments for temperature (Panel a) and humidity (Panel c). Panels b and d display vertical cross-sections of the analysis increments for the same variables, taken along the dashed black line shown in Panel a.

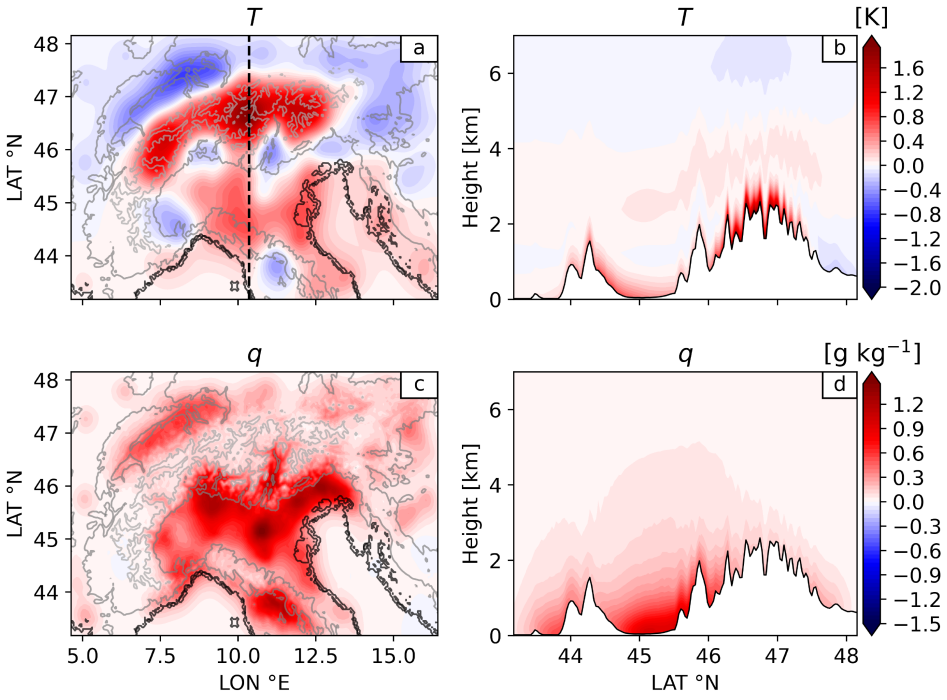

FIGURE 6 | Average analysis increments at the first model level from the 12 UTC experiments for temperature (Panel a) and humidity (Panel c). Panels b and d display vertical cross-sections of the analysis increments for the same variables, taken along the dashed black line shown in Panel a.

The vertical cross-sections (Figures 5b,d and 6b,d) show that the average analysis increments follow the terrain-conforming model levels rather than adapting vertically to orographic variations, which are not ideal for capturing fine-scale atmospheric dynamics in mountainous terrain.

The domain-averaged analysis increments at the first model level for temperature and specific humidity during the experimental period are presented in Figure 7. While for specific humidity (Panel b), the average analysis increment shows a clear positive bias both at 00 and 12 UTC, for temperature (Panel a), the bias

signal for 00 UTC runs is superimposed on a pronounced day-to-day variability. The domain-averaged temperature analysis increments are negative for all experimental days except 3, indicating that the average analysis increment is consistently biased. A clear, domain-wide bias in temperature analysis increments is not visible for 12 UTC runs.

To summarize, the average analysis increments exhibit biases and fail to capture orographic variability adequately, suggesting that the background error model used in the experiments is not optimal for effectively assimilating observations in regions with complex terrain. However, despite these limitations, observations are being successfully assimilated into the model. The critical question is whether these analysis increments ultimately translate into improved forecasting skills.

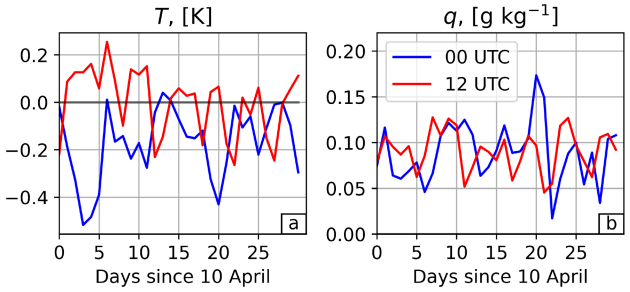

FIGURE 7 | Domain-averaged analysis increment at the first model level for temperature (Panel a) and specific humidity (Panel b) for the entire experimental period.

3.3 | Verification of the Experiments Against METAR Reports

First, the forecasts are verified against surface observations of an independent observing network that were not assimilated. We considered METAR reports available in the forecasting domain to show the beneficial effect of the assimilation of surface observations on the forecasts of surface temperature and humidity. On average, 70–80 observations per variable were utilized in the hourly verification of the forecasts. The metrics of the forecasts' verification against METARs are presented in Figure 8a,b (STDV) and Figure 8c,d (ME) as a function of the forecast lead time.

Figure 8 shows that the CV7 experiments significantly outperform the Control experiments, as the majority of dots indicating better skill and a statistically significant difference between experiments lie on the curves corresponding to the CV7 experiments. A notable characteristic observed across all variables and all experiments is the strong diurnal cycle of the verification metrics. More detailed discussion follows separately for temperature and humidity in the next two subsections.

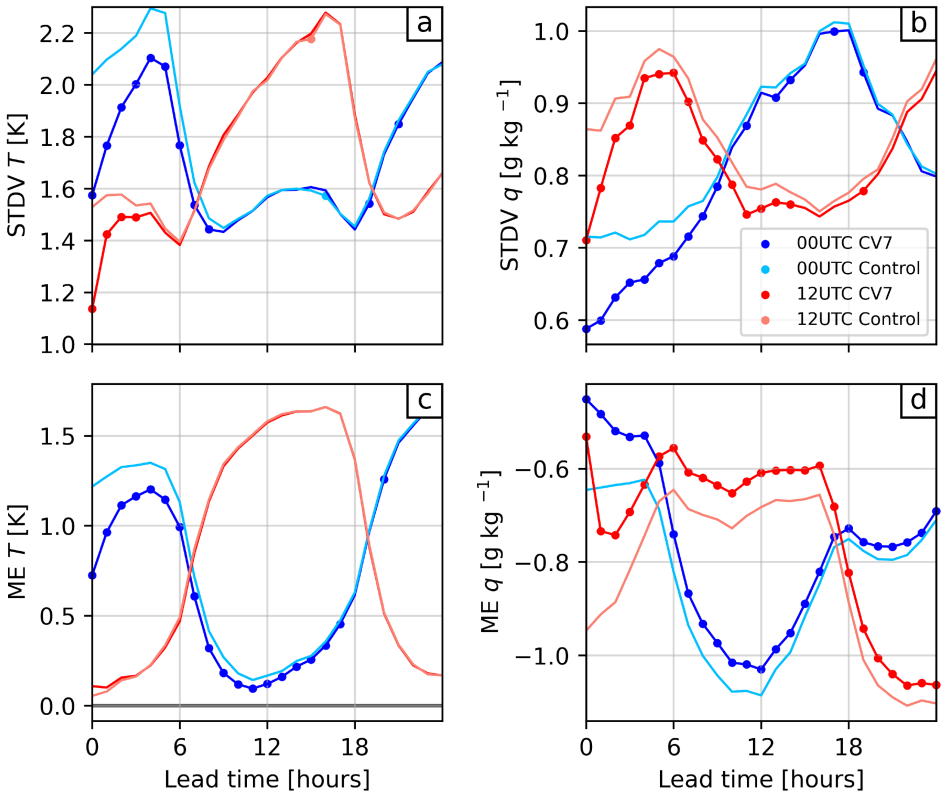

FIGURE 8 | Verification against METAR observations. In the upper row, the STDV is reported for temperature (a) and specific humidity (b). In the lower row, the ME is reported for the same variables (c and d). The dots on the curves indicate that the score difference between the CV7 and control experiments differ significantly from zero at the 95% confidence level. The dots are reported only on the curve of the experiment that performs better.

3.3.1 | Temperature

The temperature STDV in the CV7 experiments is reduced compared to the control experiment in both the 00 and 12 UTC runs (Figure 8a), with an average STDV reduction over the first six forecast hours of 0.26 K (12.6%) and 0.27 K (7.8%), respectively. The significant, positive impact lasts up to 8 h in the 00 UTC run and 4 h in the 12 UTC run. This STDV reduction indicates that the assimilation step successfully decreases the random component of model error. As shown in Figure 8c, the assimilation step also reduces the temperature ME. In the 00 UTC experiments, a statistically significant reduction in ME is observed for up to 18 h of lead time. The improvement is substantial in the first 6 h, where the ME in the CV7 experiments is reduced by an average of 0.26 K with respect to the Control experiments. In contrast, the 12 UTC experiments show no significant differences in temperature ME between CV7 and Control experiments (0.01 K difference).

These results suggest that DA has a more pronounced and longer-lasting impact at 00 UTC than at 12 UTC. The 00 UTC temperature innovations suffer from a nocturnal temperature bias that could result from multiple factors. Regarding the background, its limited vertical resolution could lead to poorly resolved near-surface temperature inversions, PBL parameterization inadequately representing stable boundary layers (García-Díez et al. 2013), and inaccuracies in surface properties such as soil state. These limitations result in weaker simulated temperature inversions than observations, causing negative temperature biases in the innovations. At the same time, innovation bias could also arise from observations. However, without additional information that can be considered unbiased with respect to the true state of the atmosphere (i.e., anchor observations, Eyre (2016)), it is not possible to evaluate whether the innovation bias is entirely model-dependent or instead partly due to observation bias. As previously shown (Figure 5), when such biased innovations are assimilated, DA efficiently performs a bias correction of the model state, resulting in a large impact of the assimilation step on the forecast scores in 00 UTC experiments. Concurrently, during the nocturnal hours, there is very limited mixing in the boundary layer, so the near-surface correction persists longer than during the day.

Another at least partially relevant factor in the pronounced innovation bias observed in our experiments is related to representativeness problems. A DA system can only reliably incorporate observations that the model can adequately represent, which is only partly true for surface observations, especially over complex terrain. When observations reflect processes, the model cannot fully resolve the assimilation shifts the model state towards the observations in a manner compatible with the model's parameterizations rather than with the actual observed phenomena. For example, assimilating surface observations collected in the presence of nocturnal temperature inversions, which the model poorly represents, induces cooling in the lowest levels. However, since the inversion structure in the model differs significantly from the real one, the analysis increments adjust the model state even at levels where no correction would be necessary. Essentially, representativeness errors combined with inaccurate spatial covariances make it harder to correct model bias.

Overall, the assimilation of surface observations in the present setup has beneficial effects on near-surface temperature forecasts, especially on the 00 UTC runs, consisting of a reduction of both the systematic and the random components of the differences between observations and forecasts. However, for the 12 UTC runs, the beneficial effect is limited to the random component.

3.3.2 | Humidity

For near-surface specific humidity forecasts (Figure 8b,d), the assimilation of surface observations has a consistent and positive impact in both the 00 UTC and 12 UTC runs. The STDV of OMF differences is significantly smaller in the CV7 experiments compared to the Control experiments, up to a lead time of 10 h. Over the first 6 h, the average reduction in STDV for the CV7 experiments is 0.09 g kg^{-1} (11.9%) in the 00 UTC runs and 0.06 g kg^{-1} (7.2%) in the 12 UTC runs, relative to the control experiments. The ME of OMF differences is consistently smaller in absolute value for the CV7 experiments than for the control ones across all lead times. During the first six forecast hours, the average ME reduction between the control and the CV7 experiments is 0.13 g kg^{-1} for the 00 UTC and 0.18 g kg^{-1} for the 12 UTC experiments. These results demonstrate that the assimilation step improves the near-surface specific humidity field in the model's ICs, according to METAR reports.

The significant reduction in specific humidity ME after the assimilation suggests that the observed humidity is systematically higher than the simulated humidity (Table 2), which implies systematically positive analysis increments (Figures 5 and 6).

Two factors could contribute to moisture biases in the analysis increments and innovations. First, as for temperature, the simulated near-surface specific humidity may not accurately represent the observed one near the surface (García-Díez et al. 2013) because of limitations in the model physics or the representation of surface properties, such as incorrect PBL mixing, surface fluxes, or soil moisture specification. Second, the moisture bias may to some extent depend on the moisture control variable, that is, the pseudo-relative humidity. This is defined as:

$$\mathrm{RH}^{*} = \frac{q}{q_{\mathrm{s}}\left(T_{\mathrm{b}}\right)} \tag{6}$$

where q is specific humidity and $q_{\mathrm{s}}\left(T_{\mathrm{b}}\right)$ is the saturation specific humidity with respect to the background temperature T_{b}. Since RH^{*} depends on T_{b}, biases in the background temperature directly impact moisture analyses. For instance, a nocturnal warm bias in the model leads to a positively biased saturation specific humidity, which in turn causes pseudo-relative humidity innovations to exhibit negative biases. These negative biases in pseudo-relative humidity innovations, which would typically suggest a model state that is too dry, result in positively biased specific humidity analysis increments. This relationship is consistent with Figure 5. As for the case of temperature, it is not possible to exclude that moisture innovation biases also contain a component of observational bias.

Overall, CV7 experiments outperformed the control runs across most metrics and variables. The positive impacts of DA were particularly evident in the temperature and humidity forecasts, where significant improvements were observed. The results

highlight the potential added value for NWP following the assimilation of surface observations in near-real time.

3.4 | Verification Against Radiosondes

To evaluate the impact of the assimilation of surface observations on upper air analyses and forecasts, we consider the verification of temperature and specific humidity forecasts against radiosonde observations at 00 and 12 UTC in the Po Valley (Figure 1b), drawn from the IGRA dataset.

3.4.1 | STDV

In Figure 9, the STDV of the OMF values between soundings and forecasts are reported for mandatory pressure levels up to 400 hPa. For the 00 UTC experiments (Figure 9a,b), the CV7 and control STDV show that the assimilation of surface observations does not significantly impact the forecasts, suggesting that the assimilation neither improved nor degraded the low-level specific humidity and temperature in the model. In the 12 UTC experiments (Figure 9c,d), the CV7 STDV for temperature and specific humidity is slightly but significantly smaller than that of the Control experiment at analysis time up to the 850 hPa level (solid lines). At 1000 and 925 hPa, the temperature STDV difference between 12 UTC CV7 and control experiments is, respectively, 0.19 and 0.08 K; the differences for specific humidity are 0.15 and 0.08 g kg^{-1}. This indicates a beneficial effect of the assimilation of surface observations on the temperature and specific humidity analyses in the PBL for the 12 UTC experiments. Furthermore, for all experiments, at +12 h (dashed lines), surface observation assimilation had no discernible effect on the upper-air temperature and humidity STDV profiles for all the considered experiments.

3.4.2 | ME

Considering instead the ME between soundings and forecasts, presented in Figure 10, the assimilation of surface observations had a mixed effect on temperature and humidity profiles in the model state. For the 00 UTC experiments at analysis time (Figure 10a, solid lines), the temperature ME for the CV7 experiment is decreased by 0.65 K at the 1000 hPa level when compared to the control ME; at this vertical level, the CV7 experiments are consistently colder than the sounding observations. This figure is comparable to the 1 K difference in temperature ME between the control and the CV7 experiments for the METAR observations corresponding to the sounding locations, represented in Figure 10a, as the full dots on the horizontal axis. This indicates that while the assimilation of surface observations reduced the mean differences between surface observations and the model state, it also introduced a temperature bias between soundings and forecasts in the PBL. Above 925 hPa, the differences in ME between CV7 and control experiments are small

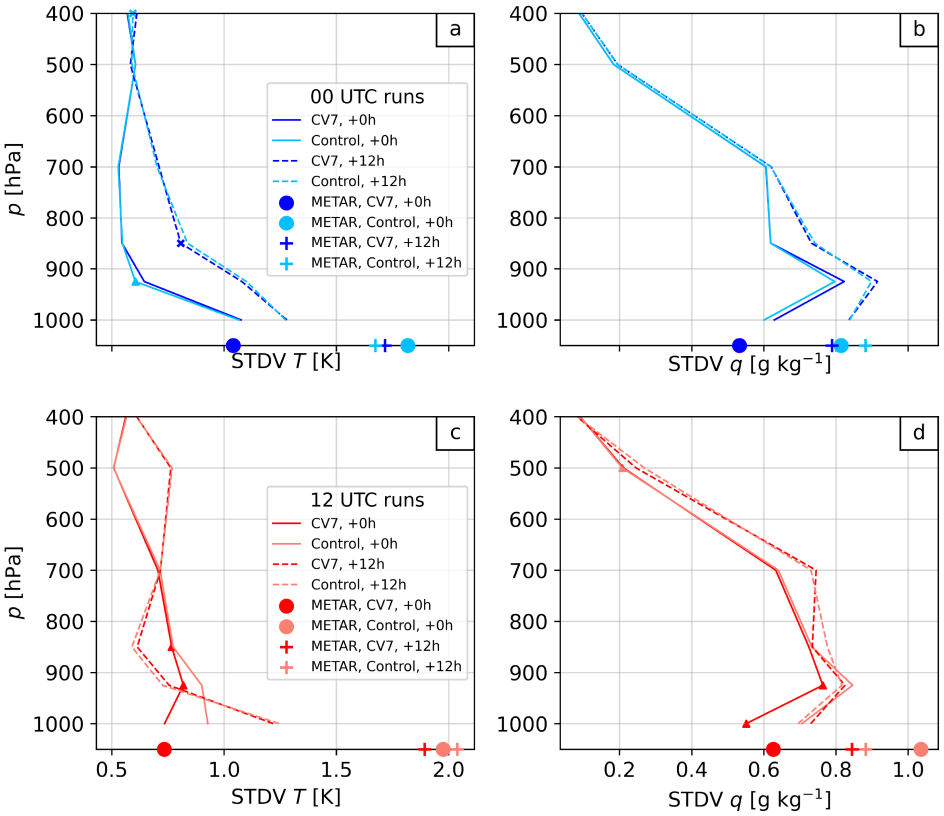

FIGURE 9 | Forecast verification against radiosondes. STDV is shown for the 00 UTC experiments (Panels a and b) and 12 UTC experiments (Panels c and d) for temperature T (Panels a and c) and specific humidity q (Panels b and d). Significant differences between experiments are marked on the curves of the better-performing experiment with triangles for analysis time and crosses for +12 h. Markers on the x-axis indicate the STDV of the OMF values for surface METAR observations at the sounding sites, at analysis time (bullets) and +12 h (crosses).

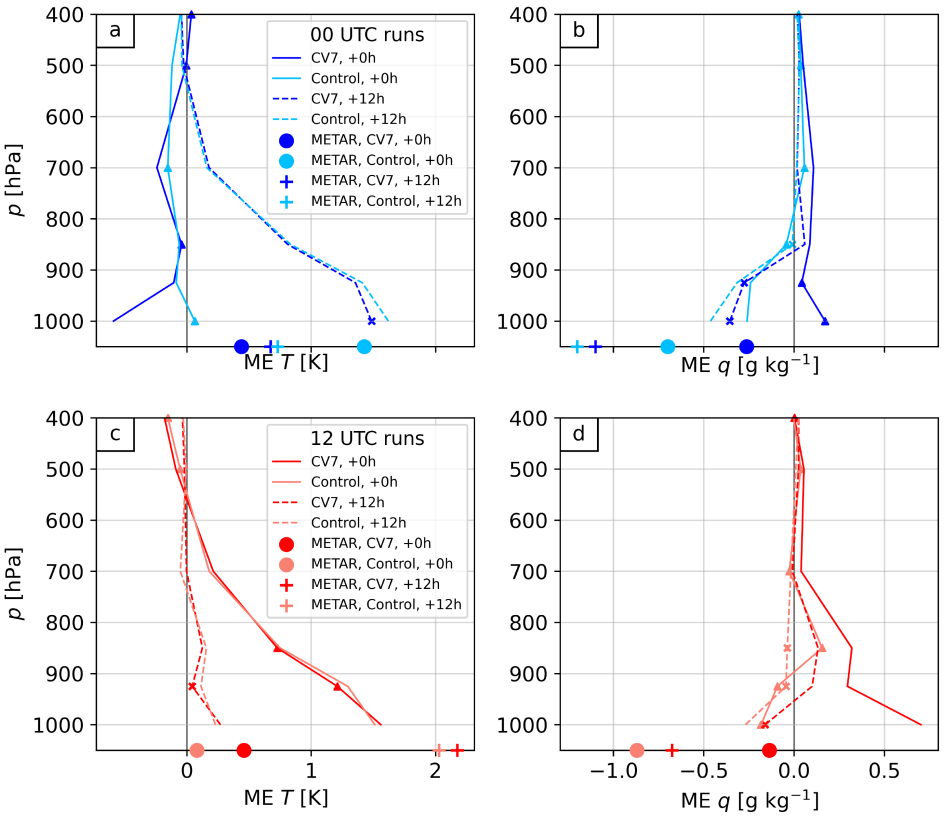

FIGURE 10 | Forecast verification against radiosondes. ME is shown for the 00 UTC experiments (Panels a and b) and 12 UTC experiments (Panels c and d) for temperature T (Panels a and c) and specific humidity q (Panels b and d). Significant differences between experiments are marked on the curves of the better-performing experiment with triangles for analysis time and crosses for +12 h. Markers on the x-axis indicate the ME of the OMF values for surface METAR observations at the sounding sites, at analysis time (bullets) and +12 h (crosses).

(i.e., below 0.1 K) but significant, meaning that the assimilation of surface observations marginally propagates also to the upper-air temperature fields. These small differences are consistent with the vertical structure of the average temperature analysis increment structure within the Po Valley (located at latitudes around 45° N) reported in Figure 5b. At 12 h of lead time (dashed lines), the assimilation had a small but significant positive effect on temperature ME at the 1000 hPa level, with a ME reduction of 0.14 K.

The 12 UTC experiments presented no significant and relevant temperature ME differences (i.e., all below 0.07 K) between CV7 and control experiments (Figure 10c). This is consistent with the results of the verification of the experiments against METAR observations, where little to no difference in T ME was observed between the CV7 and control for the 12 UTC experiments.

The results for specific humidity ME between soundings and forecasts exhibit a marked effect of the assimilation on the vertical distribution of this variable. In the 00 UTC CV7 experiments, specific humidity was better captured below 850 hPa than in the control (Figure 10b). In the 00 UTC experiments at analysis time, CV7 showed MEs of 0.17 g kg^{-1} (1000 hPa) and 0.04 g kg^{-1} (925 hPa), while the control had −0.26 and −0.24 g kg^{-1}, indicating an improved moisture representation in CV7. However, slight degradation occurred at 850 and 700 hPa in CV7. This improvement persisted at 12-h lead time (00 UTC), with smaller MEs in CV7 below 850 hPa, though minor degradation remained at 850 hPa.

For the 12 UTC runs (Figure 10d), CV7 showed larger MEs in absolute value below 700 hPa compared to control, indicating a degradation in moisture fields: At analysis time (12 UTC), CV7 had MEs of 0.7 g kg^{-1} (1000 hPa) and 0.3 g kg^{-1} (850 hPa) versus −0.18 g kg^{-1} and −0.16 g kg^{-1} for control. At the 12-h lead time, control consistently showed lower MEs in absolute value below 700 hPa, except at 1000 hPa, where CV7 performed slightly better.

The difference in analysis time at 1000 hPa in specific humidity ME between CV7 and Control experiments is 0.43 g kg^{-1} for the 00 UTC runs, and 0.89 g kg^{-1} for the 12 UTC runs; these figures are comparable with the difference in ME for the METAR stations corresponding to the sounding at 00 UTC (0.44 g kg^{-1}) and at 12 UTC (0.75 g kg^{-1}). As in the case of temperature, for specific humidity, the beneficial reduction of the mean differences between surface observations and the model state provided by DA introduced biases between soundings and forecasts in the PBL.

Overall, the assimilation of surface observations has a mixed impact on the atmospheric temperature and humidity fields above the surface. The effects of the assimilation step on the STDVs of the temperature and humidity differences between forecasts and soundings are either null or slightly beneficial. The same is not valid for the averages of these differences, which show either a null or slightly negative impact of the DA step.

The suboptimal effects of the assimilation step on the ME of specific humidity and temperature OMF values in the lower

troposphere are likely due to the presence of either model or observation biases in the temperature and humidity innovations from surface observations. As shown in Figures 5 and 6, the assimilation of biased innovations from surface observations leads to systematic analysis increments, which propagate from the surface up to the mid-troposphere with the largest impact in the PBL (see Figures 9 and 10), eventually adding systematic differences in the ME between soundings and model analyses and forecasts. Another possible reason for the observed degradation of the ME is a suboptimal specification of the background error covariance model, which influences the spatial structure of analysis increments.

3.5 | Skill Difference Between Mountains and Plains

We now consider the verification of the forecasts against temperature observations in Hypermeteo S.r.l. database that were not included in the assimilation. These observations provide a dense and homogeneous coverage of the area of interest, particularly over Italy, as shown in Figure 1b.

The results from the verification of temperature forecasts against the sets of non-assimilated observations and the subset from mountains and plains observations, respectively, are shown in Figure 11 as a function of the forecast lead time.

The impact of the assimilation step on OMF STDV for the full dataset (Figure 11a) aligns with the results presented in Section 3.3, showing a generally positive and significant effect, particularly in the first six forecast hours. In this forecast window, the average reduction in STDV between the Control and the CV7 experiments is 0.21 K (12.5%) for the 00 UTC runs and 0.16 K (9.9%) for the 12 UTC runs. The main distinction from the METAR verification (Section 3.3) is a small but statistically significant degradation in STDV between 9 and 13 h of lead time for the 00 UTC forecasts. This degradation appears in the verification metrics against the full dataset and plain station (Figure 11a,b) but not in the verification against mountain stations (Figure 11c). However, the degradation is rather small, with a maximum STDV difference of only 0.03 K (2%).

The ME of OMF values considering the full dataset, shown in Figure 11d, differs significantly from that obtained in the verification against METARs. While the ME of METAR OMF values ranged 1.65–0.1 K, the ME of non-assimilated Hypermeteo S.r.l. database observations OMF ranges from 1.7 K to −0.6 K, indicating a more pronounced diurnal cycle in the ME according to this dataset; this is driven mainly by a larger daytime bias in the verification against non-assimilated Hypermeteo S.r.l. database observations, which was previously observed in Figure 3b.

The verification against non-assimilated Hypermeteo S.r.l. database observations reveals that the assimilation step results in an improvement in the temperature OMF ME (average ME reduction of 0.3 K) in the first 6 h for the 00 UTC experiments and a smaller effect for the 12 UTC runs (0.04 K). This improvement is consistent with the one obtained in the verification against METARs. As seen with STDV, the 00 UTC experiments show a small but significant degradation in ME between 9 and 13 h of lead time, particularly in the plains (Figure 11e), where the maximum degradation reaches 0.16 K at +9 h. This suggests that the assimilation of surface observations at 00 UTC may be cooling the boundary layer too much,

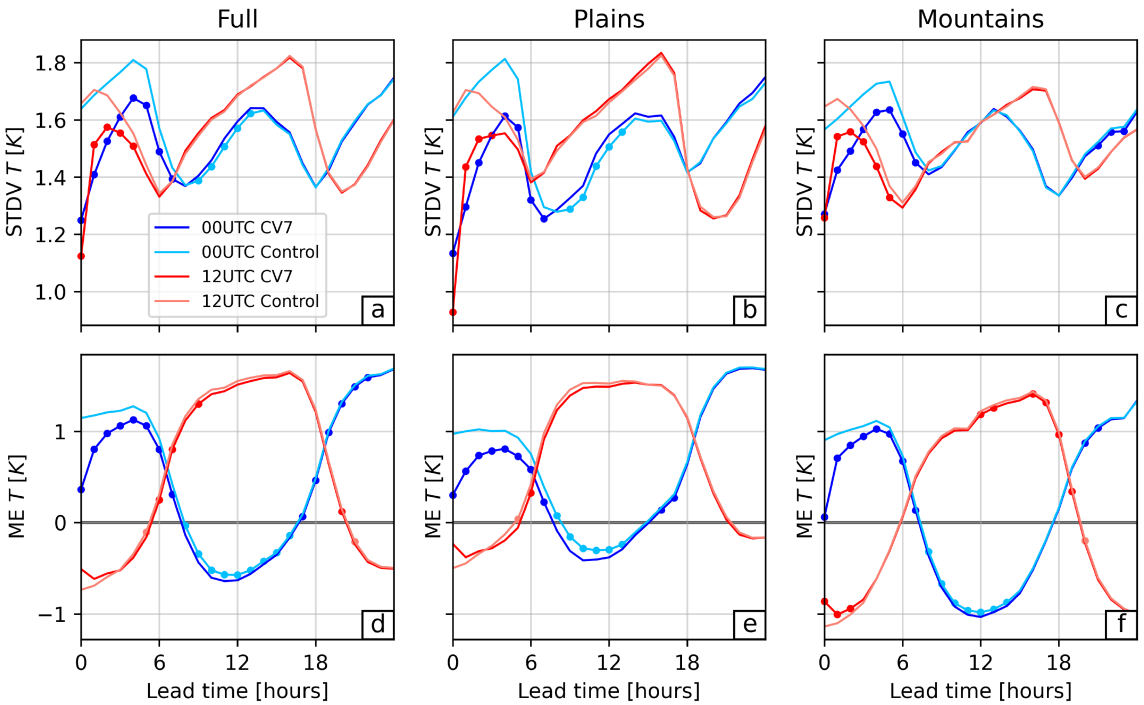

FIGURE 11 | Verification against database temperature observations. Upper panels: STDV for the full set of observations (a), observations in the plains (b), and the mountains (c). Lower panels: ME for the same observation sets (d–f). The dots on the curves indicate that the difference between the experiments differs significantly from zero at the 95% confidence level. The dots are reported only on the curve of the experiment that performs better.

especially in the Po Valley, where most of the assimilated stations and verifying plain stations are located (Figure 1a,b). This is also indicated by the negatively biased average analysis increment in the Po Valley presented in Figure 5a.

Another possible reason that could be responsible for this ME degradation is a different OMF bias for METAR and Hypermeteo S.r.l. database observations during daytime. This is substantiated by the discussion in Section 3.1, in which the different medians of the distribution of temperature OMB values for conventional and nonconventional observations are noted: while at 12 UTC, METAR observations have a slightly negative bias, Hypermeteo S.r.l. database observations have a (small) positive bias. The cooling caused by the assimilation at 00 UTC persists in the forecasts and is beneficial for the ME between forecasts and METAR observations since it corrects the model consistently with the (negative) OMF bias of METAR observations at 12 UTC. On the other hand, this cooling takes the model state further away from verifying Hypermeteo S.r.l. database observations at 12 UTC, which are, on average, warmer than the model state.

The OMF STDV varies by terrain and time of day. This can be observed considering the blue curves relative to the 00 UTC runs, shown in Figure 11b,c, for which the lead time matches the time of day. Later at night and in the morning, the STDV is lower in the plains compared to the mountains, while during other times of the day, it is slightly smaller in the mountains. However, the STDV differences between the two regions remain relatively small, below 0.2 K in absolute terms. The ME also exhibits a diurnal pattern. Considering the blue curves relative to the 00 UTC runs, shown in Figure 11e,f, for which the lead time matches the time of day, it is possible to see more pronounced ME differences between mountains and plains during the daytime (06–18 UTC). During this period, the ME in the mountains is up to 0.8 K higher in absolute value (more negative) than in the plains. This might indicate that the model seemingly underestimates diurnal warming in the mountains more than over the plains. In contrast, in the early nighttime hours (18 UTC–00 UTC), the plains have a larger ME by up to 0.6 K in absolute value (more positive). Between 00 and 06 UTC, the ME differences between the mountains and plains are smaller, remaining below 0.2 K. The overall amplitude of the daily cycle of ME is more pronounced in the mountains (2.5 K) than in the plains (2 K), driven by more negative daytime ME in the mountains. The nighttime maximum ME is comparable between the two regions, around 1.5 K. The above considerations are also valid for the 12 UTC runs upon a 12-h shift of the verification metrics.

To better visualize the impact of the assimilation step on STDV in the mountains versus plains, the differences in STDV in the mountains and plains are reported in Figure 12a: the assimilation of surface observations reduces the STDV more significantly in the plains for both 00 UTC and 12 UTC experiments. Thanks to the assimilation step, the STDV in the mountains decreases in the first six forecast hours on average by 9.7% and 7.8% for the 00 and 12 UTC runs, respectively, while in the plains it decreases by 17% and 13.3%. The beneficial effect of the assimilation on STDV persists for a comparable lead time in both regions, lasting 6 h in the 12 UTC runs and 8 h in the 00 UTC runs. The larger reduction of STDV in the plains than in the mountains indicates that the assimilation step is more effective in the plains than in the mountains.

The effect of the assimilation step on the ME in different terrains and for different initialization times is shown in Figure 12b. As previously discussed, the effect of DA on the ME for the 12 UTC experiments is minimal. Concerning the 00 UTC experiments, the assimilation has a slightly higher impact at analysis time in the mountains, with a 0.84 K ME reduction at analysis time, than in the plains, where a reduction of 0.67 K is observed. The positive impact decreases with lead time, and in about 6 h, it is only 25% of the initial impact in the plains and 7% in the mountains. In the plains, the reduction of the ME is slightly lower at analysis time but persists more with forecast lead time and is not beneficial for all lead times (blue lines, Figure 11d). As discussed above, the degradation of the temperature ME in the plains for lead times beyond 7 h could be caused by the pronounced cooling of the PBL by the assimilation of surface observations at 00 UTC or by a different bias in the verifying observations.

Overall, the more persistent and more pronounced effect observed in the plains with respect to the plains could be driven by at least two aspects. First, the density of the assimilated stations

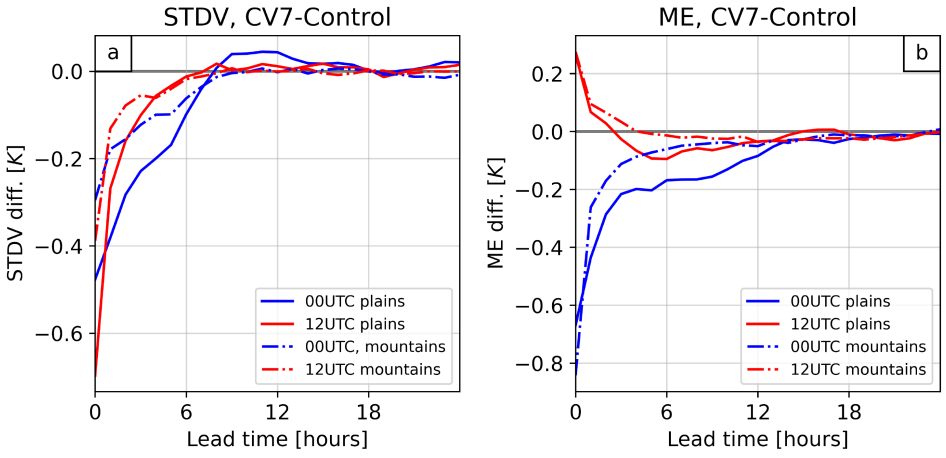

FIGURE 12 | Differences in STDV (a) and ME (b) between the experiments with (CV7) and without (control) the assimilation of surface observations in the plains (solid lines) and in the mountains (dashed-dotted lines).

is higher in the plains than in the mountains, meaning that more observational information is ingested in the analysis. Then, the background error covariance model leads to smooth increments that are horizontally isotropic. This could negatively affect the impact of the assimilation in mountains, which are characterized by horizontally inhomogeneous terrains.

3.6 | Spatial Impact of the Assimilation

The average, station-specific OMF RMS differences in the CV7 experiment with respect to the Control run in the first six forecast hours for METARs and non-assimilated DB stations are presented in Figure 13. These are calculated considering Equation (5) and are used here to evaluate the spatial differences in the impact of the assimilation step on the first forecast hours.

First, it is possible to notice a domain-wide improvement in the RMS following the assimilation step. There are only a few scattered observations for which the RMS difference between control and CV7 experiments is larger than 0, which indicates degradation in the CV7 experiment with respect to the control.

Considering the results for temperature, the improvement at 00 UTC is more marked than at 12 UTC. This can be explained in view of the fact that, as discussed in the previous sections, the analysis step is acting also as a bias correction step at 00 UTC but not at 12 UTC. The correction of the mean component of the OMF differences at 00 UTC results in an increased impact of the assimilation step on the total forecast error when compared to 12 UTC runs.

The results for specific humidity show that the impact of the assimilation step is comparable and covers the same regions at 00 and 12 UTC.

Another interesting aspect emerging from Figure 13 is the general consistency between verification results obtained from METAR observations (stars in Figure 13) and non-assimilated DB observations (dots in Figure 13). There are no evident differences between the improvements observed for the different verification datasets. This is consistent with the similar quality between the observational datasets noted in Section 3.1.

Concerning the spatial impact of the assimilation, it is possible to see that the largest effects for both variables are observed where the density of the assimilated observations is the largest (Figure 1a), namely in the Po Valley. The impact on the nearby mountainous regions and regions where the observational density is less pronounced is less marked. This is consistent with the findings presented in Section 3.5, where it was shown that the impact in the plain regions was more pronounced than in mountainous regions and could partially explain the reported regional differences.

In Sections 3.3–3.6, the forecast scores for temperature and specific humidity have been analyzed in detail. This focus was chosen because these variables exhibit the greatest impact from the assimilation step. To complement these findings, verification scores based on METAR observations for wind components and pressure are presented in Appendix B.

4 | Discussion and Conclusions

A WRF-based prediction system assimilating near real-time surface observations with the 3D-Var algorithm was implemented to improve short-range NWP from LAM forecasts over northern Italy. The assimilated observations were obtained from both conventional (GTS) and nonconventional (Hypermeteo S.r.l. database) sources. The IC and BC for the WRF integrations were

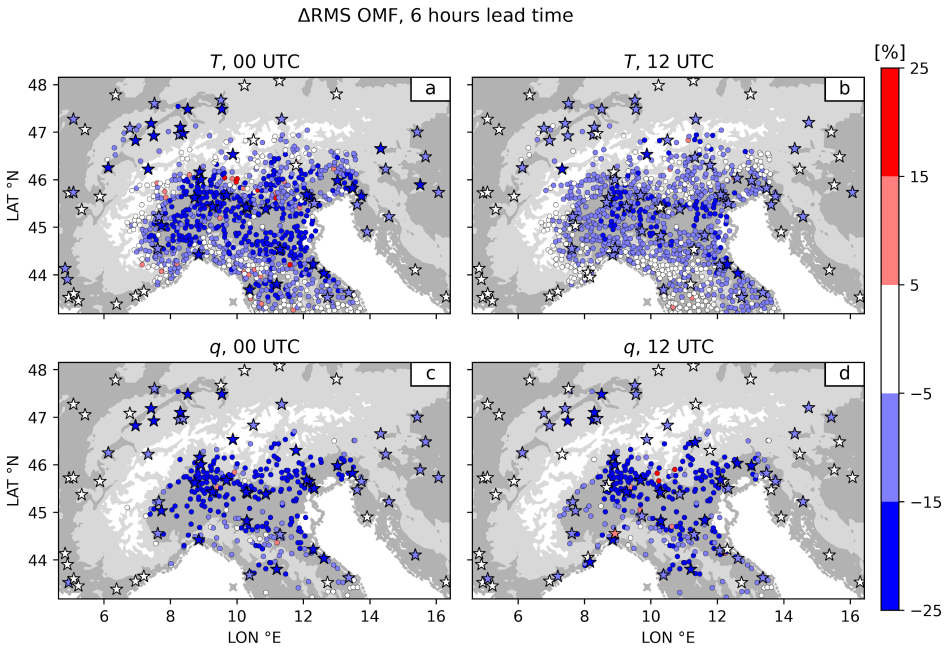

FIGURE 13 | Average of the first six forecast hours of the differences of OMF RMS values between CV7 and control, expressed as a percentage of the control RMS. Negative values indicate improvement. Circles indicate stations from the Hypermeteo S.r.l. database, and stars indicate METAR stations. Panels a and c and b and d present the results from the 00 and 12 UTC runs, for temperature (a and b) and specific humidity (c and d).

obtained from the ICON-EU model. The model chain was optimized to minimize computational cost for operational deployment (i.e., regular production of forecasts) by meteorological agencies and private companies, such as Hypermeteo S.r.l, with limited computing resources.

Two 1-month-long sets of experiments (April 10–May 10, 2024) were performed, one with the assimilation of observations and one without, as a control forecast. The experiments were initialized twice a day, at 00 and 12 UTC. Only surface observations were assimilated, providing useful insight into the impact of such observations on the forecasts.

DA diagnostics (i.e., averages and STDVs of innovations and residuals, averages of the analysis increments) revealed that surface observation innovations and the resulting analysis increments exhibited significant biases. Moreover, the average analysis increments tended to be very smooth, and their spatial variability was not consistent with the orographic features within the domain.

The experiments were then verified against three sets of non-assimilated observations, namely METAR reports, radiosoundings in the Po Valley, and other non-GTS surface observations. The verification against METARs revealed that the forecasts with the assimilation step resulted in overall better performance in predicting surface moisture and temperature with respect to the control ones. The impact on temperature forecasts was larger for 00 UTC than for 12 UTC forecasts.

Verification against radiosoundings showed mixed effects from the assimilation of surface observations onto the lower atmospheric temperature and humidity fields. While the STDV of the differences between forecasts and soundings is either decreased or unaffected by the assimilation step, their ME is slightly larger in absolute value, particularly for humidity below 700 hPa. This slight degradation indicates that the additional assimilation of surface observations can have detrimental effects on vertical temperature and humidity profiles in the lower troposphere. The vertical propagation of biases from surface observation innovations is likely responsible for these systematic discrepancies. This conclusion is also supported by the structure of the average analysis increments presented in Figures 5 and 6.

Finally, the experiments were verified against the dense surface observation dataset hosted by Hypermeteo S.r.l., which was divided into two subsets, one over the mountains and one over the plains. The impact of the additional assimilation step on the evolution of forecast errors for surface variables was assessed separately in the two subsets. This verification showed two main differences with the verification against METARS. First, the diurnal ME cycle is more pronounced for this set than for METARS; this is caused by a more negative daytime bias according to the Hypermeteo S.r.l. observational dataset. Second, verification scores for the 00 UTC runs are subject to some degradation for lead times between 9 and 13 h. This is possibly caused by the cooling of the Po Valley by the assimilation step at 00 UTC. As discussed, the assimilation at 00 UTC is cooling the model state to reduce the model bias; however, during daytime, the model develops a bias that is opposite to the nocturnal one according to verifying Hypermeteo S.r.l. observations, so the nocturnal cooling due to assimilation eventually becomes detrimental in the forecast. This excessive cooling may not have been detected in the verification against METAR observations due to their spatial distribution, which is relatively even across the domain (Figure 1b) and their different diurnal bias with respect to other observation datasets. In contrast, the observations from the Hypermeteo S.r.l. database, used for verification, are primarily located within Italian borders and are concentrated in regions where the density of assimilated stations is highest (Figure 1a,b).

Overall, the effect of the assimilation was more pronounced and lasted longer in the plains than in the mountains, suggesting that the adopted background error model is not able to accurately refine the atmospheric fields in the mountains. At the same time, it is possible that this difference is driven by the higher density of assimilated observations in the plains than in the mountains.

The differences in verification metrics between mountainous and plain regions arise from several factors affecting both surface observations and the model's performance in mountains and plains. One contributing factor could be the systematic mismatch between the actual station height and the station height represented in the model's orography, which affects observations in the mountains. In addition, complex atmospheric and surface processes in mountainous areas are often not accurately captured by the model, particularly at the given resolution.

Our results reveal that the assimilation of surface observations with the proposed setup leads to better surface temperature and humidity forecasts at a limited computational cost; however, such a system is not free of challenges, and further development and testing are envisaged to further characterize and improve its performance.

- *Biases*. We have shown that the innovations given by surface observations are biased, which is suboptimal for DA. Innovation biases could originate either from the observations or from the background. Without the addition of an unbiased estimate of the near-surface atmospheric state (i.e., anchor observations that can be considered unbiased with respect to the true state of the atmosphere), it is impossible to apportion the bias to the two sources of information. One of the main sources of biases in the system is possibly due to the model's state-limited representativity of near-surface processes that determine near-surface temperature and humidity fields, which are instead well captured by surface observations. This systematic mismatch between the resolved state in the two sources of information could indeed result indeed in innovation biases. There are several possible ways to address this issue. First, a different observation operator could be used to calculate the innovations based on the model's 2-m temperature and humidity instead of the temperature and moisture on the first model level; this could reduce the systematic misfit between the observations and their model equivalent. Then, a bias correction scheme could be applied to the innovations to remove the systematic

contribution to the innovations of the under-resolved processes in the model and observational bias. However, bias correction of the innovations is also not free of dangers, as the assimilation of observations adjusted to the model climatology can make the assimilation less effective at reducing model bias. Finally, to address the observed degradation of the forecast scores away from the surface following the assimilation of biased innovations, the vertical localization scale of surface observations could be reduced, limiting their analysis increments to near-surface atmospheric fields. However, the localization must not become too small; otherwise, the surface temperature can become inconsistent with the air mass above. A combination of these actions will be investigated in future works.

- *Smooth increments*. The average analysis increments in our system are rather smooth and do not capture the spatial variability in the atmospheric state induced by the orography. The analysis increments spread horizontally along the terrain-following model levels; this means that increments in valley floors are spread to the top of nearby mountains. This is not physically consistent and should be addressed by choosing a background error model capable of producing analysis increments that respect the physical constraints imposed on the atmospheric state by the orography. For example, an analysis increment given by an observation in a valley should not extend to the top of nearby mountains or to adjacent valleys. Future work will explore the effectiveness of an alternative model for the B matrix, based on the alpha control variable transform (Wang et al. 2014). This specification of B is still climatological, but it provides an inhomogeneous, anisotropic and multivariate analysis increment that respects the constraints imposed by orography. Thus, it could lead to a better overall performance of the assimilation of surface observations in mountainous terrain.
- *Suboptimal weighting of observations and background*. We have used the default values prescribed by WRFDA for the assimilation experiments for R and B, but this choice might not result in optimal assimilation of surface observations. Future work will focus on optimizing the values of both R and B to achieve the best use of the provided observational information for regional forecasts.
- *Seasonality and dependence on model setup*. The results presented here are possibly dependent on the model's physical parameterizations and on the period used to perform the experiments. Not only the innovation biases and spread, but also the forecast verification scores are expected to vary with the season and with the chosen physical parameterizations. This is documented, for example, in García-Díez et al. (2013), where this dependence is evaluated considering different PBL schemes and seasons for surface temperature and humidity. Future work will be devoted to evaluating the skill of the system in different seasons.
- *Observational network*. In this work, we test the assimilation of a set of surface observations, which provide information on the atmospheric state in the PBL. Other near-real-time, open-source, observational networks providing information complementary to that of surface observations are available for assimilation (e.g., radar reflectivities), and their addition could prove beneficial for the developed system, which will be tested in the future. Other modifications to the assimilated datasets of surface observations could also be tested, including, for example, variations of the thinning length and/or the addition of further observations in the mountains and in under-observed regions in the domain.

Author Contributions

Giorgio Doglioni: writing – original draft, writing – review and editing, conceptualization, methodology, data curation, investigation, formal analysis, visualization, software. **Stefano Serafin:** writing – review and editing, conceptualization, methodology, formal analysis, supervision, investigation. **Martin Weissmann:** writing – review and editing, conceptualization, methodology, formal analysis, supervision. **Gianluca Ferrari:** writing – review and editing, funding acquisition, conceptualization, resources. **Dino Zardi:** writing – review and editing, supervision, funding acquisition, conceptualization, resources.

Acknowledgments

This research was supported by the European Union - NextGenerationEU through the Italian National Recovery and Resilience Plan (PNRR), PRIN 2022, Grant code: 2022NEWP4J, CUP E53D23004450006. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the granting authority can be held responsible. Dino Zardi acknowledges the support by the strategic partnerships "Space It Up!", which is funded by the Italian Space Agency and the Ministry of University and Research - Contract No. 2024-5-E.0 - CUP No. I53D24000060005 and "iNEST" (Interconnected Nord-Est Innovation Ecosystem) iinitiative, funded by the European Union under NextGenerationEU (PNRR, Mission 4.2, Investment 1.5, project no. ECS 00000043). We also thank Nicola Carlon (Radarmeteo S.r.l.), Tullio Degiacomi (Hypermeteo S.r.l.), and Lucia Cisco (Hypermeteo S.r.l.) for their valuable technical support in managing the numerical experiments. Open access publishing facilitated by Universita degli Studi di Trento, as part of the Wiley - CRUI-CARE agreement.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Ancell, B. C., C. F. Mass, K. Cook, and B. Colman. 2014. "Comparison of Surface Wind and Temperature Analyses From an Ensemble Kalman Filter and the NWS Real-Time Mesoscale Analysis System." *Weather and Forecasting* 29, no. 4: 1058–1075.

Ancell, B. C., C. F. Mass, and G. J. Hakim. 2011. "Evaluation of Surface Analyses and Forecasts With a Multiscale Ensemble Kalman Filter in Regions of Complex Terrain." *Monthly Weather Review* 139, no. 6: 2008–2024.

Anderson, J., T. Hoar, K. Raeder, et al. 2009. "The Data Assimilation Research Testbed: A Community Facility." *Bulletin of the American Meteorological Society* 90, no. 9: 1283–1296.

Bannister, R. N. 2008. "A Review of Forecast Error Covariance Statistics in Atmospheric Variational Data Assimilation. I: Characteristics and Measurements of Forecast Error Covariances." *Quarterly Journal of the Royal Meteorological Society* 134, no. 637: 1951–1970.

Barker, D., W. Huang, Y.-R. Guo, A. Bourgeois, and Q. Xiao. 2004. "A Three-Dimensional Variational Data Assimilation System for MM5: Implementation and Initial Results." *Monthly Weather Review* 132: 897–914.

Barker, D., X.-Y. Huang, Z. Liu, et al. 2012. "The Weather Research and Forecasting Model's Community Variational/Ensemble Data Assimilation System: WRFDA." *Bulletin of the American Meteorological Society* 93, no. 6: 831–843.

Benjamin, S. G., D. Dévényi, S. S. Weygandt, et al. 2004. "An Hourly Assimilation–Forecast Cycle: The RUC." *Monthly Weather Review* 132, no. 2: 495–518.

Chen, I.-H., J.-S. Hong, Y.-T. Tsai, and C.-T. Fong. 2020. "Improving Afternoon Thunderstorm Prediction Over Taiwan Through 3DVar-Based Radar and Surface Data Assimilation." *Weather and Forecasting* 35, no. 6: 2603–2620.

Demortier, A., M. Mandement, V. Pourret, and O. Caumont. 2024. "Assimilation of Surface Pressure Observations From Personal Weather Stations in AROME-France." *Natural Hazards and Earth System Sciences* 24, no. 3: 907–927.

Demortier, A., M. Mandement, V. Pourret, and O. Caumont. 2025. "Assimilation of Temperature and Relative Humidity Observations From Personal Weather Stations in AROME-France." *Natural Hazards and Earth System Sciences* 25, no. 1: 429–449.

Durre, I., X. Yin, R. S. Vose, S. Applequist, and J. Arnfield. 2018. "Enhancing the Data Coverage in the Integrated Global Radiosonde Archive." *Journal of Atmospheric and Oceanic Technology* 35, no. 9: 1753–1770.

Eyre, J. R. 2016. "Observation Bias Correction Schemes in Data Assimilation Systems: A Theoretical Study of Some of Their Properties." *Quarterly Journal of the Royal Meteorological Society* 142, no. 699: 2284–2291.

García-Díez, M., J. Fernández, L. Fita, and C. Yagüe. 2013. "Seasonal Dependence of WRF Model Biases and Sensitivity to PBL Schemes Over Europe." *Quarterly Journal of the Royal Meteorological Society* 139, no. 671: 501–514.

Giazzi, M., G. Peressutti, L. Cerri, et al. 2022. "Meteonetwork: An Open Crowdsourced Weather Data System." *Atmosphere* 13, no. 6: 928.

Ha, J.-H., and D.-K. Lee. 2012. "Effect of Length Scale Tuning of Background Error in WRF-3DVAR System on Assimilation of High-Resolution Surface Data for Heavy Rainfall Simulation." *Advances in Atmospheric Sciences* 29, no. 6: 1142–1158.

Ha, S.-Y., and C. Snyder. 2014. "Influence of Surface Observations in Mesoscale Data Assimilation Using an Ensemble Kalman Filter." *Monthly Weather Review* 142, no. 4: 1489–1508.

Hacker, J., C. Draper, and L. Madaus. 2018. "Challenges and Opportunities for Data Assimilation in Mountainous Environments." *Atmosphere* 9, no. 4: 127.

Hamill, T. M. 1999. "Hypothesis Tests for Evaluating Numerical Precipitation Forecasts." *Weather and Forecasting* 14, no. 2: 155–167.

Hong, S.-Y., Y. Noh, and J. Dudhia. 2006. "A New Vertical Diffusion Package With an Explicit Treatment of Entrainment Processes." *Monthly Weather Review* 134, no. 9: 2318–2341.

Huang, X.-Y., Q. Xiao, D. M. Barker, et al. 2009. "Four-Dimensional Variational Data Assimilation for WRF: Formulation and Preliminary Results." *Monthly Weather Review* 137, no. 1: 299–314.

Iacono, M. J., J. S. Delamere, E. J. Mlawer, M. W. Shephard, S. A. Clough, and W. D. Collins. 2008. "Radiative Forcing by Long-Lived Greenhouse Gases: Calculations With the AER Radiative Transfer Models." *Journal of Geophysical Research* 113: D13103.

Liu, Z., J. Ban, J.-S. Hong, and Y.-H. Kuo. 2020. "Multi-Resolution Incremental 4D-Var for WRF: Implementation and Application at Convective Scale." *Quarterly Journal of the Royal Meteorological Society* 146, no. 733: 3661–3674.

Lorenc, A. C. 1986. "Analysis Methods for Numerical Weather Prediction." *Quarterly Journal of the Royal Meteorological Society* 112, no. 474: 1177–1194.

Maggioni, E. C., T. Manzoni, A. Perotto, et al. 2023. "WRF Data Assimilation of Weather Stations and Lightning Data for a Convective Event in Northern Italy." *Bulletin of Atmospheric Science and Technology* 4: 8.

NCEP. 2008. *NCEP ADP Global Upper Air and Surface Weather Observations*. National Centers for Environmental Prediction, National Weather Service, NOAA, U.S. Department of Commerce. https://doi.org/10.5065/z83f-n512.

Parrish, D. F., and J. C. Derber. 1992. "The National Meteorological Center's Spectral Statistical-Interpolation Analysis System." *Monthly Weather Review* 120, no. 8: 1747–1763.

Pu, Z., H. Zhang, and J. Anderson. 2013. "Ensemble Kalman Filter Assimilation of Near-Surface Observations Over Complex Terrain: Comparison With 3DVAR for Short-Range Forecasts." *Tellus Series A: Dynamic Meteorology and Oceanography* 65, no. 1: 19620.

Rotach, M. W., S. Serafin, H. C. Ward, et al. 2022. "A Collaborative Effort to Better Understand, Measure, and Model Atmospheric Exchange Processes Over Mountains." *Bulletin of the American Meteorological Society* 103, no. 5: E1282–E1295.

Ruggiero, F. H., K. D. Sashegyi, R. V. Madala, and S. Raman. 1996. "The Use of Surface Observations in Four-Dimensional Data Assimilation Using a Mesoscale Model." *Monthly Weather Review* 124, no. 5: 1018–1033.

Schraff, C., H. Reich, A. Rhodin, et al. 2016. "Kilometre-Scale Ensemble Data Assimilation for the COSMO Model (KENDA)." *Quarterly Journal of the Royal Meteorological Society* 142, no. 696: 1453–1472.

Sgoff, C., W. Acevedo, Z. Paschalidi, et al. 2022. "Assimilation of Crowd-Sourced Surface Observations Over Germany in a Regional Weather Prediction System." *Quarterly Journal of the Royal Meteorological Society* 148, no. 745: 1752–1767.

Skamarock, W. C., and J. B. Klemp. 2008. "A Time-Split Nonhydrostatic Atmospheric Model for Weather Research and Forecasting Applications." *Journal of Computational Physics* 227, no. 7: 3465–3485.

Skamarock, W. C., J. B. Klemp, J. Dudhia, et al. 2019. "A Description of the Advanced Research WRF Model Version 4." Technical Report NCAR/TN-556+STR. UCAR/NCAR.

Sun, J., H. Wang, W. Tong, Y. Zhang, C.-Y. Lin, and D. Xu. 2016. "Comparison of the Impacts of Momentum Control Variables on High-Resolution Variational Data Assimilation and Precipitation Forecasting." *Monthly Weather Review* 144, no. 1: 149–169.

Sun, W., Z. Liu, D. Chen, P. Zhao, and M. Chen. 2020. "Development and Application of the WRFDA-Chem Three-Dimensional Variational (3DVAR) System: Aiming to Improve Air Quality Forecasting and Diagnose Model Deficiencies." *Atmospheric Chemistry and Physics* 20, no. 15: 9311–9329.

Tewari, M., F. Chen, W. Wang, et al. 2004. "Implementation and Verification of the Unified Noah Land Surface Model in the WRF Model." Paper presented at 20th Conference on Weather Analysis and Forecasting/16th Conference on Numerical Weather Prediction, Seattle, WA.

Thompson, G., P. R. Field, R. M. Rasmussen, and W. D. Hall. 2008. "Explicit Forecasts of Winter Precipitation Using an Improved Bulk Microphysics Scheme. Part II: Implementation of a New Snow Parameterization." *Monthly Weather Review* 136, no. 12: 5095–5115.

Wang, H., X.-Y. Huang, J. Sun, et al. 2014. "Inhomogeneous Background Error Modeling for WRF-Var Using the NMC Method." *Journal of Applied Meteorology and Climatology* 53, no. 10: 2287–2309.

Xu, Q. 2019. "On the Choice of Momentum Control Variables and Covariance Modeling for Mesoscale Data Assimilation." *Journal of the Atmospheric Sciences* 76, no. 1: 89–111.

Zängl, G., D. Reinert, P. Rípodas, and M. Baldauf. 2015. "The ICON (ICOsahedral Non-Hydrostatic) Modelling Framework of DWD and MPI-M: Description of the Non-Hydrostatic Dynamical Core." *Quarterly Journal of the Royal Meteorological Society* 141, no. 687: 563–579.

Appendix A

Pseudo-Observation Test

PSOTs consist of the assimilation of an innovation d_k of a given model state variable at point k of the model state. The analysis increment $\delta x = x_b - x_a$ at every other point l of the model state following the assimilation of such innovation d_k can be expressed as (Bannister 2008, eqs. 11 and 12):

$$\delta x_l = \frac{B_{lk}}{\sigma_o^2 + B_{kk}} d_k \tag{A1}$$

where σ_o^2 is the observation error variance, and B_{lk} is the background error covariance between elements l and k of the model state. In Figure A1, the analysis increment following a 1 K temperature PSOT at the first model level is presented. Considering the CV7 specification of the background error, the analysis increment for all other control variables is null. The horizontal temperature analysis increment (Panel a) shows how the information from the increment location (blue dot) is spread horizontally along the first model level. A vertical curtain of the analysis increment is also presented in panel b. The analysis increment spreads isotropically in all directions and follows the terrain conforming to WRF model levels, generating unrealistic analysis increments even in the mountains close to the PSOT location.

The estimates of the background error standard deviation and of the correlation length scale of the background error, reported in Table 1, are based on PSOTs. The background error variance is estimated using Equation (A1) with $l = k$, while the correlation lengthscale of the background error is estimated as the standard deviation of the Gaussian function fitting the analysis increment at the first model level, considering a section passing through the location of the PSOT.

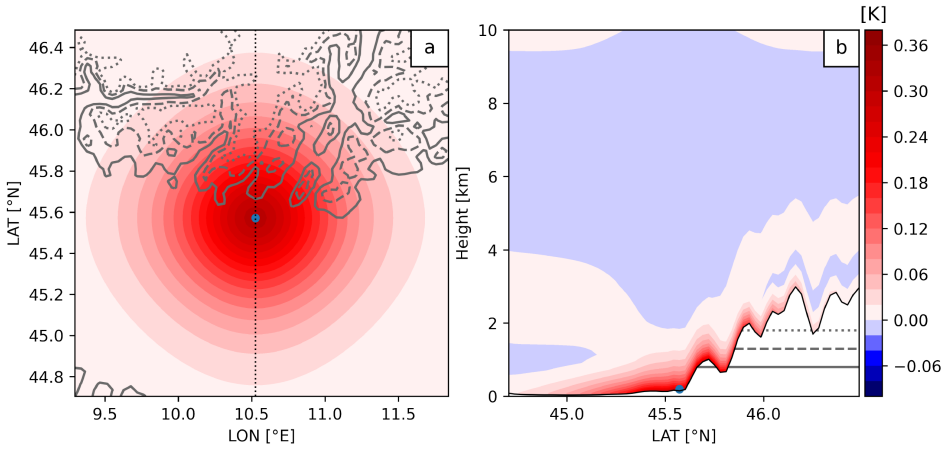

FIGURE A1 | Temperature analysis increment on the first model level (a) and on a vertical curtain (b) corresponding to the dotted line in Panel a following a temperature PSOT located at the first model level in the model point corresponding to the blue dot.

Appendix B

Forecast Scores for Pressure and Wind Components

In Section 3, we focused on the presentation of the verification results for temperature and humidity. We include here the forecast scores, calculated considering METAR observations, for the other assimilated observations: wind components and pressure.

Figure B1 presents the impact of the assimilation of surface observations on the forecast scores for wind components and surface pressure calculated considering METAR observations (Panels a–c).

The RMS of OMF differences is considered, meaning that the random and mean components of the forecast error are considered together. The RMS is calculated as reported in Equation (5).

The forecast scores for the wind components (Figure B1a,b) indicate that the assimilation has an overall positive impact, although this effect lasts only up to 3 h of lead time. On average over the first six forecast hours, the u (v) OMF RMS is reduced by the assimilation by 6.7% (4.5%) for 00 UTC runs and 5.8% (6.1%) for 12 UTC runs with respect to the control.

For pressure, the forecast scores reveal a small but beneficial effect from the assimilation during the first 2–3 h. This is followed by a brief detrimental impact on 00 UTC runs on the RMS of OMF differences, with increases of up to 5 Pa with respect to "control" forecasts. This detrimental signal is not visible for 12 UTC runs. After 5 h of lead time, no notable effects are observed. There are minor indications of improvement after 6 h, particularly for the 12 UTC runs, but these are marginal, on the order of 1 Pa. On average over the first six forecast hours, the improvement in RMS of OMF differences is 5.9% for 00 UTC runs and 5.8% for 12 UTC runs with respect to the "control".

These results show that, despite having a positive impact on wind and pressure forecasts, the assimilation step leads to limited benefit for these variables. Further research is needed to maximize the beneficial effects and possibly reduce the detrimental effects observed for pressure in 00 UTC runs.

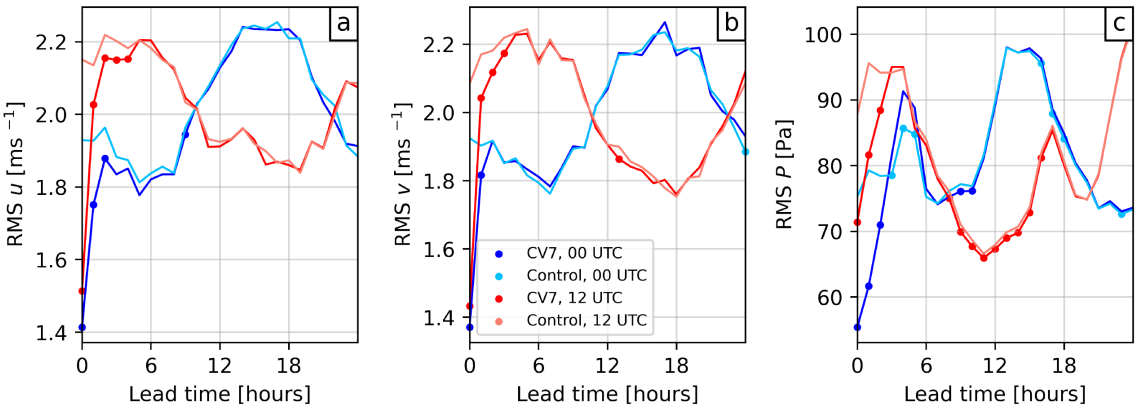

FIGURE B1 | RMS of OMF values for horizontal wind components u (a) and v (b) and pressure (c) considering METAR observations. The dots on the curves indicate that the difference between the experiments differs significantly from zero at the 95% confidence level. The dots are reported only on the curve of the experiment that performs better.