Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Home Affairs FDHA
Federal Office of Meteorology and Climatology MeteoSwiss

MeteoSwiss

Technical Report MeteoSwiss No. 288

Applying COMFORT to automatic weather forecasts: impacts and lessons learned

Adrien Chabert, Jonas Bhend, Mathieu Schaer, Christoph Spirig

ISSN: 2296-0058

Technical Report MeteoSwiss No. 000

Applying COMFORT to automatic weather forecasts: impacts and lessons learned

Adrien Chabert, Jonas Bhend, Mathieu Schaer, Christoph Spirig

Recommended citation:

Chabert, A. et al. (2025): Applying COMFORT to automatic weather forecasts: impacts and lessons learned *Technical Report MeteoSwiss,* **288,** 35 pp, doi: 10.18751/pmch/tr/288.comfort/1.0

Editor:

Federal Office of Meteorology and Climatology, MeteoSwiss, © 2025

MeteoSwiss
Operation Center 1
CH-8044 Zürich-Flughafen
T +41 58 460 99 99
www.meteoschweiz.ch

Abstract

MeteoSwiss monitors and documents the quality of its public weather forecasts with the COMFORT score. In May 2021 the forecasting system producing the public weather forecasts was fully automated. A significant drop in forecast quality as measured by the COMFORT headline score was observed in the aftermath of this change. In this report, we document the changes to the forecasting system, the ensuing score degradation and the measures taken to mitigate the adverse impact. With all the modifications in place, the scores are now better than before the change in 2021. We conclude by collecting the lessons learned and by sketching how the score might evolve in the coming years.

MeteoSwiss
Operation Center 1
CH-8044 Zürich-Flughafen
T +41 58 460 99 99
www.meteoschweiz.ch

Contents

Abstract		**V**
1	**Introduction**	**1**
2	**Change to the forecasting system**	**3**
2.1	Data4Web as the new data source	4
3	**Decline of COMFORT during 2021**	**7**
3.1	Precipitation	8
3.2	Relative sunshine duration	8
3.3	Wind	8
3.4	Temperature	9
4	**Modifications to Data4Web**	**11**
4.1	Order of operations	11
4.2	Post-processing	11
4.3	Introduction of reforecast capability	12
5	**Recovery of COMFORT in 2021 and 2022**	**13**
5.1	Influence of the weather conditions	14
6	**Towards a new headline score**	**19**
6.1	Score saturation	19
6.2	Proper scoring rules for probabilistic forecasts	21
7	**Summary**	**25**
8	**Lessons learned**	**17**
8.1	Planning	17
8.2	Improved monitoring	17
8.3	Headline score	18
References		**27**
Acknowledgements		**29**
A	**Release history of Data4Web**	**31**
B	**COMFORT score definitions**	**33**

MeteoSwiss
Operation Center 1
CH-8044 Zürich-Flughafen
T +41 58 460 99 99
www.meteoschweiz.ch

1 Introduction

MeteoSwiss has the duty to document the quality of its services. Several key performance indicators are used to evaluate different services provided to users such as the general public, the aviation or the Swiss authorities responsible to take action in case of severe meteorological events.

To measure the quality of the local weather forecasts as distributed through the mobile application and the website, the verification scheme COMFORT (COntinuous MeteoSwiss FORecast qualiTy) was created in 2013 (Cattani, et al., 2015). This verification framework was designed to provide management, but also external entities such as policy makers and media with a summary measure of the quality of public weather forecasts provided by MeteoSwiss. As such, COMFORT is one of MeteoSwiss' headline scores.

In COMFORT, the forecast error of daily forecast parameters is evaluated for 27 regions, 7 forecast lead days and 5 different parameters, namely precipitation, daily minimum and maximum temperature, sunshine duration and wind speed (Cattani, et al., 2015). These partial scores for individual forecast parameters are combined to produce the global COMFORT score used as the headline score. Following MeteoSwiss' commitment to continuously improve its forecasts, performance targets of increasing COMFORT score values are agreed upon together with the governing ministry Federal Department of Home Affairs.

The COMFORT score was originally calculated from regional forecasts produced by a semi-automated system supervised by MeteoSwiss' forecasters. As the computing power and performance of numerical models improved, it was decided in 2020 to move away from the semi-automated to a fully automated system. The introduction of the fully automated system in mid-2021 lead to a significant drop in the COMFORT scores. This report analyses the causes that lead to the degradation of COMFORT scores and documents the measures taken to mitigate the adverse effects on the headline score and forecast quality. Since October 2021, quality targets are met again and the modifications to the forecasting system resulted in better scores than before the introduction of fully automated forecasts.

In the following section, we outline the changes to the forecasting system and data sources used to compute the COMFORT score. In Section 3, we present the problems encountered after the automation of the forecasting system and highlight likely causes followed by a documentation of modifications to mitigate the degradation of the headline score in Section 4. In Section 5, the effects of these modifications on the COMFORT score are discussed. In Section 6, we provide an outlook on ideas for a redesign of the MeteoSwiss headline score for public weather forecasts. This short report concludes with lessons learned and a summary in Sections 7 and 8.

MeteoSwiss
Operation Center 1
CH-8044 Zürich-Flughafen
T +41 58 460 99 99
www.meteoschweiz.ch

2 Change to the forecasting system

When COMFORT was introduced in 2013, the public weather forecasts of MeteoSwiss were produced by a semi-automated system. This consisted of a part called the Matrix, combining a 'first guess' prediction from numerical weather prediction models with manual modifications by forecasters on a regional scale with daily temporal resolution. Subsequently, these regional, daily forecast quantities were disaggregated to hourly forecasts at many individual points of interest (POIs) for use in the mobile app and on the web by a system called Data4Web.

COMFORT, the MeteoSwiss headline score for local forecasts, was evaluated against output from the Matrix. Thereby COMFORT was based on regional forecasts with daily resolution and not the hourly point forecasts communicated through the mobile app and web. Consequently, COMFORT was computed for 27 or fewer regions only, for some parameters only at single reference sites within each of the forecast regions assessed (Table 1).

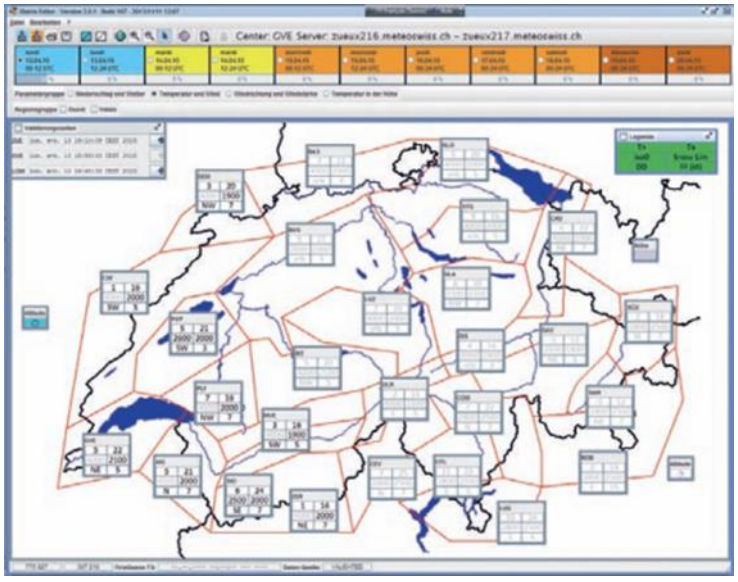

Figure 1: The Matrix graphical user interface that allowed forecasters to modify first guess forecasts from numerical weather prediction models. Editable parameters were sunshine duration, precipitation, minimum and maximum temperature, wind speed, height of zero degree line and the altitude of stratus top. The first two parameters are considered as average values for the corresponding region, while temperature and wind are point forecasts at single reference sites within the region.

In 2021, a fully automated system for producing public weather forecasts for the MeteoSwiss app and website (point forecasts at around 6000 locations) was introduced with the upgrade of the

MeteoSwiss
Operation Center 1
CH-8044 Zürich-Flughafen
T +41 58 460 99 99
www.meteoschweiz.ch

Data4Web system. It removed the need for manual input by forecasters and allowed us to integrate new forecast information from Numerical Weather Prediction (NWP) systems more quickly. This change in forecast processing also affected the headline score, in that the data source originally used to compute the COMFORT score – manually edited regional forecasts from the Matrix - was no longer available. With the introduction of fully automated processing for point forecasts, COMFORT is now computed with the same output as shown in the mobile app and on the web. This change also allowed us to evaluate the quality of the forecasts at a much larger number of sites for better representativeness. The error metrics and temporal resolution of the forecast used in the headline score, however, have not changed.

To distinguish the two versions of COMFORT, we refer to the original setup using manually modified regional forecasts as COMFORT2013, and the new setup evaluated at a larger number of sites as COMFORT2020 from here on. An overview of the changes to COMFORT introduced with the change in forecast production is shown in Table 1.

Table 1: Number of stations used to produce the forecasts and observations in COMFORT2013 and COMFORT2020. A higher number of stations is used for COMFORT2020 as it uses forecasts from Data4Web. The number of stations in COMFORT2020 represents the state of the MeteoSwiss observing network SMN in September 2022. This network may potentially grow in the future (MeteoSwiss, 2015).

	COMFORT2013		COMFORT2020	
Parameter	**Forecast**	**Observation**	**Forecast**	**Observation**
Daily total precipitation	Regional means by forecasters	CombiPrecip regional mean	Regional means from forecasts at ~300 SMN stations	CombiPrecip regional mean
Daily relative sunshine duration	Regional means by forecasters	Regional means from 80 SMN stations	Regional means from forecasts at 80 SMN stations	Regional means from 80 SMN stations
Daily minimum and maximum temperature	27 reference stations	27 reference stations	Forecasts at 152 SMN stations	152 SMN stations
Maximum hourly wind speed during the day	6 reference stations	6 reference stations	Forecasts at 89 SMN stations with heights lower than 1000m	89 SMN stations with heights lower than 1000m

2.1 Data4Web as the new data source

Data4Web is the operational production system of MeteoSwiss that produces seamless weather forecasts for seven days ahead. It is based mainly on output from the numerical weather prediction systems COSMO and IFS. The output of these ensemble systems is interpolated to produce hourly forecasts for more than 5000 POIs in Switzerland. These forecasts are displayed on the mobile application and the website. Data4web already produced almost all the parameters used for the calculation of COMFORT2013, but at that time the computation of COMFORT was still exclusively based on the regional forecasts produced in the Matrix.

Unlike the deterministic forecasts in the legacy system, which consist of a single scenario, ensemble forecasting produces a multitude of slightly different scenarios per forecast, which constitute the ensemble forecast (MeteoSwiss, 2016). For example, COSMO-2E produces 21 different scenarios per forecast. This ensemble of equi-probable forecast scenarios allows to issue probabilistic forecasts, for which we can quantify forecast uncertainty and assess forecast. A large proportion of the forecasting products still uses deterministic forecast information (i.e. forecasts without quantified uncertainty). In Data4Web, the median is used to obtain a single forecast from an ensemble.

MeteoSwiss
Operation Center 1
CH-8044 Zürich-Flughafen
T +41 58 460 99 99
www.meteoschweiz.ch

3 Decline of COMFORT during 2021

When the new COMFORT2020 and Data4Web have been released, there was a clear drop in the scores after April 2021. The drop of global scores for Switzerland is more pronounced for higher lead times (Figure 2). The performance loss was very strong and the scores dropped well below the targets.

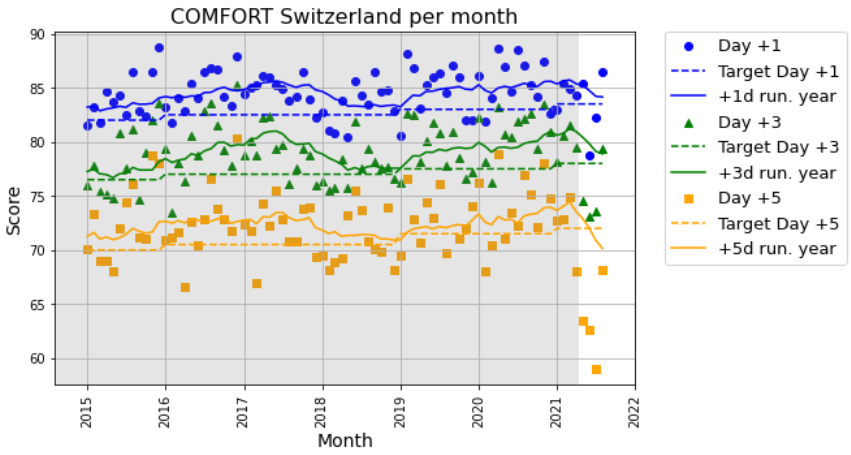

Figure 2: Global COMFORT score between January 2015 and August 2021. The grey area corresponds to the period before the transition to COMFORT2020. The dots, triangles and squares correspond to the monthly COMFORT values for day +1, day +3 and day +5. The solid lines correspond to the annual averages and the dashed lines correspond to the agreed objectives. A decrease in performance can be noticed from April to August 2021 for all lead times.

Table 2 shows the relative difference of the results obtained for the months of May, June and July 2021 to their corresponding averages between 2015 and 2020. A larger decrease is clearly visible for day 5. In the following, we analyze the score drop in more detail based on the partial scores for individual forecast parameters.

Table 2: Relative difference of the global COMFORT scores of May, June and July 2021 to their corresponding average between 2015 and 2020. The values are in percentage and show a significant decrease for the day +5.

	Day +1	Day +3	Day +5
May 2021	+0.9%	-5.7%	-11.8%
June 2021	-7.4%	-9.2%	-14.7%
July 2021	-3.9%	-8.1%	-20.0%

MeteoSwiss
Operation Center 1
CH-8044 Zürich-Flughafen
T +41 58 460 99 99
www.meteoschweiz.ch

3.1 Precipitation

The precipitation score dropped sharply during the months of May and July 2021 over the full forecast time range (see Figure 4). Two main factors influenced these poor results. The main reason is the aggregation of hourly forecast values to produce daily amounts. Data4web produces hourly quantiles (i.e. q0.1, median, q0.9) as required by the App and website to provide hourly probabilistic information to the users. The daily aggregation for COMFORT was done a posteriori with another system (in the centralized DataWareHouse of MeteoSwiss) by summing the median of the hourly rainfall from Data4Web. While summing hourly values to produce a daily quantity is correct for traditional deterministic forecasts, it leads to a strong underestimation when done with the hourly median. With ensemble forecasts, order of operations matters. The median of daily total precipitation – the target quantity used for verification – is not the same as the daily sum of the hourly median of precipitation. The second reason is the very unpredictable weather in the summer of 2021, which was mostly rainy (MeteoSuisse, 2022). The COMFORT score is highly dependent on weather events (Cattani, et al., 2019).

3.2 Relative sunshine duration

The partial scores for relative sunshine duration degraded exclusively for longer lead times (Figure 3). The analyses showed that this loss of performance was greater for the east and west regions than the south region, which generally has higher sunshine duration (not shown). Forecasts beyond Day 4 are based on the coarse-resolution global NWP system IFS from the European Centre of Medium Range Weather Forecasts (ECMWF). The poor relative sunshine results for longer lead times can be attributed to a large positive sunshine duration bias of ECMWF-IFS (ECMWF, 2022).

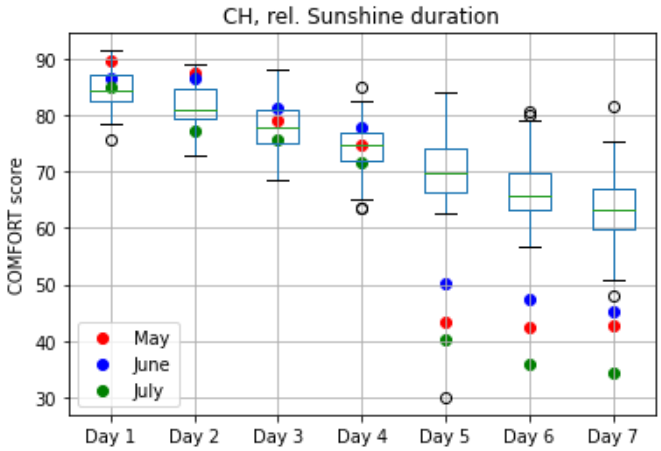

Figure 3: Monthly relative sunshine duration scores as a function of lead time for whole Switzerland; May 2021 (red points), June 2021 (blue points), July 2021 (green points), boxplots of monthly scores (2014.04-2021.04).

3.3 Wind

The partial scores for the wind forecasts are very low for lead day 5 (Figure 4). This is attributed to the forecasts produced by IFS-ENS, which is used as input for Data4Web from day 5 onwards. First, due to its coarse resolution (18km) and thus a poor representation of alpine valleys, the IFS-ENS model is not able to generate local wind systems such as valley breeze, which are dominant in the summer period. Potentially even more importantly, the orographic drag, parameterized in IFS-ENS, reduces wind speed over topography which leads to considerable underestimation (Sandu, et al., 2017). Consequently, we observe a negative bias with IFS-ENS daily maximum hourly wind speed which is prevalent for all SMN stations, the negative bias is strongest for exposed high-altitude sites with an underestimation of up to 8 m/s on average, and weakest for low-altitude stations of the Swiss plateau with underestimation of around 1 m/s (not shown).

For the original COMFORT based on Matrix forecasts, the six stations used were located in the plains and valleys. Moreover, the wind forecasts came from MosMix post-processing as first guess and were further improved by meteorologists, thus local effects were fully taken into account. For COMFORT2020, all SMN stations below 1000m.a.s.l. are used to produce the partial scores for the 27 regions, which contain many alpine regions for which the topography strongly affects the local winds.

3.4 Temperature

Unlike the other partial scores, the switch to COMFORT2020 and Data4Web 4 forecasts slightly improved the minimum and maximum temperature scores (Figure 4, Table 3). The good results of the partial score for minimum and maximum temperatures are expected due to the postprocessing for temperatures already available in Data4Web previous to the switch (see Section 4.2). Due to the continued availability of postprocessing, quality of temperature forecasts from Data4Web has not been affected by the switch.

MeteoSwiss
Operation Center 1
CH-8044 Zürich-Flughafen
T +41 58 460 99 99
www.meteoschweiz.ch

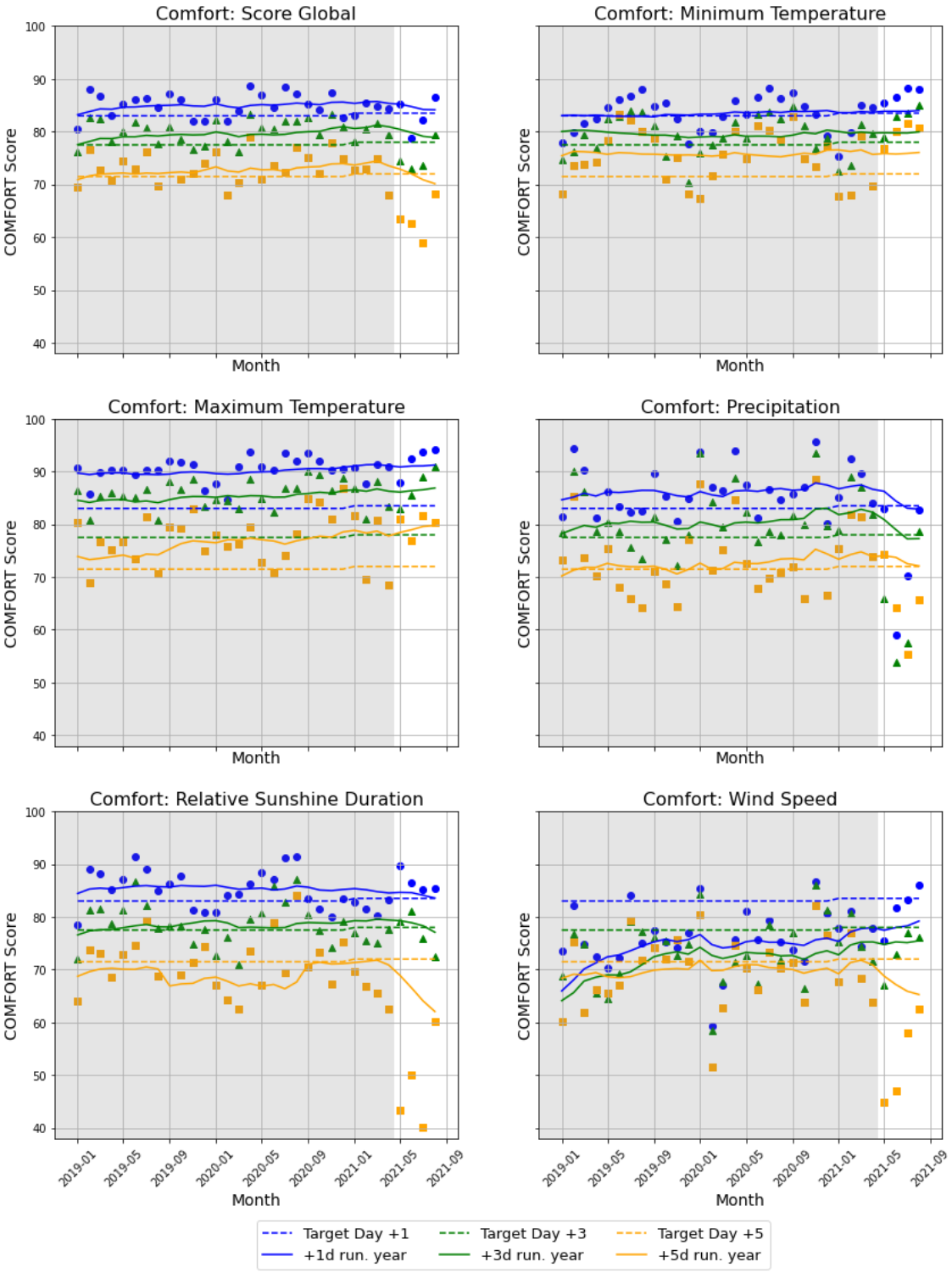

Figure 4: COMFORT scores between January 2019 and August 2021. Due to the poor performance of sunshine and precipitation partial scores and their high weighting when calculating the global score, the global score decreased significantly in the summer of 2021. The dots, triangles and squares correspond to the monthly COMFORT values for day +1, day +3 and day +5. The solid lines correspond to the annual averages and the dashed lines correspond to the agreed objectives. The grey area corresponds to the period before the transition to COMFORT2020.

4 Modifications to Data4Web

To address the degradation in the COMFORT score after introduction of the new processing chain, immediate action was taken to identify the teething problems and eliminate the weaknesses, and several changes were made in Data4Web to improve the score as much as possible. These consist of both technical modifications and qualitative changes to the forecasting data. Therefore, these changes have also improved the quality of data provided to services using Data4Web.

In addition to the modifications presented below, daily forecasts are now produced in both UTC and local time for precipitation, wind and temperature. COMFORT requires UTC times whereas the application and the website require local times (i.e. UTC+1 in winter and UTC+2 during daylight saving).

4.1 Order of operations

The COMFORT scores are calculated using deterministic forecasts. The data output of the new processing system, however, is hourly probabilistic (i.e. ensemble) forecasts. Initially, the daily aggregations needed for COMFORT were calculated by a separate system, using the hourly ensemble median forecasts provided by Data4Web. The daily aggregate of hourly quantiles however is not the same as the quantile of daily aggregated ensemble quantities, with the latter approach being the correct way to compute ensemble statistics on aggregations. With precipitation in particular, computing the daily aggregations on the median hourly values led to a strong underestimation of daily totals, especially in summer convective situations with considerable uncertainty in the timing of precipitation.

To produce meaningful daily statistics from probabilistic forecasts, separate data streams for each aggregation had to be established in Data4Web. As of Data4Web v4.1.8, a separate data stream for daily aggregates exists for daily maximum and minimum temperature, daily precipitation and maximum hourly wind speed during the day. Relative daily sunshine duration is still aggregated in the separate system to benefit from the improved orographic shading at hourly time scales.

4.2 Post-processing

In order to improve the overall performance of the forecasts, some post-processing steps have been added to compensate for the lack of manual modifications by the forecasters as previously provided in the Matrix.

To reduce the IFS-ENS overestimation of relative sunshine duration, a post-processing step was applied, which consists of an empirical formula using cloud layers to approximate sunshine duration.

MeteoSwiss
Operation Center 1
CH-8044 Zürich-Flughafen
T +41 58 460 99 99
www.meteoschweiz.ch

This eliminated the bias in IFS-ENS. In addition, more precise estimates of orographic shading are used. Realistic topographic shading further reduces the positive biases in particular in complex topography, for which direct output from NWP systems underestimates the shading in the valleys due to the limited resolution and therefore smoother topography.

To fix the degradation of wind speed forecast scores beyond Day 4, a similar post-processing as for temperature was introduced in Data4Web. Forecasts from the coarse-resolution IFS-ENS are locally calibrated using MosMix data. Note that at the time of writing of this report, a new post-processing methodology based on machine learning was introduced in Data4Web, which improves the quality of wind forecast as well as the COMFORT score throughout the forecast range.

These postprocessing steps complement the already existing postprocessing of temperature in Data4Web that combines altitude-corrected direct model output from the different NWP systems with locally postprocessed temperature forecasts from the MosMix system (Deutscher Wetterdienst, n.d.) available for a subset of station locations only. To postprocess forecasts everywhere, residuals between MOSMIX and NWP temperature forecasts at stations are interpolated in space. These interpolated residuals are then added to the NWP forecasts to result in spatially complete, calibrated temperature forecasts.

4.3 Introduction of reforecast capability

To be able to evaluate the long-term impact of any methodological change in Data4Web on COMFORT and on forecast quality in general, it was essential to implement the possibility to produce forecasts for past situations, so called reforecasts. This reforecasting capability was implemented so that Data4Web can be calculated back over several years and its output evaluated in terms of COMFORT scores in semi-automated reforecast experiments. This allowed to test different methodological choices and take informed decisions. To secure continuous improvement in quality and avoid surprises in the future, for each new release of Data4Web, a reforecast will be produced to assess the effect of the modifications on several verification scores including COMFORT. This capability was missing during the transition to COMFORT2020 and it was therefore impossible to experiment and evaluate the impact of the switch from the Matrix to Data4Web as input data for COMFORT. With such tools, it would have been possible to anticipate the decline of the score and thus make the modifications presented above before the operational introduction.

5 Recovery of COMFORT in 2021 and 2022

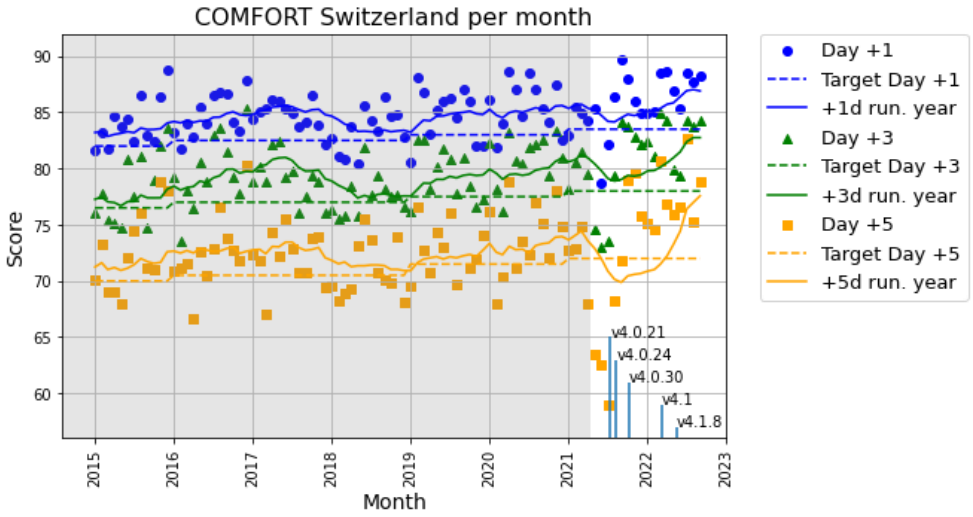

Figure 5: Global COMFORT score between January 2015 and September 2022. The grey area corresponds to the period before the transition to COMFORT2020. In the bottom right corner, you can find indications about the Data4Web releases that improved the COMFORT score

After multiple corrections and updates to Data4Web (Table 4 in appendix), the scores fully recovered from September onward, meeting again the agreed-on targets for all three forecast lead times (Figures 5 and 6). In 2022, scores were higher than at any time in the COMFORT2013 period (Figure 5).

MeteoSwiss
Operation Center 1
CH-8044 Zürich-Flughafen
T +41 58 460 99 99
www.meteoschweiz.ch

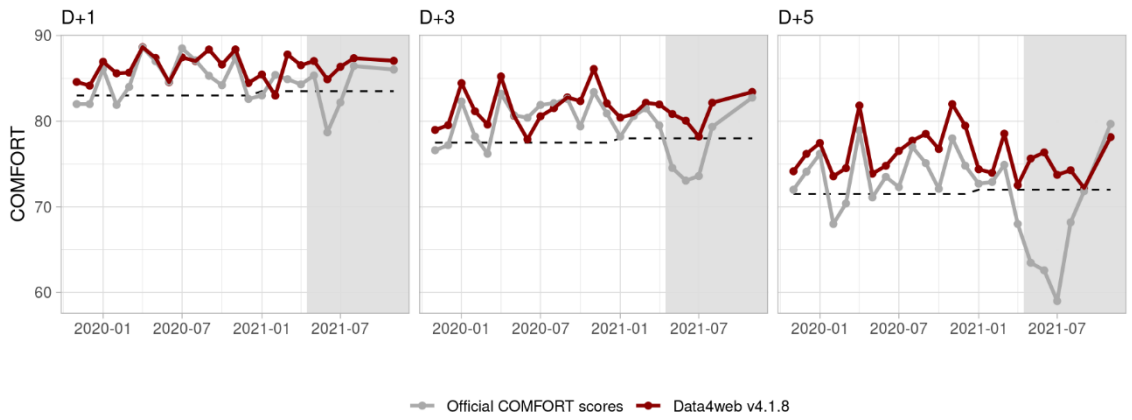

Figure 6: Global COMFORT score computed with the Data4Web release v4.1.8 as well as the official score for the three forecast lead times 1, 3 and 5 days. (November 2019 - November 2021). The last modification concerning the forecasts for COMFORT was made in vers 4.1.8 of Data4Web.

Figure 6 compares the official score with the one recalculated using Data4Web version 4.1.8 (the latest release of Data4Web concerning COMFORT at the time of writing), for a two-year period from November 2019 to October 2021. Experiments using reforecasts showed that if Data4Web had been released in production immediately with all the updates and corrections presented in this report, there would have been no loss of performance during mid-2021. In addition, according to our analysis using reforecasts, the new system based on Data4Web version 4.1.8 performs better than the previous system used before May 2021 for all parameters but wind which was calculated for only 6 stations (Table 3).

Table 3: Change in COMFORT scores with Data4Web v4.1.8 with respect to the official ones based on the Matrix before May 2021. Positive values indicate an improvement of the score computed on the fully automatic system Data4Web for the period November 2020 and May 2021.

Lead	*Global score*	*Precipitation*	*Sunshine*	*Tmax*	*Tmin*	*Wind*
D+1	1.3	1.3	1.3	0.5	1.6	2.1
D+3	1.3	2.3	0.0	1.6	2.7	-0.5
D+5	3.0	3.2	5.4	2.4	1.1	-0.4

5.1 Influence of the weather conditions

The score of COMFORT is highly dependent on weather conditions (Cattani, et al., 2019), since some meteorological situations are much more predictable than others. The variability of the score depends on the predictability of day-to-day weather. For instance, when weather is dominated by large-scale high-pressure systems, dry conditions can be forecast with high predictability days in advance. As a result, the partial score for precipitation can easily reach 100%. On the other hand, during dynamic convective situations in summer, the correct amounts and locations of precipitation can

be very hard to predict days in advance, even on regional scales. The same applies for relative sunshine duration, which is easier to predict during persistent clear sky conditions assured by high-pressure systems (except in fall and winter during stratus situations), compared to periods with highly variable cloud cover.

Therefore, during months with positive high-pressure anomalies, the score is usually higher than the one of a month with frequent showers. This may partly explain the excellent results in 2022, a very dry year (Figure 5) as well as the drop in the score during the summer of 2021, one of the wettest summers on record for the regions north of the Alps (MeteoSuisse, 2022). This drop in the COMFORT score for summer 2021 is also visible in the reforecasts with the latest version of Data4Web (Figure 6).

MeteoSwiss
Operation Center 1
CH-8044 Zürich-Flughafen
T +41 58 460 99 99
www.meteoschweiz.ch

6 Lessons learned

6.1 Planning

MeteoSwiss has learned a lot from the transition to the fully automated forecast production system and the resulting problems in COMFORT scores. The decision to stop the Matrix and switch the data basis for the headline score to Data4Web came halfway through a project involving a full redesign of Data4Web, adding new challenges to an already ambitious task. It also added pressure on the project to respect a hard deadline as it affected the operational forecasting services shift planning and thus needed coordination. Given the importance of the COMFORT score, not enough resources and time were planned to assess the effect of this transition. Furthermore, the impact of such changes on the score was clearly overlooked. Too much trust was given to the raw NWP models and the added value of the forecaster to manually adjust known biases of the NWP was underestimated. The dip in the headline score could have been avoided if more resources and management precautions had been taken. Indeed, the score transition was a very intense period and required the coordination of many heterogeneous services and departments with deadlines to meet in a context of mandatory home-office because of the pandemic, making it even more difficult.

In hindsight, it would have been wise to allow more time for the transition and set the reforecast capability as a strict requirement before going into production. It would also have been easier to make the transition in several smaller steps and be more flexible with the deadlines for the different forecasting services. Of course, studies were carried out on the effect on COMFORT of the switch to a fully automated system. But these were in hindsight misleading due to the lack of reforecasting capabilities of the candidate system and as a consequence, important aspects were overlooked such as the need for separate data streams for daily quantities to respect the order of operations. Also, Data4Web was designed to fulfill the technical requirements of the mobile app and the website with less focus on forecast quality. As a consequence of missing the performance targets, careful evaluation of the impacts forecast quality and on the headline score COMFORT has become a top priority. All these points are now considered when the forecasting system to produce local forecasts at MeteoSwiss is changed.

6.2 Improved monitoring

Following the degradation of forecast performance in May 2021, several tools have been implemented to facilitate monitoring and quality control for the automatic local forecasting system. Data4Web's reforecast capability, released in Data4Web on March 8, 2022, now allows changes in

Data4Web to be evaluated over a long period of time before changes are introduced in production. Such reforecasts not only allow to quantify the effect on the headline score, but also to monitor the impact of changes on all aspects of the forecasts with a variety of forecast quality metrics. In addition, daily COMFORT scores are monitored to identify problems with the COMFORT score calculation with automated alerting for missing values in partial scores. A comprehensive visualization tool for the COMFORT score (Figure 12) provides quick overviews and drill-downs for partial scores and individual regions and thereby helps to understand and fix degradations to the headline. The improved monitoring and quality assurance is a key component for the robust operations of our automated system for local forecasts.

6.3 Headline score

COMFORT, the current MeteoSwiss headline score for public point forecasts, has been designed at a time when the forecast production and dissemination was radically different from the current situation. In particular, COMFORT is based on the fact that a human forecaster can issue forecasts only for a very limited number of locations and in a deterministic manner. In contrast, probabilistic forecasts at thousands of locations in Switzerland are currently being provided to end users. Apart from COMFORT not assessing the probabilistic forecasts fully and thus missing an important part of the story, past design choices also pose problems throughout the processing chain, which has to be updated to take the specifics of probabilistic ensemble forecasts into account. In addition, with its focus on daily regional values, COMFORT fails to capture the increasing resolution of our point forecasts. Furthermore, the risk of a saturation is continuously increasing. By definition, it becomes more and more difficult to improve.

These limitations lead to the COMFORT score not measuring the relevant aspects of the forecasts as communicated to the public. The MeteoSwiss headline score thus needs to be redesigned and a few first ideas are presented in the following section.

7 Towards a new headline score

COMFORT, the headline score used to assess forecast quality of MeteoSwiss' public weather forecasts has been designed at a time when the forecasting system was radically different from now. As such the headline score doesn't align well with the format and presentation of MeteoSwiss' public weather forecasts, i.e. seamless probabilistic hourly forecasts for many thousand distinct locations in Switzerland. In addition, the COMFORT score suffers from a few additional shortcomings:

- **Saturation:** COMFORT scores become increasingly difficult to improve in particular for forecasts at short lead time as the score is bounded to 100.
- **Adoption:** COMFORT is not widely adopted. Although the COMFORT framework was created with the aim of being reusable and modifiable by other weather organizations, it is only used by MeteoSwiss.
- **Variability:** The score is affected by variability in prevailing weather. For example, the score is lower in rainy conditions.
- **Representative:** The score is evaluated at stations that are also used for the postprocessing including MOSMIX. Therefore, the score does not describe the real quality of the forecast for any arbitrary location (POI).
- **Consistency:** Precipitation and sunshine duration are currently treated not as local, but as regional quantities. Improvements in spatial resolution for these parameters are thus not reflected in COMFORT.
- **Update frequency and forecast range:** With its focus on forecasts multiple days ahead, COMFORT does not allow to measure relevant aspects of our local forecasts including the quality and timeliness of short-term forecasts for the next minutes and hours.

Various alternatives have been explored to address some of these issues and to replace the COMFORT score in the future.

7.1 Score saturation

To address the saturation of the score, a transformation of the score values could be used. The skillful days representation corresponds to the number of lead days for which the scores are above a certain target value. In other words, applied to COMFORT, it is the forecast range at which the COMFORT scores are above a threshold defined by the user or the administration depending on their requirements (Figure 7). This transformation of (skill) scores into forecast lead time by which a certain (skill) score is achieved has been pioneered by ECMWF and is used in their headline score.

MeteoSwiss
Operation Center 1
CH-8044 Zürich-Flughafen
T +41 58 460 99 99
www.meteoschweiz.ch

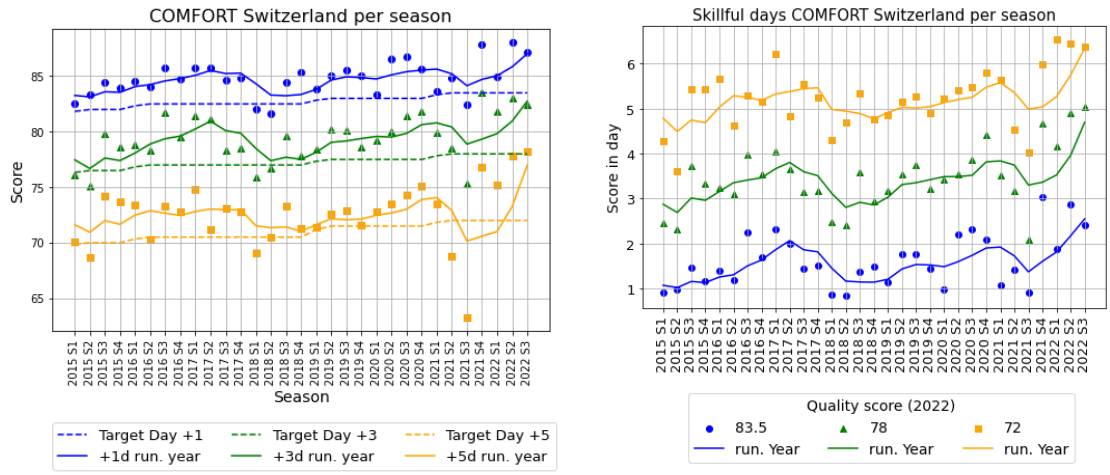

Figure 7: On the left, a graph showing the Comfort score for day 1, 3 and 5 between 2015 and August 2022. On the right, a graph representing the skillful days score for the targets 72, 78 and 83.5 between 2015 and March 2022

This approach is not limited to the COMFORT score and could be applied to other forecast performance measures such as the headline score for NWP forecast (Kaufmann, 2019). One of the requirements for using this skillful days representation is that the scores must be calculated for a sufficiently long lead-time period (e.g., 10 days) and the target has to be well chosen. Indeed, the score should reach the target value within the forecast horizon, otherwise an extrapolation would be necessary to compute skillful days. Extrapolation is problematic and has been encountered in the calculation of the skillful days score on the headline score for COSMO-2E (MOQUAS) with its forecast horizon of 120 hours, see Figure 8.

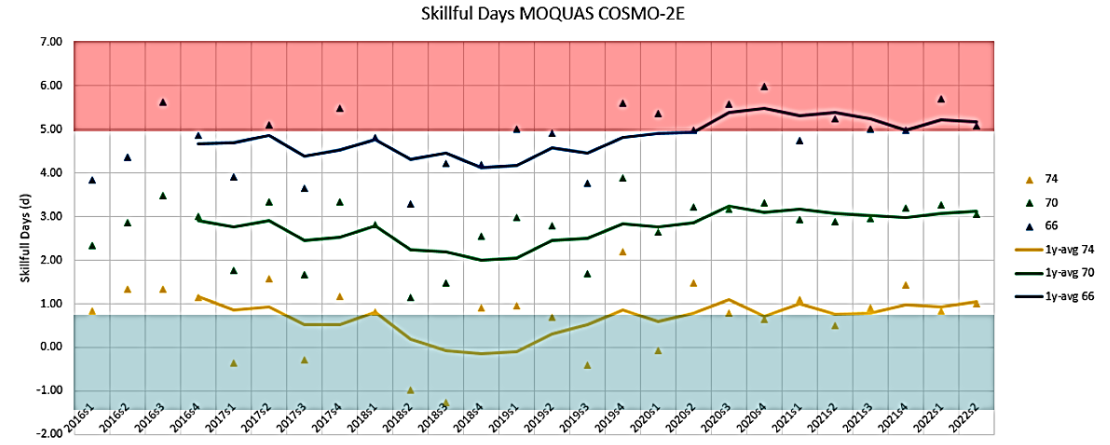

Figure 8: Skillful days scores of the MOQUAS on COSMO-2E for the thresholds 74% (yellow), 70% (green), and 66% (blue). The calculation technique to compute the skillful days score applied on the MOQUAS uses regression. The red zone corresponds to the extrapolation for values higher than 5 days, the maximum range of COSMO-2E. The light blue zone corresponds to the extrapolation of values smaller than 18 hours. It contains even negative skillful days for seasons with all MOQUAS scores below the chosen threshold value. (Figure by P. Kaufmann)

The choice of threshold can be non-trivial and is highly dependent on the parameter and time period considered. As the wind score is generally much lower than the precipitation or temperature scores, it could be worth considering using different thresholds for each parameter, which would complicate the interpretation of the score but help to avoid extrapolation.

7.2 Proper scoring rules for probabilistic forecasts

Probabilistic forecasts are known to better support decisions as such forecasts not only provide a most likely outcome, but also quantify forecast uncertainty. With the increasing adoption and dissemination of probabilistic forecasts, scores that are able to quantify accuracy of such probabilistic forecasts need to be used. One such candidate is the Continuous Ranked Probability Score (CRPS), a widely adopted and well-established proper scoring rule for probabilistic forecasts (Hersbach, 2000).

The CRPS is the generalization of the Mean Absolute Error for probabilistic forecasts and as such measures the forecast error of a probabilistic forecast in units of the data used. The CRPS evaluates both the sharpness (how uncertain is a forecast) and the calibration (are there systematic biases in the forecast) of probabilistic forecasts, with forecasts that are less uncertain (sharper) and closer to the verifying observation (unbiased) being awarded a better (lower) score.

7.3 Weather-induced variability

All forecast verification scores are affected by prevailing weather in the sense that different weather generally results in different scores. This is particularly noticeable in a combined score such as COMFORT when the variability in various components is correlated. For example, during sunny and dry weather, the COMFORT score is higher (meaning the forecasts are better) due to elevated partial scores for precipitation and sunshine duration (predictably dry and sunny conditions result in perfect scores) and wind speed (dry, sunny conditions are often associated with low wind speeds). Thus the COMFORT score measures the prevalence of weather conditions that are easier to forecast rather than forecast quality. The effect of weather variability on forecast quality metrics can be reduced by averaging across many forecasts as is done in COMFORT, for which performance targets only apply to annual global scores.

To better understand and diagnose deficiencies of the forecasting system and to more closely monitor the evolution of forecast quality, it would be preferrable to further limit the effect of weather variability. One way to achieve this is to measure forecast performance relative to a reference forecasting system. These reference forecasts are themselves affected by weather variability, and by measuring relative performance, the effect of weather variability may cancel out. To be able to monitor the long-term evolution of our forecasting system, reference forecasts have to be produced with the same system for the entirety of the time. Operational forecasts (e.g. from NWP systems) therefore cannot be used, as these typically evolve and improve over time as well. In contrast, reanalyses are produced with systems that are as stable over time as possible and reanalysis systems sometimes also produce long forecasts (i.e. forecasts beyond the analysis interval that is necessary to run the reanal-

ysis anyway). The usefulness of such reanalysis forecasts from ERAinterim was assessed in a collaboration with Mario Gerber and Daniela Domeisen (ETHZ) in 2019-20. The findings are summarized below.

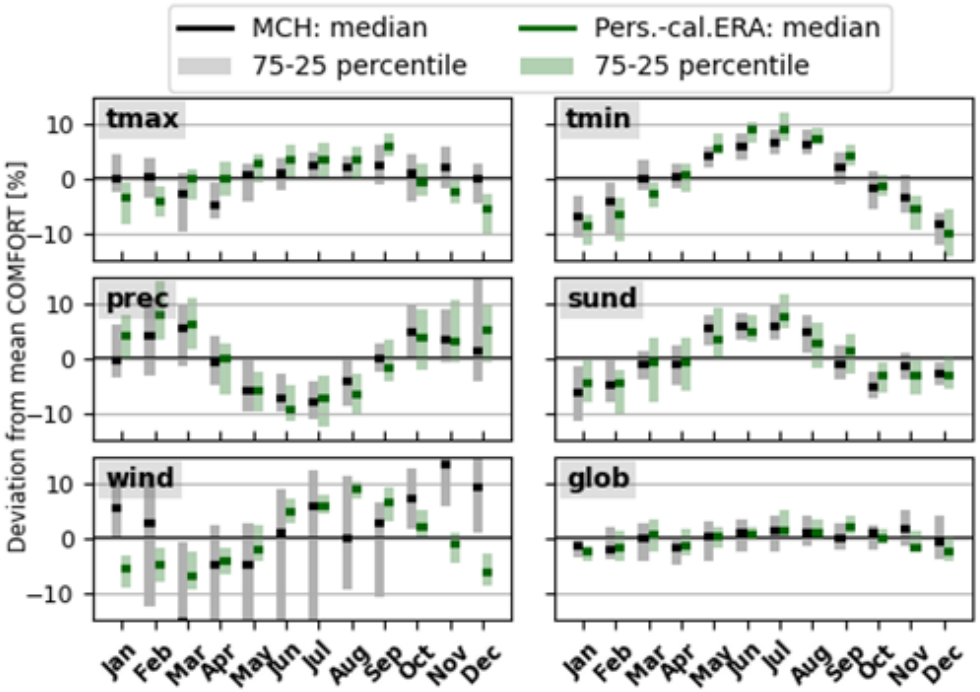

Figure 9: Distribution of monthly global COMFORT anomalies with respect to mean from 2010-2019. MeteoSwiss official COMFORT scores are shown in black, COMFORT scores computed with bias corrected ERAinterim reforecasts are shown in green. The deviation from the overall mean COMFORT of the respective lead time is expressed in percentage and thereafter aggregated over lead times D+1 to D+7.

Using reforecasts from ERAinterim to reduce weather-induced variability in COMFORT (and other scores) is not straight-forward for various reasons. ERAinterim forecasts exhibit significant biases due to the coarse resolution of the global NWP. This affects forecast performance in general, but also the relationship and sensitivity to prevailing weather conditions. While various alternatives for debiasing (postprocessing) exist, even after correcting for the first-order biases, the relationship of COMFORT with prevailing weather is different from that of the official COMFORT scores (Figure 9). While using a reference forecast can reduce weather-induced variability, often this only works well for highly aggregated quantities such as annual global COMFORT scores. When the score is broken down for individual parameters in individual regions, the scores do not vary in sync due to slightly different characteristics of the forecasts and thus computing score differences or relative improvement with respect to the reference forecast increases variability rather than reducing it. Please note that these are the conclusions from the comparison of the legacy local forecasting system at MeteoSwiss (semi-automated) and a coarse resolution reforecast from ERAinterim. These two forecasting systems differ enough to make them behave fairly independently for all but the most aggregated quanti-

ties. If on the other hand a reference forecast can be found or constructed that is very similar in quality to the target forecast (i.e. the MeteoSwiss local forecasts), normalization with the reference forecast might be more promising.

Apart from the scientific challenges, there are also operational challenges that may render reference forecasts difficult to use in practice. First and foremost, the timely computation and dissemination of reforecasts needs to be guaranteed to be able to use in a headline score. If reforecasts are only updated episodically, real-time monitoring using the headline score becomes impossible. Furthermore, continuity of the reforecasts needs to be guaranteed to allow to compute performance measures for multiple years (decades) of operational forecasts. With ECMWF using reforecast-based reference forecasts (Haiden, et al., 2021), these operational challenges seem solvable.

MeteoSwiss
Operation Center 1
CH-8044 Zürich-Flughafen
T +41 58 460 99 99
www.meteoschweiz.ch

Summary

The year 2021 has been a period of important changes for the forecasting and verification procedure of MeteoSwiss. Data4Web, the data backend providing local forecasts to the MeteoSwiss mobile App and Website, has been redesigned and refactored. This allowed to decommission the Matrix, a part of the semi-automated legacy forecasting system supervised and manually updated by forecasters. With this change from the semi-automated legacy system to the fully-automated system to produce local forecasts, also the data basis to assess forecast quality with the MeteoSwiss headline score COMFORT changed. The data basis changed from regional forecasts for 27 regions in Switzerland in the legacy system to local forecasts at the more than 130 stations of the MeteoSwiss automatic measurement network. The MeteoSwiss headline score is now computed with the same forecasts that are fed to the MeteoSwiss mobile app and website and thus more closely measures the quality of the forecasts as exposed to the general public.

In the process of the transition to the automated system to produce local forecasts, several shortcomings in the forecast production chain were exposed and fixed. A degradation of COMFORT was observed after the switch during summer 2021 and performance targets were missed in particular for forecasts for day +5 which are based on output from the coarse resolution global weather forecasting model IFS-ENS. Various improvements have been implemented in Data4Web to again reach the performance targets. Most notably, postprocessing has been introduced to adjust systematic forecast biases of the weather forecasting models. Also, separate data streams for daily quantities evaluated by the headline score had to be introduced to fully benefit of the probabilistic forecasts. Since these improvements, performance targets have been reached again. Analysis also shows that the current system performs better than the legacy one.

MeteoSwiss
Operation Center 1
CH-8044 Zürich-Flughafen
T +41 58 460 99 99
www.meteoschweiz.ch

References

Cattani, D., Faes, A., Giroud Gaillard, M. & Matter, M., 2015. *COMFORT: Continuous MeteoSwiss Quality Score,* s.l.: MeteoSwiss, Scientific report.

Cattani, D. et al., 2019. *Analysis of Variations in Forecast Quality during 2018,* s.l.: Federal Office of Meteorology and Climatology, MeteoSwiss.

Deutscher Wetterdienst, n.d. *Model Output Statistics-MIX (MOSMIX).* [Online]
Available at:
https://www.dwd.de/DE/leistungen/met_verfahren_mosmix/met_verfahren_mosmix.html

ECMWF, 2022. *Known IFS forecasting issues.* [Online]
Available at: https://confluence.ecmwf.int/display/FCST/Known+IFS+forecasting+issues
[Accessed 1 December 2022].

Golding, B. W., 1998. *Nimrod: A system for generating automated very,* s.l.: s.n.

Haiden, T. et al., 2021. *Evaluation of ECMWF forecasts, including the 2021 upgrade,* s.l.: ECMWF.

Hersbach, H., 2000. Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. In: *Weather and Forecasting.* s.l.:s.n., pp. 559-570.

Kaufmann, P., 2019. *Model Quality Score for Switzerland MOQUAS,* s.l.: MeteoSwiss.

MeteoSuisse, 2022. *Bulletin climatologique année 2021,* s.l.: MeteoSuisse.

MeteoSwiss, 2014. *Données spatiales CombiPrecip.* [En ligne]
Available at:
https://www.meteosuisse.admin.ch/home/recherche.subpage.html/fr/data/products/2014/donnees-spatiales-combiprecip.html?query=combiprecip&topic=0&pageIndex=0&tab=search_tab#

MeteoSwiss, 2015. *SwissMetNet – le réseau national de mesures au sol pour la météorologie et la climatologie.* [En ligne]
Available at:
https://www.meteosuisse.admin.ch/home/actualite/infos.subpage.html/fr/data/news/2015/10/swissmetnet-le-reseau-national-de-mesures-au-sol-pour-la-meteorologie-et-la-climatologie.html

MeteoSwiss, 2016. *Qu'est-ce qu'une prévision d'ensemble et comment est-elle générée ?.* [En ligne]
Available at: https://www.meteosuisse.admin.ch/home/actualite/meteosuisse-blog.subpage.html/fr/data/blogs/2016/5/Qu-est-ce-qu-une-pr%C3%A9vision-d-nsemble-et-comment-est-elle-generee.html

Sandu, I., Zadra, A. & Wedi, N., 2017. *Impact of orographic drag on forecast skill.* [Online]
Available at: https://www.ecmwf.int/en/newsletter/150/meteorology/impact-orographic-drag-forecast-skill
[Accessed 1 December 2022].

MeteoSwiss
Operation Center 1
CH-8044 Zürich-Flughafen
T +41 58 460 99 99
www.meteoschweiz.ch

Acknowledgements

The authors would like to thank all colleagues at MeteoSwiss who contributed materially or conceptually to the transition to the new COMFORT2020 and to the writing of this report. We thank Colombe Siegenthaler and Josué Gehring for their attentive readings of a preliminary version of this text. Their willingness to spend time on the report has been very much appreciated. We thank Pirmin Kaufmann for preparing and providing a graph of the new headline score visualization on MOQUAS and his meticulous reading of a final version of the report. Moreover, we would like to express our gratitude to Daniel Cattani for his valuable work on COMFORT during the last ten years. His constructive feedback has always been appreciated. Finally, we would like to thank Lionel Moret for his continuous support as the team leader.

MeteoSwiss
Operation Center 1
CH-8044 Zürich-Flughafen
T +41 58 460 99 99
www.meteoschweiz.ch

A Release history of Data4Web

Table 4: The history of the Data4Web releases modified calculations of the forecasts used in COMFORT.

Release version	Data	Description
4.0	April 27th, 2021	Release of the new Data4Web
4.0.21	July 13th, 2021	• Use CombiPrecip to produce daily precipitation observations instead of the station measurements. • Replace formulas to produce daily forecasts from probabilistic hourly forecasts. Replace sum of the hourly medians by the median of the daily sums.
4.0.24	August 10th, 2021	• Adjust IFS-ENS sunshine duration to match MCH 200 W/m^2 definition and improve COMFORT (mandatory). • Calibration with MosMix wind data and high-resolution topographic descriptors to correct IFS-ENS wind biases due to its low resolution and absence of valley winds.
4.0.30	October 5th, 2021	• Produce daily (24h) precipitation in UTC (used for COMFORT) and LT (used for App & Web). • Improve the algorithm to derive sunshine duration from cloud cover, by using varying weights for cloud layers according to the sun elevation.
4.1	March 8th, 2022	• Production of daily minimum/maximum temperature for UTC days instead of LT. • Automatized the reforecast capabilities of Data4Web which allowed us to run various experiments and make a systematic assessment of COMFORT over two full years. • Fine tuning: interpolation methods, spatial smoothing, correction of sunshine duration according to observed topographical shading.
4.1.8	May 17th, 2022	• New aggregation of maximum hourly wind speed between 06:00 and 18:00 UTC. This aggregated parameter used for COMFORT verification was computed in DWH from the hourly wind speed values. Now it is produced directly by Data4Web which increases the partial wind COMFORT score by about 2.5 points for D+3 (based on 2 years of reforecasts).

Acknowledgements

MeteoSwiss
Operation Center 1
CH-8044 Zürich-Flughafen
T +41 58 460 99 99
www.meteoschweiz.ch

B COMFORT score definitions

The goal of the score is to describe general forecast quality by a single value, considering the most relevant parameters (temperature, precipitation, sunshine duration and wind speed). It is an index from 0 to 100 evaluating the forecasts on a daily basis. The score measures for 7 days lead time the quality of the forecasts for daily minimum and maximum temperature (T_{min} and T_{max}), daily precipitation amounts (P), daily relative sunshine (RS) and maximum hourly average wind speed between 06:00 and 18:00 (V). Each parameter is evaluated independently with a corresponding partial score ($S_{T_{min}}, S_{T_{max}}, S_P, S_{RS}\ and\ S_V$). These partial scores are then aggregated into a weighted average to obtain a global score. The weights differ depending on the measured score (Cattani, et al., 2015).

$$COMFORT = 0.3 * S_P + 0.3 * S_{RS} + 0.15 * S_{T_{min}} + 0.15 * S_{T_{max}} + 0.1 * S_V$$

Partial scores

Each partial score is calculated with specific formulas classifying the forecasts as correct, useful or useless. Correct corresponds to a score of 100, it is obtained for forecasts very close to the observation, a maximum error (tolerance) is defined to qualify a forecast as perfectly correct. The useless classification has a score of 0 and is given to forecasts considered too erroneous to be of any value, i.e. with errors larger than a certain threshold. Between these two error-based thresholds, the forecast is called useful and the score is measured as a continuous quantity. All partial scores follow this same concept. There is a tolerance threshold for the correct class and a utility threshold for the useful class. The thresholds are the same for all regions but differ depending on the parameter. For precipitation and relative sunshine, they vary according to the forecasted value, while for all other parameters they are absolute tolerance thresholds. All the metrics for the five parameters are presented in more detail in a previous report (Cattani, Faes, et al. 2015).

Precipitation

The precipitation partial score measures the daily precipitation between 00:00 and 24:00 UTC. The continuous metric is inspired from the root mean squared fraction score (Golding, 1998). The tolerance for the *correct* and *useful* categories is defined in a relative sense. The score is maximum when the difference between the forecast and the observation is less than 30% of the forecast.

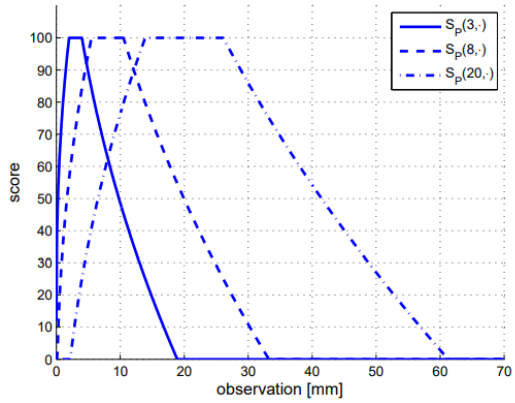

Figure 10: Behavior of the precipitation partial score $S_P(f, o)$ with respect to the observation o, for three different forecast values: f = 3, f = 8 and f = 20 mm.

Relative Sunshine

The daily relative sunshine duration (i.e. relative to the maximum possible for a given day with cloud free conditions) in [%], is classified in 5 sunshine categories: cloudy, mostly cloudy, partly sunny, mostly sunny and sunny. The score maximizes whenever the observation falls into the forecasted category and the score decreases continuously (and linearly) from the bounds of the forecasted category.

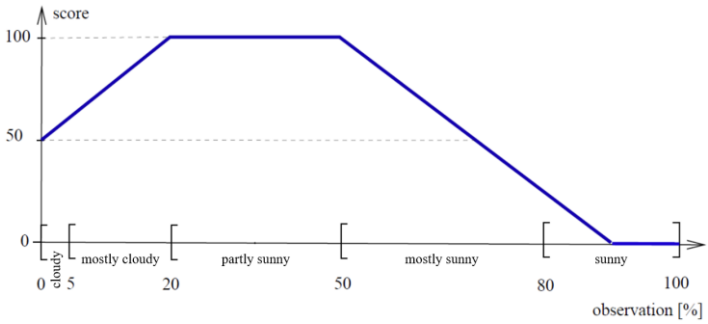

Figure 11: Behavior of the relative sunshine partial score $S_{RS}(f, o)$ with respect to the observation o when the "partly cloudy" category [20, 50[% is forecasted.

MeteoSwiss
Operation Center 1
CH-8044 Zürich-Flughafen
T +41 58 460 99 99
www.meteoschweiz.ch

Minimum temperature, maximum temperature and wind speed

The formula below is used for the partial scores of daily minimum temperature [°C], daily maximum temperature [°C] and wind speed [kt]. For the wind, the quantity evaluated is the maximum hourly average wind speed at 10m between 06:00 and 18:00, while for temperature it is the minimum and maximum temperature over the entire calendar day. The calculation of the COMFORT score corresponds to a transformation of the absolute error into a bounded index from 0 to 100 with a tolerance threshold [μ] and a utility threshold [α] (Figure 11). Unlike precipitation, the utility and the tolerance thresholds are independent of the magnitude of the verified quantity.

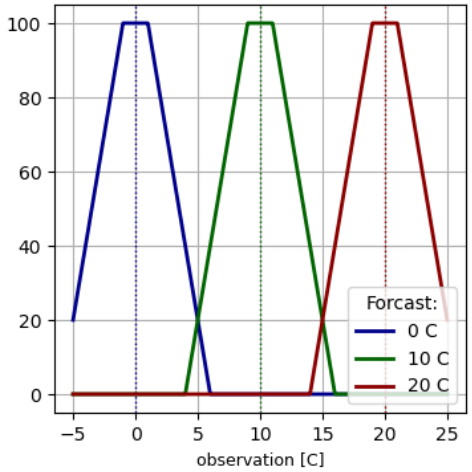

Figure 12: Illustration of COMFORT temperature partial score for three different example forecasts.

Spatial aggregation of COMFORT

COMFORT scores are computed for 27 forecast regions (Figure 1) at daily resolution and for 7 lead days. COMFORT distinguishes two methods for computing regional averages of partial scores, referred to as ‘local’ or ‘regional’. In the regional approach, the predicted and observed quantity represents a spatial average over the forecast region. Therefore, the forecasts and observations are first averaged over all the locations within the region and then the partial score is calculated on these regional averages. In the local approach, the predicted and observed values represent individual sites. Thus, with the local method, the partial score at each location is computed first, then averaged over the entire region to obtain the regional score. Temperature and wind are treated as local quantities due to their strong dependency with local topography, while precipitation and relative sunshine as regional quantities tend to be more homogeneous within regions. Observation data are mainly obtained from MeteoSwiss’ surface observation network, SwissMetNet (SMN, (MeteoSwiss, 2015)). The following sections give the details on the data and aggregation procedures for each parameter at the time of COMFORT introduction in 2013.

Precipitation

Precipitation is defined as a regional quantity. The observed regional precipitation average is computed from a gridded precipitation dataset combining continuous radar images and sparser measurements provided by the SMN rain gauge network (CombiPrecip, (MeteoSwiss, 2014)).

Relative sunshine

Like precipitation, the relative sunshine is a regional quantity, but instead of a gridded observation data set, an average of observations at about 80 SMN stations is used to compare to the regional sunshine duration forecast.

Temperature

Temperature is treated as a local quantity. The maximum and the minimum daily temperature forecasts at 27 SMN reference stations (i.e. one for each region) and corresponding observations are used to compute the partial score.

Wind

Wind speed is a local quantity, and the partial score is evaluated based on the maximum hourly averages (in [kt]) between 6am and 6pm. In contrast to the other quantities, wind is only verified at six SMN stations, mainly because of the poor quality of middle-range and long-range wind forecast in Alpine valley due to coarse resolution models. The six locations (Payern, Sion, Buchs/Aarau, Glarus, Lugano, Acquarossa / Comprovasco) were chosen to capture the dominating winds blowing in Switzerland.

MeteoSwiss
Operation Center 1
CH-8044 Zürich-Flughafen
T +41 58 460 99 99
www.meteoschweiz.ch

MeteoSchweiz
Operation Center 1
CH-8044 Zürich-Flughafen
T +41 58 460 99 99
www.meteoschweiz.ch

MeteoSvizzera
Via ai Monti 146
CH-6605 Locarno Monti
T +41 58 460 97 77
www.meteosvizzera.ch

MétéoSuisse
7bis, av. de la Paix
CH-1211 Genève 2
T +41 58 460 98 88
www.meteosuisse.ch

MétéoSuisse
Chemin de l'Aérologie
CH-1530 Payerne
T +41 58 460 94 44
www.meteosuisse.ch