Seamless Multimodel Postprocessing for Air Temperature Forecasts in Complex Topography

REGULA KELLER,^{a,b,c} JAN RAJCZAK,^a JONAS BHEND,^a CHRISTOPH SPIRIG,^a STEPHAN HEMRI,^a
MARK A. LINIGER,^a AND HEINI WERNLI^c

^a Federal Office of Meteorology and Climatology, MeteoSwiss, Zurich, Switzerland

^b Centre for Climate Systems Modelling, ETH Zurich, Zurich, Switzerland

^c Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland

(Manuscript received 11 August 2020, in final form 16 March 2021)

ABSTRACT: Statistical postprocessing is applied in operational forecasting to correct systematic errors of numerical weather prediction models (NWP) and to automatically produce calibrated local forecasts for end-users. Postprocessing is particularly relevant in complex terrain, where even state-of-the-art high-resolution NWP systems cannot resolve many of the small-scale processes shaping local weather conditions. In addition, statistical postprocessing can also be used to combine forecasts from multiple NWP systems. Here we assess an ensemble model output statistics (EMOS) approach to produce seamless temperature forecasts based on a combination of short-term ensemble forecasts from a convection-permitting limited-area ensemble and a medium-range global ensemble forecasting model. We quantify the benefit of this approach compared to only postprocessing the high-resolution NWP. The multimodel EMOS approach (“mixed EMOS”) is able to improve forecasts by 30% with respect to direct model output from the high-resolution NWP. A detailed evaluation of mixed EMOS reveals that it outperforms either one of the single-model EMOS versions by 8%–12%. Temperature forecasts at valley locations profit in particular from the model combination. All forecast variants perform worst in winter (DJF); however, calibration and model combination improves forecast quality substantially. In addition to increasing skill as compared to single-model postprocessing, it also enables us to seamlessly combine multiple forecast sources with different time horizons (and horizontal resolutions) and thereby consolidates short-term to medium-range forecasting time horizons in one product without any user-relevant discontinuity.

KEYWORDS: Postprocessing; Complex terrain; Statistical techniques; Forecast verification/skill; Operational forecasting; Model output statistics

1. Introduction

Weather forecasts are a key element to support decision-making for a broad range of applications. Thus there is a high demand for accurate weather forecasts from a wide range of stakeholders including the general public, the private sector, and authorities issuing weather warnings. Over the last decades, forecasts have been steadily improving largely driven by advances in numerical weather prediction (NWP) including the data assimilation procedure (Bauer et al. 2015; Raftery et al. 2019). The advent of ever more powerful high performance computers allows simulating weather with increasing detail. In addition, multimodel ensemble prediction systems are run to quantify the uncertainty of forecasts. Despite these improvements in NWP, forecasts from physics-based models are not free from systematic bias, and ensemble predictions are often underdispersive (Buizza 1997; Wilks and Vannitsem 2018; Haiden et al. 2019). At the same time, the rapidly

increasing data volume produced by state-of-the-art NWP systems poses a significant challenge to end-users aiming for accurate and easily interpretable products (Morss et al. 2018; Nunley and Sherman-Morris 2020). Furthermore, users usually require one forecast but they may have to choose between different models depending on the time horizon of the forecast they need. And a last challenge is the availability of short-range predictions and the time it takes the users to receive and evaluate them for their specific purpose; therefore often, users cannot profit from short-range high-resolution forecasts.

Statistical postprocessing is an attractive tool to further refine, improve and calibrate NWPs and at the same time generate end-user tailored products. The principle of statistical postprocessing is to describe empirical relationships and (or) error characteristics of past forecast–observation pairs, which then are used to correct the most recent forecasts. The goal of statistical postprocessing is to maximize sharpness subject to calibration (Gneiting et al. 2007).

The pioneering work on so-called model output statistics (MOS) goes back to Glahn and Lowry (1972) and was applied successfully in e.g., the Netherlands for improving various forecast parameters including maximum and minimum temperature (Lemcke and Kruizinga 1988). Since then, MOS and other postprocessing methods have become increasingly popular and many different approaches and variants have been proposed for deterministic (e.g., Glahn and Lowry 1972; Yang et al. 2019) and probabilistic forecasts (e.g., Hamill and Colucci 1997; Wilks 2011; Delle Monache et al. 2013; Vannitsem et al. 2018).

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/WAF-D-20-0141.s1>.

Corresponding author: Regula Keller, regula.keller@meteoswiss.ch

DOI: 10.1175/WAF-D-20-0141.1

© 2021 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

Nonparametric ensemble postprocessing often considers quantiles (Björnarn Bremnes 2004; Taillardat et al. 2016). The most common parametric ensemble postprocessing methods are Bayesian model averaging (BMA; Raftery et al. 2005) and ensemble model output statistics (EMOS; Gneiting et al. 2005). A further EMOS-related method is standardized anomaly model output statistics (SAMOS; Dabernig et al. 2017). Recent research efforts increasingly exploit machine-learning approaches, such as neural networks (NN; Rasp and Lerch 2018). EMOS can be seen as a NN with no hidden layer and a simple linear activation function. Usually, common NN configurations add complexity by incorporating several hidden layers that allows us to incorporate nonlinear relationships between predictors. NNs can be seen as an extension of EMOS, as they add complexity with the hidden layers.

EMOS has been found to be a simple yet skillful approach to postprocess ensemble forecasts enabling the generation of calibrated probabilistic forecasts (Gneiting et al. 2005). The principle of EMOS is a regression that corrects for errors in the mean (e.g., systematic biases) and spread (e.g., under- or overdispersion) of an ensemble forecast. In analogy to multiple regression, EMOS offers flexibility to be extended to a multi-predictor and (or) multimodel framework.

This study explores a high- and a coarse-resolution NWP ensemble and their seamless combination, and investigates the accuracy of probabilistic 2-m air temperature forecasts at measurements stations spread across Switzerland. High-resolution NWP is essential to provide forecasts of local weather up to a few days ahead, especially in regions with complex topography as the Alps. Global, coarser-resolution NWP provide longer-range forecasts. Here, we implement a straight-forward EMOS approach to calibrate the two NWP ensembles and to combine the information from both NWP models into a single data stream. Typically, multimodel ensemble calibration, weighting, and combination are done in separate steps (e.g., Johnson and Swinbank 2009; Beck et al. 2016). Our approach performs these different tasks in one step and includes a seamless transition between multimodel and single-model prediction beyond the forecast horizon of the high-resolution ensemble. Often, simply adding a longer-range forecast at the end of a shorter-range forecast is already called “seamless” (Wastl et al. 2018; Wetterhall and Di Giuseppe 2018). Here, we propose two simple approaches to smooth the transition from a multimodel to single-model prediction that leads to no user-relevant discontinuity and thereby enables to exploit the full range of lead times provided by the NWP model with the longest forecasting horizon. There is a strong need of such consolidated seamless forecasts in operational applications, both for meteorological services and for impact modelers.

The paper is structured as follows: section 2 provides an overview of the data and section 3 describes the methods used in this study, then the results are presented in section 4, and finally section 5 presents a discussion and the conclusions.

2. Data

a. Observational reference

The observational reference in this study are 2-m air temperature measurements from 290 sites in Switzerland (see

Fig. 1a). The majority of these automatic measurement stations are operated by the Swiss Federal Office of Meteorology and Climatology (MeteoSwiss). These observational data are from high-quality instruments and have gone through extensive quality control (automatic and manual). The dataset also includes measurements from several partner networks operated by public authorities, research institutes, and private weather services. The quality of these measurements is lower, as partner data stems from various instrument (mostly high quality) and has been subject to only basic quality control. Nevertheless, our forecasts show comparable scores for data from both origins, hence both are used in this study. The complete station set includes a large variety of locations within the pre-Alpine lowlands as well as in topographically highly complex settings within the Alps, including valley-floor and mountain-top stations. The majority of observations come from below 1000 m (55%), with the lowest at 200 m, and the highest at 3571 m.

b. Numerical weather prediction models

Two state-of-the-art operational NWP ensembles are used in this study, which are based on the high-resolution numerical weather prediction model from the Consortium for Small-Scale Modeling (COSMO-E) operated by MeteoSwiss, and on the coarser-resolution Integrated Forecasting System (IFS-ENS) from the European Centre for Medium-Range Weather Forecasts (ECMWF). IFS-ENS is a 51-member global ensemble at about 18-km horizontal grid spacing (Owens and Hewson 2018). It is initialized four times daily with forecasts out to 6 days (initializations at 0600 and 1800 UTC) and 15 days (initializations at 0000 and 1200 UTC). COSMO-E is a limited-area model with 2.2-km grid spacing for the greater Alpine region offering twice a day (0000 and 1200 UTC) a set of 21 members with forecasts extending to 5 days (Klasa et al. 2018). At the lateral boundaries, COSMO-E forecasts are forced by the 1800 and 0600 UTC IFS-ENS simulations, respectively. Figures 1b and 1c visualize the representation of the topography of Switzerland in both models. This study relies on an archive of the operational 0000 and 1200 UTC runs from both ensemble systems for the time period from 1 January 2017 to 27 October 2019 (i.e., 2 years and 300 days).

Analyses have been carried out for 0000 and 1200 UTC runs separately and show consistent results; thus this paper will merely focus on the results for the 0000 UTC model runs. The focus is on a 120-h forecast horizon, which is the time period covered by both models, and the few following hours for the transition from a multimodel to a single-model system. As the availability of the IFS-ENS output changes from 1- to 3-hourly time steps after a 90-h lead time of, hourly time steps have been obtained by linear interpolation. While a linear interpolation is sufficient during the majority of cases, linear interpolation might not be ideal in cases where short-term rapid changes take place (e.g., frontal passages, convection). A detailed overview of the model attributes is given in Table 1. In the following, COSMO-E is called COSMO and IFS-ENS just IFS.

3. Methods

a. Ensemble model output statistics

Statistical postprocessing aims at correcting systematic biases in NWP output. We here consider the well-established

FIG. 1. (a) Topography of Switzerland at 500-m resolution (horizontal grid spacing). The circles mark the location of the stations and their elevation is indicated by color. The white crosses mark (from left to right) Adelboden, Zürich, and Säntis. (b) Model topography of COSMO-E in and around Switzerland. (c) As in (b), but for IFS-ENS.

EMOS methodology, also termed nonhomogeneous Gaussian regression (Gneiting et al. 2005). EMOS allows us to calibrate probabilistic forecasts by correcting for errors in the mean and variance. An EMOS forecast is characterized by the parameters of a probability density function (PDF). The PDF should best match distributional characteristics of the predictand. In the case of 2-m air temperature, a Gaussian distribution is used (Gneiting et al. 2005; Scheuerer and König 2014). Other variables may require different distributional assumptions (Wilks 2018).

In a very straightforward setup, we use as parameters for our Gaussian predictive PDF the ensemble mean:

$$\mu_{\text{EMOS}}(t) = a + b\overline{x(t)}, \quad (1)$$

and ensemble standard deviation:

$$\sigma_{\text{EMOS}}(t) = \sqrt{c^2 + d^2 s^2(t)}, \quad (2)$$

where $\overline{x(t)}$ and $s(t)$ denote the mean and the standard deviation of the direct model output (DMO) at time t , respectively, and a , b , c , and d are the regression coefficients. Following the approach of Gneiting et al. (2005) the regression coefficients are estimated by minimizing the continuous ranked probability score (CRPS; see below).

The coefficients are estimated, separately for each location and lead time, by using a rolling archive (i.e., training period)

that incorporates the past 45 days. The choice of using 45 days has operational advantages because models change relatively frequently and such a short training period guarantees that training is done only with the current or very recent versions. Also it allows us to partly consider seasonality including seasonally specific weather types, but it can be prone to errors in the case of abrupt changes in weather conditions, in particular during transition seasons and in the case of snow-cover retreat in the Alpine region. Using reforecasts could reduce these errors (Wilks and Hamill 2007; Hagedorn et al. 2008); however, this is usually not operationally feasible. Sensitivity tests (not shown) with different lengths of rolling archive indicated that a window of 45 days is a good choice for 2-m air temperature, in agreement with previous findings (Gneiting et al. 2005; Hagedorn et al. 2008).

b. Multimodel combination

To merge two NWP systems, the EMOS equations are extended such that they allow incorporating an additional predictor for the mean and the standard deviation:

$$\begin{aligned} \mu_{\text{Mixed_EMOS}}(t) &= a + b_1 \overline{x_1(t)} + b_2 \overline{x_2(t)}, \\ \sigma_{\text{Mixed_EMOS}}(t) &= \sqrt{c^2 + d_1^2 s_1^2(t) + d_2^2 s_2^2(t)}, \end{aligned} \quad (3)$$

where $\overline{x_1(t)}$ and $s_1^2(t)$ are the mean and standard deviation of the raw COSMO forecast, and $\overline{x_2(t)}$ and $s_2^2(t)$ are the mean and

TABLE 1. Overview of the model data.

Model	COSMO-E	IFS-ENS
Name	Consortium for Small-Scale Modeling–Ensemble	Integrated Forecasting System–Ensemble
Operated by	MeteoSwiss	ECMWF
Ensemble members	21	51
Domain	Regional	Global
Horizontal grid spacing	2.2 km	18 km
Vertical levels	60	91
Lead time	120 h (5 days)	360 h (15 days)
Temporal resolution	1 h	1/3/6 h
Initialization	2 per day	2 per day

standard deviation of the raw IFS forecast. This combination is termed mixed EMOS in the following. In this study, we combine two models only, but the approach could easily be extended to include more models.

Coefficients b_1 , b_2 , d_1 , and d_2 are constrained to be positive so the weight of a single predictor (for instance, the “importance” of an NWP model ensemble) can be determined as follows:

$$\begin{aligned} \text{weight for mean} &= \frac{b_1}{(b_1 + b_2)}, \\ \text{weight for standard deviation} &= \frac{d_1}{(d_1 + d_2)}, \end{aligned} \quad (4)$$

where weight for mean and weight for standard deviation is the weight of predictor 1, and the weights for predictor 2 are one minus the weight of predictor 1. In the present case, COSMO serves as predictor 1 and is compared to its fractional weight with respect to predictor 2, which is the IFS.

c. Transition

The forecast period of mixed EMOS is limited by the maximum common lead time of both models. In our case, the forecast period of COSMO is 120 h, while IFS extends up to 360 h. Thus, mixed EMOS can only be applied up to 120-h lead time, and thereafter, forecasts can only be based on IFS. To facilitate a smooth transition between mixed EMOS and IFS EMOS we test two approaches and compare them to the version with no transition. While the first approach modifies the predictions during the 3 h before the transition, the second approach affects the three following hours.

For the first approach (referred to as transition 1), upper bounds are defined for coefficients b_1 and d_1 to limit the weight of COSMO. For lead time $t = 118, 119, 120$ h, the upper bounds are defined as

$$b_1(t) = w(t)b_1(117\text{ h}), \quad (5)$$

and analogously for d_1 , where the weights decrease linearly with lead time and attains the values $w = 0.75, 0.5, 0.25$ for $t = 118, 119$, and 120 h, respectively. The second approach (transition 2) prolongs the influence of COSMO to the 3 h beyond its forecast horizon. The difference d :

$$d(\mu, 120\text{ h}) = \mu_{\text{Mixed_EMOS}}(120\text{ h}) - \mu_{\text{IFS_EMOS}}(120\text{ h}), \quad (6)$$

is taken and added to the following 3 h like

$$\mu_{\text{IFS_EMOS}}(t) = \mu_{\text{IFS_EMOS}}(t) + w(t)d(\mu, 120\text{ h}), \quad (7)$$

with the same decreasing weights $w = 0.75, 0.5, 0.25$ for $t = 121, 122, 123$ h. The same is done for σ . Transition 2 adjusts the forecast itself after it predicted by EMOS, while transition 1 influences the parameter estimation before the application of EMOS.

We only present results from the transition of the 0000 UTC runs for consistency but also because, in our

setup, it is the more complex transition. The 0000 UTC run transitions from multimodel to single-model at midnight, i.e., at a time of the day when the model that runs out (COSMO) is typically more important than the continuing IFS—as discussed below.

d. Validation

The grid point of the NWP model that is nearest to the station is used for comparison with the station observations. However, this model grid point might be at a different altitude than the station. It is to be noted that even the 2.2-km NWP system COSMO is not able to fully resolve the complex topography in Switzerland (see also Fig. 1). For instance, the largest mismatch of model topography against target elevation of the corresponding station in the present study is 934 m for the 2.2-km COSMO, and 1687 m for the 18-km IFS, which imposes large systematic biases to the DMO. To enable a fair comparison between the DMO and EMOS-forecasts, DMO is corrected for its altitudinal offset with respect to target locations using a constant lapse rate correction of $0.6^\circ\text{C} (100\text{ m})^{-1}$. Averaged over the entire archive length of 3 years $0.6^\circ\text{C} (100\text{ m})^{-1}$ should be a good estimate of the mean lapse rate, which is also consistent with the standard atmosphere definition. We are aware that spatial, seasonal and diurnal variations can be tremendous and that lapse rates are positive in the case of inversions.

The topographic position index (TPI) is used to characterize the topographic situation of the investigated sites and to assess the impact of topography on the results (Figs. 3 and 5). The TPI corresponds to the difference in elevation of a central pixel relative to the mean altitude of its surrounding eight pixels. A positive TPI indicates an elevated position, e.g., a mountain top or ridge, and a negative TPI denotes a cavity such as a valley. The TPI depends on the grid spacing of the topography dataset; we calculate the TPI with a 500-m resolution topography, as this has been found to characterize local-scale conditions fairly well (not shown).

The performance of different forecasts is assessed using the continuous ranked probability score (CRPS; Gneiting and Raftery 2007). The CRPS is a proper scoring rule and a common measure to evaluate probabilistic forecasts. It takes the integrated squared difference between the cumulative distribution functions (CDFs) of the forecast and the observation. Therefore, smaller values indicate more accurate predictions and a value of 0 denotes a perfect forecast.

To directly compare different strategies, we use the continuous ranked probability skill score (CRPSS):

$$\text{CRPSS} = 1 - \frac{\text{CRPS}}{\text{CRPS}_{\text{ref}}}, \quad (8)$$

where CRPS is the mean score of the forecast under investigation and CRPS_{ref} denotes the score of a reference forecast. A positive CRPSS stands for an improvement of the forecast compared to the reference, with a maximum value of 1 for a perfect forecast. Negative values indicate weaker

FIG. 2. Time series of the mean CRPS for the elevation-corrected direct model outputs (COSMO and IFS), the postprocessed forecasts (COSMO EMOS and IFS EMOS), and the mixed EMOS. Shown are (a) averages over all stations, (b) results for Zürich (lowland), (c) results for Säntis (mountain top), and (d) results for Adelboden (valley floor). The values in parentheses in (a) are averages over the first 120 lead times; the values in (b)–(d) are in the same order.

performance than the reference. To assess the significance of differences in the score between different models (Fig. 3), we use the Diebold–Mariano test (Diebold and Mariano 1995) as implemented in the R package “SpecsVerification” (Siegert 2017).

In addition, we evaluate the calibration of the ensemble forecast with probability integral transform (PIT) histograms. PITs show the relative position of the observation within the ensemble distribution, summed up over many forecasts. Ideally, the histogram has a uniform shape. One-sided PIT

histograms indicate a bias, and a u- or n-shape reveals underdispersion or overdispersion, respectively.

4. Results

a. Verification of forecasts

The mean forecast CRPS as a function of lead time is presented in Fig. 2a for all analyzed forecast variants. Scores for the elevation-corrected raw model output from COSMO and IFS, their postprocessed EMOS counterparts, as well as the

TABLE 2. (Skill) scores averaged over 0–120-h lead times and over spring (MAM), summer (JJA), fall (SON), winter (DJF), and all year (ANN). The CRPS of the elevation-corrected direct model outputs (COSMO and IFS), and postprocessed forecasts (COSMO EMOS and IFS EMOS), and the mixed EMOS. The CRPSS of the same forecasts relative to the elevation-corrected COSMO. The percentage of stations that have a positive CRPSS.

		COSMO	IFS	COSMO EMOS	IFS EMOS	Mixed EMOS
CRPS	MAM	1.64	1.90	1.20	1.19	1.06
	JJA	1.40	1.48	1.05	1.05	0.971
	SON	1.41	1.69	1.10	1.18	1.01
	DJF	1.73	2.34	1.26	1.40	1.17
	ANN	1.54	1.84	1.15	1.19	1.05
CRPSS	MAM	—	−0.210	0.234	0.227	0.320
	JJA	—	−0.113	0.220	0.215	0.279
	SON	—	−0.243	0.201	0.133	0.259
	DJF	—	−0.395	0.252	0.162	0.307
	ANN	—	−0.231	0.241	0.204	0.306
Stations CRPSS > 0 (%)	MAM	—	34.5	100	88.6	100
	JJA	—	50.3	98.6	85.9	99.7
	SON	—	37.9	99.0	73.4	100
	DJF	—	23.4	100	82.4	100
	ANN	—	33.1	100	90.3	100

combined mixed EMOS are shown in a comparative manner. The 2-m air temperature forecast quality of both NWP models varies strongly with time of day. While IFS itself performs best in the morning (recall that lead time 0 corresponds to 0000 UTC, which is local time 0100 in winter and 0200 in summer), COSMO tends to have its best score in the evening. COSMO is worst in the early afternoon, the only time of the day when it is on average outperformed by IFS. The discrepancy stems from a difference in the models' bias of the 2-m air temperature. IFS forecasts too low minimum temperatures and (too) high maxima, while COSMO has the opposite problem, hence its diurnal cycle of temperature is too weak. The application of EMOS lowers the CRPS of COSMO and IFS at all lead times. Both postprocessed forecasts still exhibit the same diurnal cycles but less prominently so. With EMOS, the best performance of IFS shifts to the evening. The discrepancy in forecast quality between the two models has clearly decreased with EMOS postprocessing. EMOS removes most of the bias, which is part of the reason why the CRPS is lower and the diurnal cycle of the score weaker. The diurnal cycle of the EMOS forecasts matches the observations much closer but both EMOS forecasts have a tendency for too high maximum temperatures (especially in summer) and minimum temperatures (particularly in winter). The other influence on the CRPS is the spread of the forecast, which also displays a diurnal cycle (not shown). The standard deviation of both models is too small, particularly for the first few days and for COSMO, therefore EMOS increases the standard deviation. The IFS EMOS has a larger spread in the night than COSMO while COSMO has a slightly larger standard deviation in the early afternoon, which could explain part of the remaining diurnal cycle in CRPS. The mixed EMOS outperforms both single-model EMOS at all lead times. The diurnal cycle of the mixed EMOS score is highly correlated with the score of COSMO EMOS but at lower CRPS values. As expected, the CRPS tends to increase with lead time, but this increase is comparable to the diurnal variations.

The seasonal CRPS and CRPSS of the different forecast approaches are summarized in Table 2. The skill score is calculated with the elevation-corrected COSMO as reference. The high-resolution NWP model COSMO outperforms the coarser-resolution IFS in all seasons, particularly during winter (DJF) when forecast quality is lowest for both models. EMOS is able to improve the forecast of COSMO in a distinct manner by 24% and of IFS by 20% (last row within the CRPSS section of Table 2). In the case of COSMO, EMOS improves the skill at all stations in spring and winter (see last section of Table 2), and at more than 98% of the stations in summer and fall. Forecast quality of IFS EMOS improves for more than 90% of cases at the annual scale. Seasonally stratified, values range from 73% in fall to 89% in spring. Also with EMOS, both models still exhibit weakest performance in terms of CRPS during winter (see first section of Table 2). COSMO EMOS and IFS EMOS have a similar score in spring and summer. In fall and winter, COSMO EMOS clearly outperforms IFS EMOS.

The mixed EMOS improves 2-m air temperature forecasts by ~30% with respect to elevation-corrected COSMO (middle section of Table 2), by 8.8% with respect to COSMO EMOS and by 12% with respect to IFS EMOS. The mixed EMOS outperforms all other forecasts in all seasons. The general performance is best in summer (CRPS of 0.971), followed by fall (1.01), spring (1.06), and winter (1.17) (first section of Table 2). All five forecasting variants under investigation share the same seasonal rank order in terms of forecast quality. The improvement in terms of CRPSS of mixed EMOS over the operationally used weather forecasting model COSMO is largest in spring and winter. The benefit of mixed EMOS is present in all stations in all seasons except summer, when one single station has a CRPSS of ~0.

Location-specific analyses for three geographically diverse stations are presented in Figs. 2b–d in terms of annual mean CRPS as a function of lead time. The three stations are used to illustrate three characteristic examples of locations in Switzerland

FIG. 3. Analysis of how the CRPSS of the mixed EMOS depends on (a),(b) latitude vs elevation of the station and (c),(d) TPI vs elevation. Blue–gray shading indicates the difference of the CRPSS of the mixed EMOS compared to elevation-corrected COSMO (the direct model output), averaged over 0–120-h lead times, for day and night. Red circles denote where the difference between the CRPS of mixed EMOS and COSMO EMOS (the postprocessed version) is significantly worse (with a significance level of 0.05). The difference in CRPS of mixed EMOS performs better at all other sites [only 2 sites in (a) and (c) and 13 sites in (b) and (d) with nonsignificant improvement] than COSMO EMOS. All data are separated into (left) day time and (right) nighttime.

and enable to study model behavior in detail. Zürich is a typical low-land pre-Alpine location at 556 m MSL, Säntis is a mountain-top station at 2501 m MSL that distinctly stands out into the free atmosphere, and Adelboden is an Alpine valley floor location at 1321 m MSL prone to local-scale processes like cold-pool formation, shading effects and local-scale wind-systems. It should be noted that valley stations are an inhomogeneous group and Adelboden is not representative for all of them.

It is obvious in Fig. 2b that for Zürich both models already have good scores; only IFS stands out negatively at night. EMOS can improve this shortcoming and slightly improve both forecasts at all lead times. The different EMOS variants exhibit very similar scores. At all times of day, the mixed EMOS narrowly outperforms either single-model EMOS for Zürich. For Säntis (Fig. 2c), IFS has distinct biases (CRPS of 3.17). This is primarily because of the horizontal and vertical

mismatch between the target location and the coarse grid spacing of IFS topography and is to some extent also evident for COSMO. IFS is also characterized by a large diurnal cycle in score with weakest performance in the early morning and best performance during the afternoon. EMOS is able to massively improve 2-m air temperature predictions at Säntis, particularly for IFS. The mixed EMOS is mostly consistent with COSMO EMOS, but performs slightly better on average. In Adelboden (Fig. 2d) both models show mediocre scores as compared to the entire set analyzed. Postprocessing with EMOS (and model combination) is able to remove large parts of these deficits but is not able to reach the low CRPS levels seen at low-land and mountain-top locations. Still, Adelboden is the location where the benefit of the mixed EMOS (with respect to COSMO) is largest due to important shortcomings in both models.

FIG. 4. PIT diagram of elevation-corrected IFS and COSMO, postprocessed IFS and COSMO (“EMOS”), and mixed EMOS at an 18-h lead time.

An in-depth analysis of the benefit of mixed EMOS with respect to COSMO and its postprocessed counterpart during day (0700–1800 UTC) and night (1900–0600 UTC) is shown in Fig. 3. During the day, the gain in forecast quality of mixed EMOS relative to COSMO is most prominent in elevated valleys (altitude of 1000–2000 m and negative values of TPI; Figs. 3a and 3c). There, even the high-resolution COSMO is insufficient to capture all relevant local-scale processes. The strong cooling of IFS at night tends to be more realistic in clear nights and therefore improves mixed EMOS forecasts. At night, the skill of mixed EMOS cannot be related to the geographic characteristics as easily (Figs. 3b,d). Largest improvements of mixed EMOS occur in the Alps albeit with a lot of variability between stations. Skill at stations in the Swiss plateau (characterized by altitude around 500 m and north of 47°N with small TPI magnitude) is generally limited to 10%–20%. It is, however, interesting to note that mixed EMOS performs significantly better than COSMO EMOS for all sites but two during the day (Figs. 3a,c). While the multimodel combination improves forecast quality at most stations during the night, at exposed locations (i.e., elevated stations with positive TPI) along the northern slopes of the Alps, the mixed EMOS forecasts perform significantly worse than the single-model COSMO EMOS but still improving over COSMO (Figs. 3b,d). These exposed locations reach into the free atmosphere that is decoupled from the boundary layer at night, hence experiencing a smaller diurnal cycle in temperature. In IFS, these places are located at too low altitude and the model cools too much at night. COSMO is better able to distinguish ridges and peaks and can better represent local-scale phenomena. Hence, the addition of IFS in exposed locations can be disadvantageous.

The dispersion of the different ensembles is analyzed using PIT histograms. Figure 4 shows a PIT histogram for 18-h lead time (i.e., in the early evening). Both, COSMO and IFS are strongly underdispersed and possess a systematic cold bias, in particular the IFS. PIT diagrams at other lead times show similar results; only around noon does COSMO show no or a small positive bias and then the negative bias in IFS is the smallest. Applying EMOS or mixed EMOS reduces underdispersion distinctly, which

indicates an improved calibration of EMOS-postprocessed temperature predictions.

The remaining error of the mixed EMOS depends on the season (Table 2). Winter (DJF) has the highest CRPS despite the second largest improvement in skill over COSMO. The pattern of the CRPS is related to TPI in winter (and fall) as shown in Fig. 5. Locations with flat surroundings have the best score while stations within topography have worse CRPS values. In the lowlands, where the forecast is already good to begin with, postprocessing and model combination improves it even more. In the mountainous areas, the benefit of EMOS is largest but the error remains higher than in lowland areas. In a few cases, the CRPS value is still larger than 2. These stations

FIG. 5. CRPS, averaged over 0–120-h lead times of the mixed EMOS in winter (DJF) against TPI at 500-m resolution (horizontal grid spacing). The blue marked stations are (from top to bottom) Sântis (mountain top), Zürich (lowland), and Adelboden (valley floor).

FIG. 6. The weight of COSMO direct model output in the mixed EMOS for the mean (red lines) and standard deviation (black lines). Averages over all dates and (a) all stations, (b) Zürich (lowland), (c) Säntis (mountain top), and (d) Adelboden (valley floor). The values in brackets are averages over the first 120 lead times. A value > 0.5 means that COSMO has higher weight than IFS in the mixed EMOS.

are located in midaltitude valleys (approx. 1000–2000 m), which are prone to the formation of cold air pools. (The one location in Fig. 5 with a positive TPI and high CRPS is situated in a small basin above a large valley, making the TPI at 500 m unsuitable to describe this location.) On clear winter nights, temperature at these locations can drop dramatically, which is hard to predict by NWP and only partly corrected by EMOS.

b. Model weights in mixed EMOS

Figure 6a depicts the 290 station average weight of COSMO, Weight, as a function of lead time. The weight of

IFS consequently is $1 - \text{Weight}$. The weights for the mean and for the standard deviation show a similar pattern. While IFS is more important (i.e., $\text{Weight} < 0.5$) during daytime and toward longer lead times, COSMO is the favored model during the night and at shorter lead times, especially during the first few hours. The times when IFS has more weight correspond to the cases when IFS EMOS outperforms COSMO EMOS (see Fig. 2a). The weight of COSMO for the standard deviation is generally lower (i.e., IFS is more important), especially on day one. Both raw models are underdispersed, but IFS less so, therefore has a higher weight for the spread.

FIG. 7. (a) The mean absolute difference to the previous lead time of the forecast mean, (b) the mean absolute difference to the previous lead time of the forecast standard deviation, and (c) the mean CRPS during the transition period from forecast hours 116 to 126 for IFS EMOS, COSMO EMOS, mixed EMOS without any transition, and mixed EMOS with transitions 1 and 2 (see text for explanation).

Figures 6b–d illustrate the model weighting (i.e., importance of COSMO at given locations) for the three exemplary locations discussed earlier (see Figs. 2b–d). For the two stations located within complex topography (Santis and Adelboden) COSMO is the model in favor compared to IFS. At the station of Zürich in the lowland the average weight of COSMO is 0.49 for the mean and 0.43 for the standard deviation, indicating an equal importance of the two models overall. However, the weight has a distinctive diurnal cycle with high values (Weight > 0.6) during the night when COSMO has a better score, and low values at daytime (Weight < 0.4) when IFS is the preferred model. The mountain station Santis has high weights for COSMO (0.84 and 0.66 for the mean and the standard deviation, respectively), with a strong diurnal cycle (weight for mean near 1 at night and 0.5–0.7 during the afternoon). In the case of the valley-floor location Adelboden, the weight of COSMO is at roughly 0.67 for the mean and 0.53 for the standard deviation.

c. Transition

Figure 7 shows the forecast score and “smoothness” in the transition phase from a multimodel to a single-model forecast. The “smoothness” is displayed as the mean absolute difference in the forecast between the forecast at the time of interest and 1 h earlier (i.e., the previous lead time). Both for predicted mean temperature (Fig. 7a) and its standard deviation (Fig. 7b)

the transition between the forecasting strategies is substantial, with a spurious peak in short-term changes, in the case where no specific transition smoothing is applied. The two simple transition approaches applied here (i.e., transition 1 and transition 2) provide “smooth” forecasts but still possess a signal with respect to the continuous (single-model) IFS EMOS. Note that due to the diurnal cycle of the temperature forecast, there is always a difference to the forecast at the previous lead time.

The CRPS score of the different transition methods is presented in Fig. 7c. Obviously transition 2 is the preferable approach as the overall performance is best. Transition 1 decreases the forecast quality in the last hours before COSMO fades out, while transition 2 increases the quality beyond the transition at 120 h and even improves predictions in the following hours.

5. Conclusions

The present study demonstrates that statistical postprocessing and model combination is able to distinctly improve 2-m air temperature forecasts of state-of-the-art operational NWP systems. Using the EMOS method, we show that postprocessing removes substantial parts of the systematic biases inherent to weather prediction models and further adjusts for errors in ensemble spread and thereby improves calibration. We particularly assessed differences in performance of the

global coarser-resolution IFS and the regional higher-resolution COSMO ensembles and their combination.

For the example of COSMO and IFS, it is shown that combining a higher- and coarser-resolution model enables a further improvement in forecast quality by 8%–12% (measured in terms of CRPSS) compared with the calibration of a single NWP. The average improvement amounts up to 30% as compared to the high-resolution NWP model in operation (COSMO). The benefit of the mixed EMOS over IFS EMOS is especially large at high altitudes and peak locations, where EMOS alone is insufficient to correct the poor performance of IFS (see also Fig. 2c). The improvement of mixed EMOS over COSMO EMOS is generally smallest on mountain tops and at night and largest in valleys (see Fig. 3). In short, mixed EMOS has a smaller bias and a smaller spread than either single-model EMOS in all seasons, leading to an improved score.

The multimodel combination with mixed EMOS allows us to combine forecasts with a different forecast time horizon. To provide a seamless prediction across the transition from two to one available ensembles, we proposed two simple approaches: One approach decreases the weight of the shorter NWP forecast in its last few hours, while the other approach is based on an extrapolation of the shorter NWP forecasts beyond its time horizon. While both approaches are fairly similar in terms of smoothness, the second approach is favorable as its overall skill is higher. Both approaches are motivated by the fact that we try to provide a smooth forecast without breaks or inconsistencies. The two approaches applied are straightforward and illustrative, and more sophisticated approaches can be developed in future studies. There is also the potential to smooth additional seams in the proposed framework. Most notably, one with respect to observations potentially leading to improvements in forecast skill in the nowcasting time horizon (not shown), as well as with respect to monthly, seasonal, and decadal forecasts and climate projections.

The model combination in this study relies on ideal conditions, in which both models are available for the same initialization, i.e., the 0000 UTC run from COSMO is combined with the 0000 UTC run from IFS. In an operational setup, this is usually not always the case and times of availability might differ substantially. Commonly, IFS forecasts for identical initialization times are available much later than the corresponding COSMO runs. Usually, the most recent IFS is 6–12 h older than COSMO. Testing the presented setup with lagged IFS runs (i.e., using the 12 h older IFS run) revealed identical results with only very small deterioration in skill.

The proposed multimodel combination presented in this study is attractive for several reasons. It substantially improves forecast skill compared to any single-model postprocessing, particularly in valleys and even for single-model postprocessings using a high-resolution NWP model. It allows us to combine high-resolution NWP products, which usually exhibit frequent update cycles, with coarser-resolution NWP products, which are updated less frequently but provide predictions for longer lead times into a seamless forecast. A multimodel combined prediction thus profits from frequent updates of new forecast information that is of relevance in situations with low predictability (e.g., frontal passages and summertime convection),

as well as from covering medium-range lead times (i.e., in the present example 15 days). The mixed EMOS thereby generates medium-range forecasts in a seamless manner with improved skill during the period of overlap (i.e., up to five days in the present case), thereby consolidating all information in to a single product without any discontinuity. The proposed multimodel combination additionally offers more reliable operations as the risk of outage in operation is spread across multiple model sources. Finally, the mixed EMOS approach acts as a foundation and can be extended flexibly to even more than two models, additional physical predictors or more sophisticated predictor interactions. Our analysis solely analyzed 2-m air temperature predictions that rely on a Gaussian setup of EMOS using a moving-window strategy. Transferring the applied approach to other variables is an interesting topic for future research but requires some adaptation in terms of the underlying distribution of data and the training strategy.

Acknowledgments. We acknowledge the following institutions for providing observations from their monitoring networks: Federal Office for the Environment (FOEN), numerous Cantons of Switzerland, MeteoGroup Schweiz AG, and the Swiss Federal Institute for Forest, Snow and Landscape Research (WSL). We thank Pirmin Kaufmann and Lionel Moret from MeteoSwiss for the discussion on the scores of different models and their valuable input for the manuscript.

Data availability statement. The observation data used in this work are available after registration on the data portal of MeteoSwiss (<https://gate.meteoswiss.ch/idaweb>). Model data are available from the authors upon request.

REFERENCES

- Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, <https://doi.org/10.1038/nature14956>.
- Beck, J., F. Bouttier, L. Wiegand, C. Gebhardt, C. Eagle, and N. Roberts, 2016: Development and verification of two convection-allowing multi-model ensembles over Western Europe. *Quart. J. Roy. Meteor. Soc.*, **142**, 2808–2826, <https://doi.org/10.1002/qj.2870>.
- Bjørnar Bremnes, J., 2004: Probabilistic forecasts of precipitation in terms of quantiles using NWP model output. *Mon. Wea. Rev.*, **132**, 338–347, [https://doi.org/10.1175/1520-0493\(2004\)132<0338:PFOPIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(2004)132<0338:PFOPIT>2.0.CO;2).
- Buizza, R., 1997: Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system. *Mon. Wea. Rev.*, **125**, 99–119, [https://doi.org/10.1175/1520-0493\(1997\)125<0099:PFSOEP>2.0.CO;2](https://doi.org/10.1175/1520-0493(1997)125<0099:PFSOEP>2.0.CO;2).
- Dabernig, M., G. Mayr, J. Messner, and A. Zeileis, 2017: Spatial ensemble post-processing with standardized anomalies. *Quart. J. Roy. Meteor. Soc.*, **143**, 909–916, <https://doi.org/10.1002/qj.2975>.
- Delle Monache, L., F. A. Eckel, D. L. Rife, B. Nagarajan, and K. Searight, 2013: Probabilistic weather prediction with an analog ensemble. *Mon. Wea. Rev.*, **141**, 3498–3516, <https://doi.org/10.1175/MWR-D-12-00281.1>.
- Diebold, F. X., and R. S. Mariano, 1995: Comparing predictive accuracy. *J. Bus. Econ. Stat.*, **13**, 253–263.
- Glahn, H. R., and D. A. Lowry, 1972: The use of Model Output Statistics (MOS) in objective weather forecasting. *J. Appl.*

- Meteor.*, **11**, 1203–1211, [https://doi.org/10.1175/1520-0450\(1972\)011<1203:TUOMOS>2.0.CO;2](https://doi.org/10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2).
- Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.*, **102**, 359–378, <https://doi.org/10.1198/016214506000001437>.
- , —, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, <https://doi.org/10.1175/MWR2904.1>.
- , F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc.*, **69**, 243–268, <https://doi.org/10.1111/j.1467-9868.2007.00587.x>.
- Hagedorn, R., T. M. Hamill, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-meter temperatures. *Mon. Wea. Rev.*, **136**, 2608–2619, <https://doi.org/10.1175/2007MWR2410.1>.
- Haiden, T., M. Janousek, F. Vitart, L. Ferranti, and F. Prates, 2019: Evaluation of ECMWF forecasts, including the 2019 upgrade. ECMWF Tech. Memo. 588, 56 pp., <https://doi.org/10.21957/mlvapkke>.
- Hamill, T. M., and S. J. Colucci, 1997: Verification of eta–RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, **125**, 1312–1327, [https://doi.org/10.1175/1520-0493\(1997\)125<1312:VOERSR>2.0.CO;2](https://doi.org/10.1175/1520-0493(1997)125<1312:VOERSR>2.0.CO;2).
- Johnson, C., and R. Swinbank, 2009: Medium-range multimodel ensemble combination and calibration. *Quart. J. Roy. Meteor. Soc.*, **135**, 777–794, <https://doi.org/10.1002/qj.383>.
- Klasa, C., M. Arpagaus, A. Walser, and H. Wernli, 2018: An evaluation of the convection-permitting ensemble COSMO-E for three contrasting precipitation events in Switzerland. *Quart. J. Roy. Meteor. Soc.*, **144**, 744–764, <https://doi.org/10.1002/qj.3245>.
- Lemcke, C., and S. Kruizinga, 1988: Model output statistics forecasts: Three years of operational experience in the Netherlands. *Mon. Wea. Rev.*, **116**, 1077–1090, [https://doi.org/10.1175/1520-0493\(1988\)116<1077:MOSFTY>2.0.CO;2](https://doi.org/10.1175/1520-0493(1988)116<1077:MOSFTY>2.0.CO;2).
- Morss, R. E., and Coauthors, 2018: Hazardous weather prediction and communication in the modern information environment. *Bull. Amer. Meteor. Soc.*, **98**, 2653–2674, <https://doi.org/10.1175/BAMS-D-16-0058.1>.
- Nunley, C., and K. Sherman-Morris, 2020: What people know about the weather. *Bull. Amer. Meteor. Soc.*, **101**, E1245–E1240, <https://doi.org/10.1175/BAMS-D-19-0081.1>.
- Owens, R. G., and T. Hewson, 2018: ECMWF forecast user guide. ECMWF, <https://doi.org/10.21957/m1cs7h>.
- Raftery, A., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, <https://doi.org/10.1175/MWR2906.1>.
- , —, —, and —, 2019: The ECMWF ensemble prediction system: Looking back (more than) 25 years and projecting forward 25 years. *Quart. J. Roy. Meteor. Soc.*, **145**, 12–24, <https://doi.org/10.1002/qj.3383>.
- Rasp, S., and S. Lerch, 2018: Neural networks for postprocessing ensemble weather forecasts. *Mon. Wea. Rev.*, **146**, 3885–3900, <https://doi.org/10.1175/MWR-D-18-0187.1>.
- Scheuerer, M., and G. König, 2014: Gridded, locally calibrated, probabilistic temperature forecasts based on ensemble model output statistics. *Quart. J. Roy. Meteor. Soc.*, **140**, 2582–2590, <https://doi.org/10.1002/qj.2323>.
- Siebert, S., 2017: SpecsVerification: Forecast Verification Routines for Ensemble Forecasts of Weather and Climate. Accessed 19 June 2020, <https://CRAN.R-project.org/package=SpecsVerification>.
- Taillardat, M., O. Mestre, M. Zamo, and P. Naveau, 2016: Calibrated ensemble forecasts using quantile regression forecasts and ensemble model output statistics. *Mon. Wea. Rev.*, **144**, 2375–2393, <https://doi.org/10.1175/MWR-D-15-0260.1>.
- Vannitsem, S., D. S. Wilks, and J. W. Messner, Eds., 2018: *Statistical Postprocessing of Ensemble Forecasts*. Elsevier, 362 pp.
- Wastl, C., and Coauthors, 2018: A seamless probabilistic forecasting system for decision making in civil protection. *Meteor. Z.*, **27**, 417–430, <https://doi.org/10.1127/metz/2018/902>.
- Wetterhall, F., and F. Di Giuseppe, 2018: The benefit of seamless forecasts for hydrological predictions over Europe. *Hydrol. Earth Syst. Sci.*, **22**, 3409–3420, <https://doi.org/10.5194/hess-22-3409-2018>.
- Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.
- , 2018: Univariate ensemble postprocessing. *Statistical Postprocessing of Ensemble Forecasts*, S. Vannitsem, D. S. Wilks, and J. W. Messner, Eds., Elsevier, 49–89, <https://doi.org/10.1016/B978-0-12-812372-0.00003-0>.
- , and T. M. Hamill, 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Wea. Rev.*, **135**, 2379–2390, <https://doi.org/10.1175/MWR3402.1>.
- , and S. Vannitsem, 2018: Uncertain forecasts from deterministic dynamics. *Statistical Postprocessing of Ensemble Forecasts*, S. Vannitsem, D. S. Wilks, and J. W. Messner, Eds., Elsevier, 1–13, <https://doi.org/10.1016/B978-0-12-812372-0.00001-7>.
- Yang, J., M. Astitha, and C. S. Schwartz, 2019: Assessment of storm wind speed prediction using gridded Bayesian regression applied to historical events with NCAR’s real-time ensemble forecast system. *J. Geophys. Res. Atmos.*, **124**, 9241–9261, <https://doi.org/10.1029/2018JD029590>.