Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

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

Scientific Report MeteoSwiss No. 99

COMFORT: Continuous MeteoSwiss Forecast Quality Score

Daniel Cattani, Anna Faes, Marianne Giroud Gaillard, Michel Matter

ISSN: 1422-1381

Scientific Report MeteoSwiss No. 99

COMFORT: Continuous MeteoSwiss Forecast Quality Score

Daniel Cattani, Anna Faes, Marianne Giroud Gaillard, Michel Matter

Recommended citation:
Cattani D, Faes A, Giroud Gaillard M, Matter M: 2015, COMFORT: Continuous MeteoSwiss Forecast Quality Score, *Scientific Report MeteoSwiss*, **99**, 45 pp.

Editor:
Federal Office of Meteorology and Climatology, MeteoSwiss, © 2015

MeteoSwiss
Operation Center 1
CH-8058 Zürich-Flughafen
T +41 58 460 91 11
www.meteoswiss.ch

Abstract

Since 1985, MeteoSwiss uses a global score for systematically assessing the general weather forecasts issued by the regional forecasting centers. This assessment is done for the following two main reasons: it is used for administrative purposes, as the weather centers are expected to communicate to the general public and to the government the evolution of the quality of their forecasts. On the other side, the forecasters need to know the performance of their predictions, in order to be able to improve them. In 2013, we developed a new verification scheme which allows to take more benefits of the evolution of the forecasting system as well as of the current automated observation networks. This verification scheme, called COMFORT (for COntinuous MeteoSwiss FORecast qualiTy), was designed for communication purposes and aims to provide the management, but also external entities such as policy makers, media, etc. with a measure of the quality of general forecasts provided by MeteoSwiss. The score COMFORT was developed within the MeteoSwiss operational forecasting system but, in spite of its apparent specificity, it is based on ideas that might be transposable to other weather forecasting services.

Contents

1 Introduction

The COMFORT forecast verification scheme (*Cattani et al.*, 2015) was developed at the Swiss Federal Office of Meteorology and Climatology in order to serve the following administrative purpose: provide different entities, such as hierarchy, policy makers, press, etc. with an overall measure of (some attributes of) the quality of general deterministic forecasts provided by MeteoSwiss. An important issue here was to be able to explain in a simple way the variations of a global score to non-specialists.

Weather forecast verification is a complex and multifaceted area of investigation. A variety of verification methodologies were developed in the last decades, resulting in a profusion of scores assessing various characteristics of the forecasts, see e.g. (*Jolliffe and Stephenson*, 2012) and (*Wilks*, 2011) as reference textbooks. According to Murphy's classification (*Murphy*, 1993), the goodness of a forecast can be decomposed into three types known as forecast *consistency*, *quality* and *value*. In this paper, the focus is on forecast quality, which measures the correspondence between forecasts and observations. Drawing a complete picture of the quality of a forecast leads one to consider a variety of *attributes* (*Murphy*, 1993) contributing to the characterization of the quality, such as *accuracy*, *skill*, *discrimination*, *reliability*, etc. Classical measure-oriented approaches, see e.g. (*Stanski et al.*, 1989), such as mean absolute error or mean-square error usually focus on attributes such as accuracy. A more recent verification framework attempting to assess all attributes of quality in a unified way, thus providing a comprehensive picture of the quality of forecasts, is the distribution-oriented approach (*Murphy and Winkler*, 1987) which is based on the assumption that the joint distribution of forecasts and observations contains all of the non-time-dependent information about the quality of forecasts.

In this work, we shall follow a measure-oriented approach since it has the advantage of assessing quality attributes which are intuitive and easily perceptible by standard customers, that is, persons whose private or professional activities are not crucially affected by weather. Another desirable feature is that this approach preserves temporal and spatial dependency. A requirement that COMFORT should ideally fulfill is that it should encode in a single value the general forecast quality, together with the capability to provide intuitive and intelligible explanation for a high/low global score, typically computed over a long period and on a vast territory, to people that are neither experts in verification, nor forecasters. A way of conciliating these conflicting requirements is to make possible focusing on specific periods and/or geographical areas in order to detect and analyze forecasts whose accuracy deviates from the average. When analyzing score values, placing them in relation with the corresponding meteorological context helps to interpret them. A classical approach, see e.g. (*Wilks*, 2011), consists in computing the skill of the verified forecasts with respect to some reference forecast, such as the *persistent* or the *climatological* forecast, representing the weather variability or anomaly. An alternative approach that will be explored in a future work is to associate COMFORT values with an adequate *predictability index* derived, *a posteriori*, from relevant observations.

In Section 2, we define a *Global Continuous Accuracy Score* (GCAS). A GCAS is a linear com-

bination of partial scores defined for each verified quantity. Each quantity is assumed to be continuous. The forecasting system might in fact impose a categorization of the values that a given quantity can take, e.g., when forecasters are editing forecasts by classes. As we shall see, our definition allows to take such constraints into consideration. Each partial score encompasses tunable thresholds defining what a *correct*, *useful* or *useless* forecast for the given quantity is, as well as a continuous distance-based measure of accuracy. Each partial score is defined on a daily basis, allowing focus at high temporal resolution. The coefficients in the linear combination are weights which reflect the relative importance of the involved quantities; they can be tuned in order to fit any specific requirements.

In our context, it was desirable for communication purposes to perform a verification of quantities reflecting the main features of sensible weather in Switzerland. The COMFORT score is thus a particular case of a GCAS which integrates the following quantities: precipitation (without distinction of its type), sunshine, minimum and maximum daily temperatures, and wind speed. COMFORT assesses general deterministic forecasts edited quantitatively by forecasters for a certain number of regions/locations. These forecasts are either directly delivered to clients or constitute the working basis for the preparation of more specific forecast products. It should be emphasized that some end-products such as warnings or aviation forecasts are not verified by COMFORT as these require specific verification frameworks.

A significant part of the work related to the development of COMFORT was devoted to simulations with the aim to test with real data[1] different properties of the score such as its spatial and temporal variability, its sensitivity to perturbations of different kinds, its ability to reflect theoretical enhancements to the forecasts, and its robustness against hedging. Each quantity involved in the verification was considered separately. The tested forecasts were predictions edited by forecasters, *First-Guess* forecasts obtained from different numerical outputs, and in addition different reference forecasts: persistence for temperatures and various “poor-man” predictions for relative sunshine and precipitation. A selection of results is presented in Section 5.

The robustness of the score against hedging was tested by considering different “no-skill” or “no-risk” forecasts in order to check that there is no obvious systematic way of obtaining better long-term results, at least for short-range predictions, by forecasting some predefined scheme rather than the best-judgement. This should encourage forecasters to issue their forecasts according to their best-judgement (see Subsection 5.3 for more details).

The paper is structured as follows: in Section 2, we formulate and discuss the principles on which the COMFORT score is based. Section 3 contains all the relevant definitions about the score, whereas Section 4 discusses a number of details of the deployment of the COMFORT score at MeteoSwiss. Finally, Section 5 contains examples of how the COMFORT score can be used operationally for analyzing forecasts at different levels of detail, and presents a selection of simulation results.

[1] Most of our simulations were based on past forecasts made at MeteoSwiss during a period of three years running from 2010 to 2012.

2 Verification Principles

In this work, only deterministic forecasts are considered. A frequently used approach by weather forecast providers who communicate their verification results is to compare forecasts and observations on a dichotomous basis. For a continuous quantity, e.g. temperature, this requires fixing a threshold defining the maximal acceptable error and assumes that the forecast is *correct* if it falls within the threshold, and *false* otherwise. Applying this dichotomous test to a sequence of forecast-observation pairs allows one to estimate the forecasts' accuracy by computing, for instance, the percentage correct. The method of fixing the threshold value is largely subjective. For instance, a threshold used for the verification of temperature by several weather providers is of 2°C around the forecasted temperature. Alternatively, widely used classical accuracy measures for continuous forecasts and observations such as the mean absolute error or the mean squared error present the advantage of measuring the average distance between forecasts and observations without categorizing whether the forecast is good or bad.

We propose a simple and intuitive approach which combines properties of dichotomous and continuous verification frameworks. We retain from dichotomous verification the principle of thresholds and shall split forecast's accuracy with regards to three qualifications: *correct*, *useful* and *useless*. In many contexts, it is indeed desirable to have a finer scale for estimating the accuracy of a forecast than only *correct* or *false*. For instance, when verifying a temperature forecast using dichotomy with a threshold fixed at 2°C, an error of 2.5°C has the same impact on the verification result than an error of 5°C: both forecasts are considered as equally false. However, it is likely that an error of 5°C has a worse impact to a customer than an error of 2.5°C. The categorization *useful* allows us to take this aspect into consideration. Also, the impact of an error of a given magnitude might vary depending on whether it is associated with an event close to, or far from, the climatology; or situated around some critical threshold, e.g. the temperature of freezing. From this point of view, criteria for the categorization of forecast's accuracy into the previous three qualifications might depend on the meteorological event that occurs.

The categories *correct*, *useful* and *useless* are defined by two thresholds that should be seen as tunable parameters depending on the verification context. The first threshold defines what we call a *tolerance interval* around the forecasted value. This threshold should be seen as an estimation of the maximum error below which a forecast is assumed as completely correct; this estimation is subjective and it is defined when setting up the verification framework. The second threshold is the maximum error beyond which the forecast is considered too erroneous to be of any value, and defines what we will call the *utility interval* around the forecasted value. Similarly, this subjective threshold is fixed according to the verification context. Between these thresholds, the accuracy of the forecast is measured as for a continuous quantity (for instance using mean absolute error). In the special case, if we set both thresholds equal to each other, we return to the dichotomous framework. On the other extreme, if we set the threshold delimiting the tolerance interval to zero and the threshold defining the

utility interval to infinity, we recover the classical measure-oriented framework for continuous forecasts and observations.

As discussed in the introduction, it is desirable for communication purposes to have a score based on the verification of quantities encoding sensible weather and reflecting the global accuracy of the forecasts as a single value. The simplest way to achieve this is to consider a weighted sum of partial scores computed for each verified quantity. The approach previously explained can be applied independently to each quantity. Before combining scores for different quantities, since errors might be of different magnitudes depending on the verified quantity, it is necessary to rescale them on a common scale. A score valued between 0 and 100 with higher values corresponding to better forecast accuracy, that is, a score with positive orientation, seems to be the most intuitive.

We denote a generic forecast by f and the corresponding observation by o. Let us assume that f and o are real numbers. By *(Continuous) Accuracy Score* (CAS), we mean a (continuous) function of f and o, bounded by 0 and 100, which encompasses:

- a **tolerance threshold** μ: if $|f - o| \leq \mu$ then the score obtained by the pair (f, o) takes the maximum value (=100). The tolerance threshold reflects the principle that an error which is small enough does not affect the quality of the forecast.
- an **utility threshold** α: if $|f-o| > \alpha$ then the score obtained by the pair (f, o) takes the minimum value (=0). The utility threshold reflects the principle that an error which is too large renders the forecast useless.

By definition, $\alpha > \mu$. The thresholds μ and α might depend on f and/or o. Let ERR be any (continuous) metric defined on the real line (for instance, the absolute error). For any pair forecast-observation (f, o), one defines

$$CAS(f, o) = \max\left(0, 100 \cdot \left(1 - \frac{ERR_\mu(f, o)}{d}\right)\right) \tag{1}$$

where $ERR_\mu(f, o) = \min(ERR(o, z); z \in [f - \mu, f + \mu])$ and $d > 0$ is an appropriate normalization constant (for instance, if ERR is the absolute error, then $d = \alpha - \mu$).

Then, we define a *Global (Continuous) Accuracy Score* (GCAS) as a weighted sum of CASs defined for each verified quantity according to (1):

$$GCAS = \sum_{i=1}^{n} \rho_i CAS_i \tag{2}$$

where n is the number of quantities involved and the ρ_i are weights which should always sum to 1.

The tuning of the parameters μ and α in each partial score CAS_i can be done following different approaches, depending on the verification context. For instance, in a customer-oriented system, thresholds might be imposed by each specific client according to his requirements. Differently, thresholds might be set according to the difficulty to predict quantitative values for a given parameter, due for instance to its variability; in this case, thresholds might differ from one region to another depending on the climatology of the region. The resulting score would then be rather a measure of skill than of accuracy.

The thresholds that we have fixed for our verification purposes are mostly empirical and try to represent, for each verified parameter, reasonable estimations of what a *correct*, *useful* or *useless*

forecast for the general public is. Also, we have made the choice of setting the same thresholds for all regions in Switzerland as this allows easier explanation and comparison of the forecast accuracy from one region to another.

The weights ρ_i in equation (2) represent the relative importance of each verified quantity in the global score and can be adjusted according to the verification context. If we perform a verification of parameters representing sensible weather, then the list of verified parameters would most likely enclose precipitation, cloudiness or sunshine duration, temperature and wind, as those features are the most widely communicated to the general public. As we shall see in Section 4, we give a similar weight to all these parameters except for wind, for reasons that we will explain below. Of course, the previous list can vary between countries since the impact of some weather feature, e.g. relative humidity, might be different from one region of the Globe to another. As in Switzerland relative humidity is not a crucial characteristics of sensible weather for most of people (actually, it is even not predicted by the bench forecasters), we do not consider it in our verification framework.

In the perspective of comparing results between different countries, the possibility of considering separately partial scores for commonly verified parameters allows some flexibility in defining the global score; each country might include additional parameters and set weights in definition (2) according to its own climatological and administrative specificities. Obviously, if for a given parameter different tolerance and utility thresholds are used, then direct comparison is trickier. If common thresholds are set, then one should keep in mind that comparison is made between the absolute accuracies of the forecasts rather than between their skills.

3 The COMFORT Score

We apply now the principles of Section 2 to define the global score COMFORT. The verified quantities cover the main features of sensible weather in Switzerland and are edited by MeteoSwiss bench forecasters (more details about how forecasters edit forecasts are given in Section 4). As already discussed in Section 2, the choice of the free parameters μ and α in each partial score were based on empirical estimation of what a correct/useless forecast is, and were supported by a series of simulations.

According to definitions (1) and (2), the COMFORT score has positive orientation and is bounded by 0 and 100. Thus, a score equal to 100 means that the forecast is fully correct whereas a score of 0 means that the forecast is useless. The following quantities are verified by the COMFORT score:

1. **precipitation** (denoted by P): daily (i.e., 0h-24h) amount [mm]
2. **relative sunshine** (denoted by RS) relative to the maximum daily sunshine duration in [$\%$]
3. **minimum daily temperature** (denoted by $T_{\min}$) in [°C]
4. **maximum daily temperature** (denoted by $T_{\max}$) in [°C]
5. **wind speed** at 10m above ground level (denoted by V): maximum hourly average between 6am and 6pm in [kt].

For each of the previous quantities, a partial score is defined according to (1) (see Subsections 3.1 to 3.3). The global COMFORT score is then a particular case of (2):

$$COMFORT = \rho_P S_P + \rho_{RS} S_{RS} + \rho_{T_{\min}} S_{T_{\min}} + \rho_{T_{\max}} S_{T_{\max}} + \rho_V S_V \tag{3}$$

where S_P is the partial score for precipitation, S_{RS} is the partial score for relative sunshine, $S_{T_{\min}}$ and $S_{T_{\max}}$ are partial scores for minimum and maximum daily temperatures respectively, and S_V is the partial score for wind speed. The weights which are based on the former verification system at MeteoSwiss (OPKO) were initially inspired by the *Met Office global NWP index* (*MetOffice*, 2010):

$$\rho_P = \rho_{RS} = 0.3; \quad \rho_{T_{\min}} = \rho_{T_{\max}} = 0.15; \quad \rho_V = 0.1. \tag{4}$$

Equal weights are thus given to precipitation, relative sunshine and temperature, whereas wind speed has only little influence on the global score. The main reason for setting such a smaller weight for wind is the difficulty of having representative observations especially in mountainous regions which prevail in the country. We thus have made the choice of verifying wind speed only at selected stations catching

out the dominating winds blowing in Switzerland. For countries with larger flatlands or coastlines, where measures might be more representative of the regional weather conditions, more importance shall be given to this parameter.

In the following subsections, we define the partial scores for all verified quantities.

3.1 Partial Score for Precipitation

For precipitation, we assume that an error of a given magnitude has a smaller impact on the quality of the forecast when the amount of rainfall is large than when it is small or equal to zero. Several variants have been tested among which the following one was retained, inspired from the *Root mean squared fraction* score (*Golding*, 1998) with the advantage of being well-defined for zero values. For any pair forecast-observation (f, o), we define

$$S_P(f,o) = \begin{cases} 100 & \text{if } |f-o| \leq \mu(f), \\ 100 \cdot \left(1 - \frac{o^p - (f+\mu(f))^p}{d}\right) & \text{if } 0 < o^p - (f+\mu(f))^p < d, \\ 100 \cdot \left(1 - \frac{(f-\mu(f))^p - o^p}{d}\right) & \text{if } 0 < (f-\mu(f))^p - o^p < d, \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $p > 0$. Formula (5) is thus a particular case of (1) with $ERR(f,o) = |f^p - o^p|$. The tolerance threshold μ depends linearly on the forecast as $\mu(f) = 0.3f + 0.1$, so that the score granted to the pair (f, o) is maximum whenever the observation o falls inside a neighborhood of 30% of the magnitude of the forecast f. Figure 1 shows the behavior of the partial score $S_P(f, o)$ with respect to the observation o, for different values of the forecast f. The last two rows of Table 1 indicate the use interval around forecast f for different values of f.

As we can see from Figure 1, the choice of a non-linear distance in (5) implies that, given a forecast $f > 0$ and a number $0 \leq \Delta < f$, $S_P(f, f-\Delta) \leq S_P(f, f+\Delta)$. This property might have an underforecasting influence; suppose indeed that a forecaster (let us suppose "he") expects a precipitation amount between 10 and 30 [mm], without having a strong feeling about the forecast distribution inside this interval. If he must provide a single value representing his best-judgement, he will be likely to choose the mid-point value of the previous interval, that is, 20 [mm]. However, being aware of the previous property of the score, he might be tempted to hedge his forecast in order to reduce the possible penalty, as shown in Table 2. However, letting the tolerance threshold μ depend on the forecast (rather than on the observation) mostly compensates this phenomenon. In particular, this should convince forecasters to edit, whenever justified, amounts delivering a concrete signal (≥ 1 [mm]) rather than precipitation traces (0.2 or 0.5 [mm]) often used to edit a "secure" mean forecast.

The utility threshold α is implicitly defined by the parameters p and d; it depends on f as well as on the sign of the error $|f - o|$. Different values for p and d have been tested and we have retained $p = 2/5$ and $d = 3/2$. The partial score S_P is very strict for small quantities. In particular, wrongly forecasting rain when the weather remains dry, or conversely, is severely penalized. Considering the metric $ERR(f,o) = |f^p - o^p|$ for $0 < p < 1$ in our context has similar implications as considering the error between f and o in *probability space* (see e.g. (*Jolliffe and Stephenson*, 2012), Chapter 5) for which wrongly forecasted frequent events (dry or light rain) are penalized more severely than sparse

and extreme events (heavy rainfall).

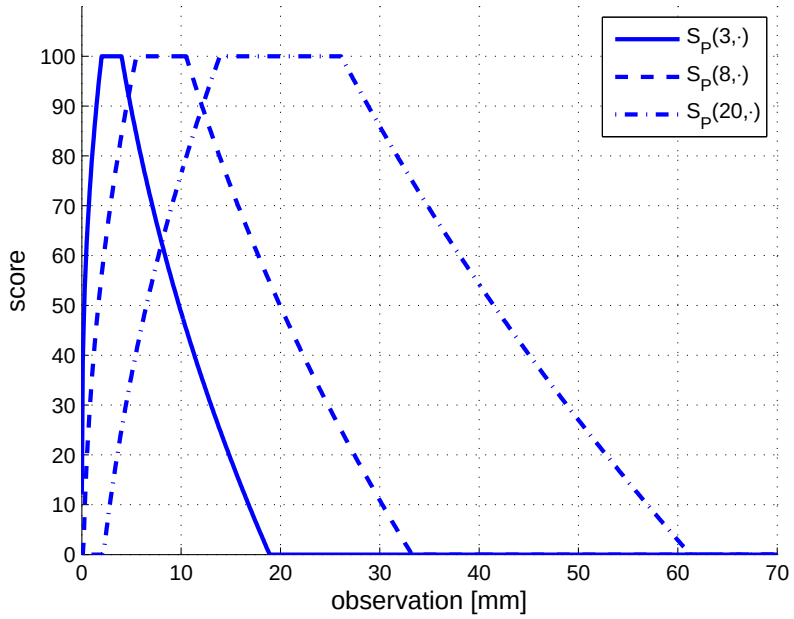

Figure 1: Behaviour of the partial score $S_P(f, o)$ with respect to the observation o, for three different values of the forecast: $f = 3$, $f = 8$ and $f = 20$. The choice of the parameters is $p = 2/5$ and $d = 3/2$.

Table 1: For different values of forecast f are shown the intervals $[a_s, b_s]$ delimiting observations which yield a partial score S_P of at least s, with $s = 0, 50, 75, 100$. The choice of the parameters is $p = 2/5$ and $d = 3/2$.

score	$f\ [mm]$	0	0.2	0.5	1	5	10	15	20	30	50
$= 100$	a_{100}	-	0.1	0.3	0.6	3.5	7	11	14	21	35
	b_{100}	0.1	0.3	0.7	1.4	6.5	13	19	26	39	65
≥ 75	a_{75}	-	-	< 0.1	0.2	2	4.5	7	10	16	28
	b_{75}	0.5	1	1.5	3	10	18	25	33	48	77
≥ 50	a_{50}	-	-	-	0.1	0.8	2.5	4.5	7	12	22
	b_{50}	1.4	2.2	3.4	5	14	23	32	41	58	90
> 0	a_0	-	-	-	-	0	0.4	1	2	5	11
	b_0	5	7	9	11	25	38	50	61	82	121

Table 2: Three variants of the partial score (5) are considered: 1) the retained one; 2) exchanging the roles of f and o in (5); 3) letting μ around f depend on o instead of f. Suppose that the forecaster's best judgement is: a) the mid-point between 10 and 30 [mm]; b) the mid-point between 0.2 and 2 [mm]. The table shows the scores when the observation falls on the bounds of the previous intervals, as well as the approximate correction that the forecaster shall bring to his forecast (column *hedging*), in order to minimize the potential penalty. The choice of the parameters is $p = 2/5$ and $d = 3/2$.

	a)			b)		
(f, o)	$(20, 10)$	$(20, 30)$	hedging	$(1, 0.2)$	$(1, 2)$	hedging
1) $\mu(f) = 0.3f + 0.1$	76	86	-1	81	88	-0.1
2) $f \leftrightarrow o$ in (5)	66	96	-3	78	93	-0.3
3) $\mu(o) = 0.3o + 0.1$	61	96	-4	73	94	-0.4

3.2 Partial Score for Relative Sunshine

The forecast for relative daily sunshine duration RS is edited by forecasters using sunshine categories according to the partition shown in Table 3:

Table 3: Partition of relative sunshine into forecast categories.

RS [%]	$0 \leq RS < 5$	$5 \leq RS < 20$	$20 \leq RS < 50$	$50 \leq RS < 80$	$80 \leq RS \leq 100$
okta	●	◐	◑	◔	○
qualification	“cloudy”	“mostly cloudy”	“partly sunny”	“mostly sunny”	“sunny”

Observations are continuous values bounded by 0 and 100. The classical approach would be to partition the observations into the same categories than forecasts before comparing them, using the multi-categorical verification framework. Instead of this, we shall avoid reducing observations into categories and use formula (1) with a tolerance threshold corresponding to the forecasted category: the score takes it maximum value whenever the observation falls into the forecasted category and the score decreases continuously (and linearly) from the bounds of the forecasted category (see Figure 2). More precisely, denoting by $[a(f), b(f)[$ the category of the forecast f (for instance, $[a(f), b(f)[= [20, 50[$), one defines

$$S_{RS}(f, o) = \begin{cases} 100 & \text{if } a(f) \leq o \leq b(f), \\ 100 \cdot \left(1 - \frac{o - b(f)}{d}\right) & \text{if } 0 < o - b(f) < d, \\ 100 \cdot \left(1 - \frac{a(f) - o}{d}\right) & \text{if } 0 < a(f) - o < d, \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

which is a particular case of (1) with $ERR(f, o) = |f - o|$. As for precipitation, the tolerance threshold

depends on the forecast since the widths of the sunshine categories are not constant (see Table 3).

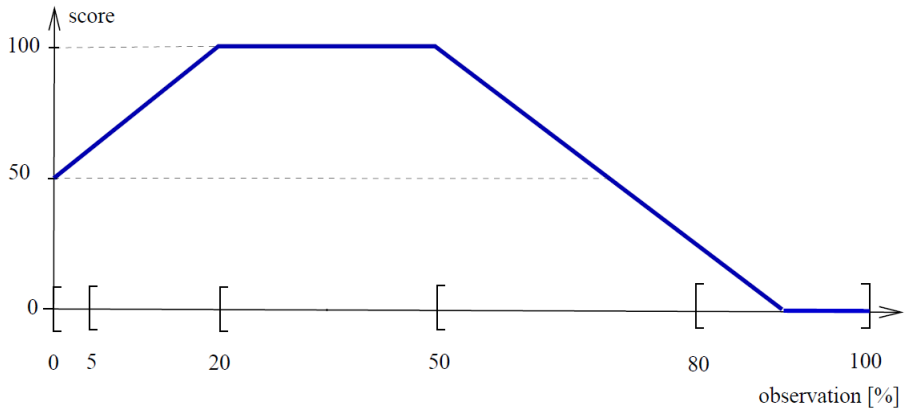

Figure 2: Behaviour of the partial score $S_{RS}(f,o)$ with respect to the observation o when the category $[20,50[$ is forecasted.

The free parameter d defines the use threshold: $\alpha = \mu + d$; if the observation falls further away than $d\%$ from the upper/lower bound of the forecasted sunshine category, then $S_{RS}(f,o) = 0$ (see Figure 2). For our verification purposes, we have retained the value $d = 40$.

3.3 Partial Scores for Minimum/Maximum Temperature and Wind Speed

For daily minimum and maximum temperatures as well as for wind speed, the utility and the tolerance thresholds are independent of the magnitude of the verified quantity. The partial scores are both obtained from formula (1) by setting $ERR(f,o) = |f-o|$. For any pair forecast-observation (f,o), this gives

$$S_{Param}(f,o) = \begin{cases} 100 & \text{if } |f-o| \leq \mu, \\ 100 \cdot \left(1 - \frac{|f-o|-\mu}{\alpha-\mu}\right) & \text{if } \mu < |f-o| \leq \alpha, \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where $Param \in \{T_{\min}, T_{\max}, V\}$. For our verification purposes, we have fixed the following thresholds:

- maximum temperature: $\mu = 1$ [°C] and $\alpha = 6$ [°C]
- minimum temperature: $\mu = 0.5$ [°C] and $\alpha = 6$ [°C]
- wind speed: $\mu = 2.5$ [kt] and $\alpha = 5$ [kt].

4 Deploying COMFORT at MeteoSwiss

This section contains details about the application of the COMFORT score in the context of the MeteoSwiss forecast operational framework. The verified forecasts are provided by forecasters with an editing tool, called the *Methods Editor*, and are assumed to represent forecaster's best judgment. Forecasts are edited out to time-ranges running from D0 (current day) to D7 (i.e., one week out). All time-ranges except for D0 are verified by COMFORT separately but in a uniform way, that is, using the same scores, the same spatial and temporal resolutions etc. This approach allows one to compare forecast accuracy for successive time-ranges. The reason for neglecting D0 is that some of the forecasts may have been edited when the corresponding observation was already known.
Every partial score is computed on a daily basis: any daily forecast f and the corresponding observation o provide a daily score $S_k(f, o)$ where $k = P$, RS, $T_{\min}$, $T_{\max}$, V. Except for temperatures, short-term forecasts (D1-D2) are edited by forecasters at higher temporal resolution than one day: rainfall amounts are edited in 6 hours intervals, sunshine and wind speed are edited for morning and afternoon. For the verification of those forecasts, the relevant daily quantity f (precipitation sum, average sunshine or maximum wind speed) is derived before computing the daily score. By convention, the forecast of day d is the forecast whose *validity is* day d. The global COMFORT score for a given time interval (e.g. a month, a season or a year) is given by formula (3) where the S_k, $k = P$, RS, $T_{\min}$, $T_{\max}$, V are the arithmetic means of the daily partial scores computed over that period.

The spatial resolution of a forecast edited in the *Methods Editor* depends on the forecast's time-range. The Swiss territory is partitioned into 27 **regions** for *short-range* forecasts (time-ranges D1 and D2), into 11 regions for *middle-range* forecasts (time-ranges from D3 to D5) and into 6 regions for *long-range* forecasts (time-ranges D6 and D7), as shown in Figure 3 and 4. Each region is assigned a **reference station** (indicated by red dots on the maps), as well as a number of **observation stations** (indicated by black dots on the maps), each reference station being an observation station itself. The exact list of observation stations used for the verification depends on the verified quantity (see the Appendix).

The choices for the temporal and especially the spatial granularities retained for the verification were of course influenced by the current forecast editing methodology. In the probable case of an evolution of the editing tool in the future, e.g., to provide forecasts for 27 (or more) regions at all time-ranges with an increased temporal resolution, we shall still keep for the verification the granularities that we are defining now. This will ensure a coherence and a continuity in the verification process, allowing comparison between the scores through the next years.

The verified quantities are of two types: temperature and wind speed are defined in the *Methods Editor* as *local* quantities, which means that the predicted values attributed by forecasters hold for the reference stations only. In contrast, precipitation and relative sunshine are defined as *regional* quantities, which means that the predicted values represent spatial averages over the forecast region.

Technically, all predicted quantities are edited at the reference station of each forecast region. For instance, the relative sunshine edited at the reference station *La Chaux-de-Fonds* (CDF) represents the spatial average of the sunshine forecasted for the corresponding forecast region *Jura* (WS2). There is however an exception when inversion phenomena are forecasted: a distinction is then made between sunshine forecasted below and above the inversion (see Subsection 4.2)

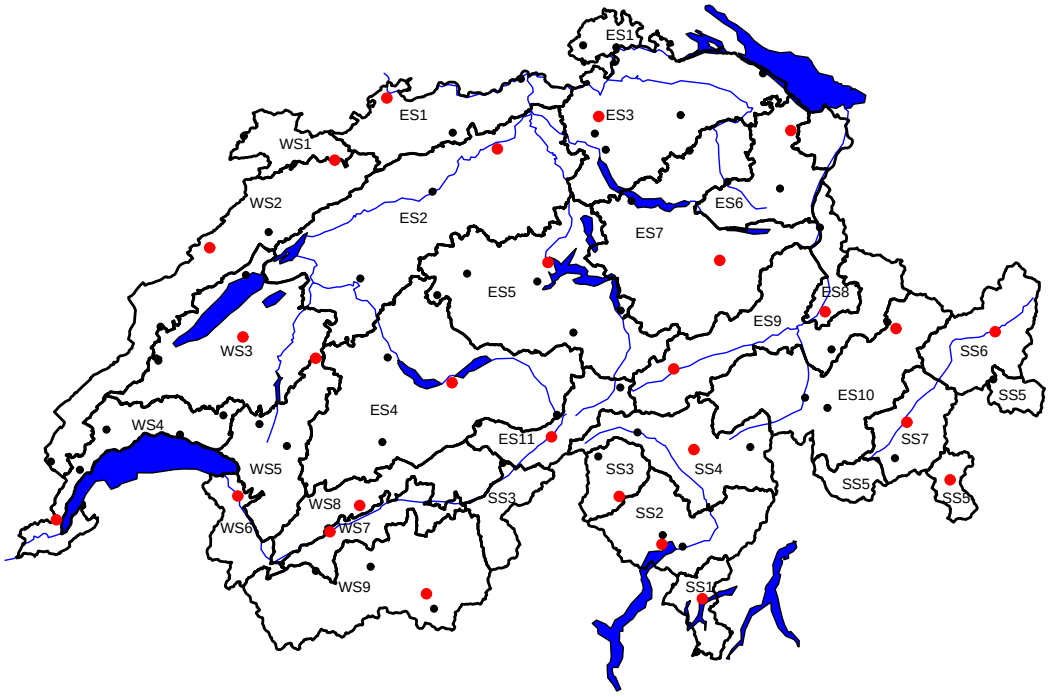

Figure 3: The 27 forecast regions of the *Methods Editor* used for short-range forecasts. These regions are used for the forecast verification at all time-ranges.

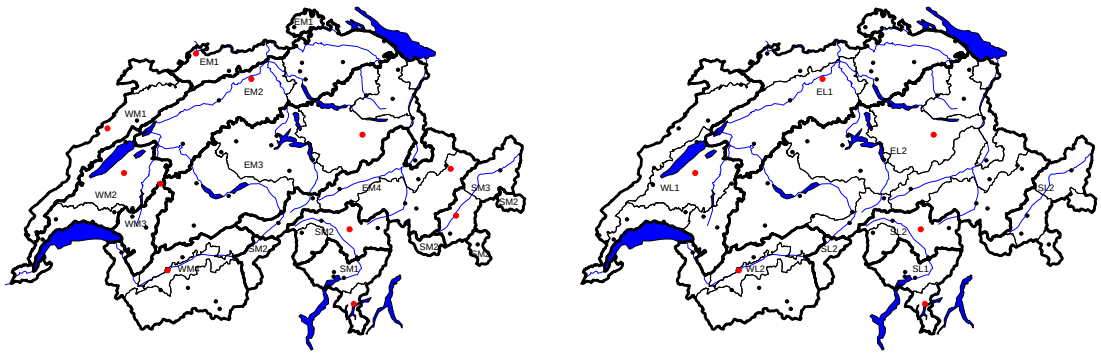

Figure 4: The 11 forecast regions used for middle-range forecasts (left) and the 6 forecast regions used for long-range forecasts (right).

If 15 [mm] of rain are edited for a given day at the reference station *Geneva* (GVE), then this forecast represents the average amount of precipitation over the whole region *Bassin Lémanique* (WS4). Consequently, local quantities are verified using observations from the reference station only, whereas regional quantities are verified using averaged observations over the corresponding region. Thus,

when writing/talking about the COMFORT score for a given region (say the *Bassin Lémanique*), then one should remember that the verification of local quantities concerns the reference station of that region, not the entire region.

As already mentioned, we have made the choice of keeping the same spatial resolution in the verification process for all (short, middle and long) time-ranges. For all quantities except wind speed, this resolution corresponds to the 27 short-range forecast regions/stations depicted in Figure 3. For regional quantities, middle-range and long-range forecasts edited for larger regions naturally induce forecasts for the sub-regions belonging to the larger region (see Figure 4). This allows one to verify forecasts for each of the 27 regions. However, we cannot use the same method for local quantities. In the following subsections, we will discuss the verification procedure for each quantity separately.

4.1 Precipitation

Since precipitation is a regional quantity, a forecast made for a whole region is compared to an average observation for that region. The simplest way of getting an average observation for a given region would be to consider the arithmetic mean of the daily precipitation amounts collected at the raingauges of observation stations belonging to the SwissMetNet (SMN) network. For availability reasons, this source of measurements has been used for the majority of the simulations performed during the development of COMFORT; for each verification region, the corresponding observation stations are listed in Table 13 in the Appendix. An alternative and much more elaborate way is to use rainfall observation data produced by a recent tool developed at MeteoSwiss, which has become operational since October 2013. This tool, called *CombiPrecip* (*Sideris et al.*, 2011) is a combination of a continuous field of precipitation provided by radar images and of sparser measurements provided by the SMN raingauge network (see Figure 5). Geostatistical techniques such as kriging with external drift are generalized and used in order to perform a smart calibration of the radar estimates. This tool provides precipitation estimates at a very high spatial and temporal resolution. *CombiPrecip* was chosen as the primary observation source for the verification, as it provides much more representative regional rainfall estimations than those solely obtained from the raingauge network.

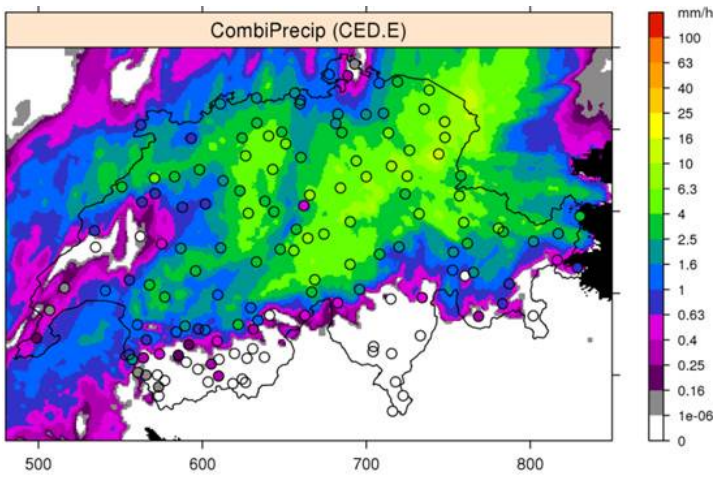

Figure 5: Example of an observation grid generated by *CombiPrecip*: a combination of radar images and measurements from the automatic rain gauge network (circles) provides regional observations used for the verification of precipitation forecasts.

4.2 Sunshine

Similarly to rainfall, sunshine is a regional quantity. A forecast edited for a whole region is thus compared with an average observation corresponding to that region which, in this case, is obtained by taking the arithmetic mean of daily relative sunshine measurements obtained at the SMN network stations (these stations are listed in Table 14 of the Appendix.

During cold season stable anticyclonic regimes north from the Alps, the presence of stratus clouds that either dissipate during the day or persist for several days can significantly impact the daily relative sunshine over large parts of the Swiss Plateau. Conversely, locations above the inversion typically experience maximum sunshine. Consequently, the daily sunshine amount in a given forecast region can significantly differ depending on the elevation of the site. Therefore, when editing the sunshine forecast for a region R, the forecaster has the possibility to include the occurrence of stratus; if he/she choose to do so, he/she can then edit the height of the stratus top as well as the sunshine above the inversion. Technically, the latter corresponds to editing the sunshine at the mountain station *Le Moléson* (MLS, located at 1974 m) if R belongs to the Western Forecasting Area, at *Saentis* (SAE, located at 2505 m) if R belongs to the Eastern Forecasting Area or at *Cimetta* (CIM, located at 1661 m) if R belongs to the Southern Forecasting Area.

The sunshine verification in region R is performed with respect to the following derived forecast: if stratus is predicted by the forecaster, then each observation station located below the inversion is attributed the sunshine amount edited at the reference station of region R[2]. Each observation station located above the inversion is attributed the sunshine amount edited by the forecaster at one of the three previously mentioned mountain stations MLS, SAE or CIM according to the geographical location of the region R. The average forecast for region R is then obtained by taking the arithmetic mean of the forecasts previously derived at each observation station of R. As already mentioned, for time-ranges D1 and D2, distinct forecasts are edited for the morning and for the afternoon; it is then the mean of the derived forecasts for the morning and for the afternoon which is compared with the daily observation.

To illustrate this, let us consider the *ZentralSchweiz* (ES5) forecast region (see Table 14); let us suppose that the forecaster predicts a persistent stratus with a top at around 1200m with a maximum sunshine above. As a result, the stations NAP (1404 m) and PIL (2106 m), both located above the inversion, receive the forecast validated for the mountain station SAE, which would be “sunny” in this case, whereas the stations LUZ (454 m), ALT (438 m), ENG (1036 m) and LAG (745 m) receive the forecast validated at the reference station LUZ, namely “cloudy”.

4.3 Temperatures

Since temperatures forecasts are defined in the *Methods Editor* as local quantities, the maximum and the minimum daily temperatures forecasts at a reference station are compared with the corresponding measurements obtained at the same station. The Table 15 in the Appendix lists the stations used for the verification. For middle-range (respectively long-range) temperature forecast verification, we

[2] If an inversion is forecasted then, by definition, the value edited in the *Methods Editor* at the reference station represents the spatial average of the predicted relative sunshine over those areas of the forecast region which are located below the inversion.

consider an interpolation extending the forecasts validated by forecasters for the 11 (respectively the 6) stations to the 27 reference stations of the short-range partition. Since this interpolation is also in operational use for production, it makes sense to include it in the verification process, with the consequence of englobing the part of forecast errors which are due to the interpolation. Currently, this interpolation uses techniques developed in (*Frei*, 2013). As discussed there, errors (mean absolute error MAE) introduced by the interpolation technique typically range from 0.5 [°C] to 1.5 [°C] depending on the orography of the area and on the season, with larger errors (MAE ≥ 3) expected in un-sampled valleys. These values should however be considered carefully in our context as there are differences between the operational algorithm and the one which yielded the cited values; for instance, the set of points used by the operational algorithm when interpolating the temperature forecasts slightly differs from that one considered in (*Frei*, 2013).

4.4 Wind speed

Wind speed forecasts are edited by MeteoSwiss forecasters as the maximum hourly averages (in [kt]) between 6am and 6pm expected to occur at the reference stations. The verification of wind speed differs a bit from the other quantities, mainly due to the lack of an efficient method for interpolating middle-range and long-range wind forecasts to all 27 reference stations. Consequently, in order to keep the spatial resolution of the verification constant through all time-ranges, the verification of wind speed is performed for 6 stations catching out the dominating winds blowing in Switzerland. These are the reference stations used for long-range forecasts, see Table 12 in Appendix 6. When computing the COMFORT score for each of the 27 forecast regions, we assign a given region R the partial score obtained at the reference station of the long-range region containing R. For instance, for the *Bassin Lémanique* (WS4) forecast region, we assign the partial score for wind speed obtained at the *Payerne* reference station (PAY).

5 Selected results

This section contains a selection of the results obtained by testing the score COMFORT on real data. In Subsection 5.1, examples of how COMFORT can be concretely used to analyse and get an insight into the verified forecasts are provided. In Subsection 5.2, we perform simulations aiming to quantify the improvement potential of the score subject to different theoretical enhancements of the forecasts. As explained there, these simulations were largely motivated by administrative purposes. In Subsection 5.3, we test the robustness of the score against hedging. Finally, in the last subsection, COMFORT's values are compared with those provided by another administrative score.

5.1 From rough values to a finer analysis

A convenient feature of COMFORT is that the global score obtained for a given period can be easily decomposed parameter by parameter making easier the interpretation of its values. Starting typically from a global representation such as on Figure 6, which shows the evolution of annual scores averaged over the entire of Switzerland, one can progressively increase the spatial and/or the temporal resolution (see Figures 7 and 8) in order to perform analysis at different levels of detail.

For instance, different observations can already be drawn out from Figure 8, such as seasonality in the forecast accuracy for some parameters; this is striking for wind but seems also to appear for precipitation or minimum temperature. As we can see from the plotted time-series, the score for wind speed lies significantly below the other scores. This reflects the difficulty of accurately predicting this parameter exhibiting high spatial and temporal variability, yet intensified by the complex orography of Switzerland. However, one also notices a clear improvement in accuracy since July 2013; this coincides with the operationalization of a new model output statistics which most of forecasters are now working with. The Table 4 contains annual (partial and global) scores, averaged over Switzerland, for different time-ranges.

As an example, if one wishes to analyse in greater detail the scores for a given region, let us say *Engadina bassa* which has obtained the poorer score for precipitation in 2014 (73 points), and for a given period, one can conveniently increase the resolution up to daily scores, as shown on Figure 10. Daily scores can then be put into relation with the meteorological context, for further analysis. For instance, during January and February 2014 which obtained scores below the average, the occurences of humid days as well as the predicted amounts were both overestimated (all humid days were detected, but the false alarm rate was equal to 45%).

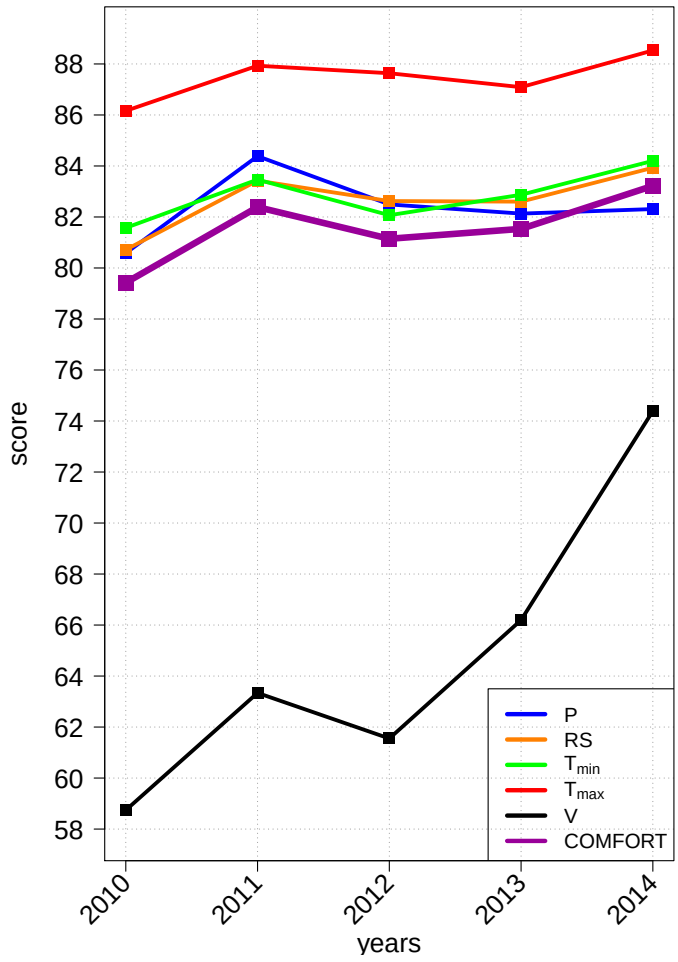

Figure 6: Yearly evolution of the COMFORT score and the partial scores composing it, averaged over Switzerland. The forecast time-range is D1.

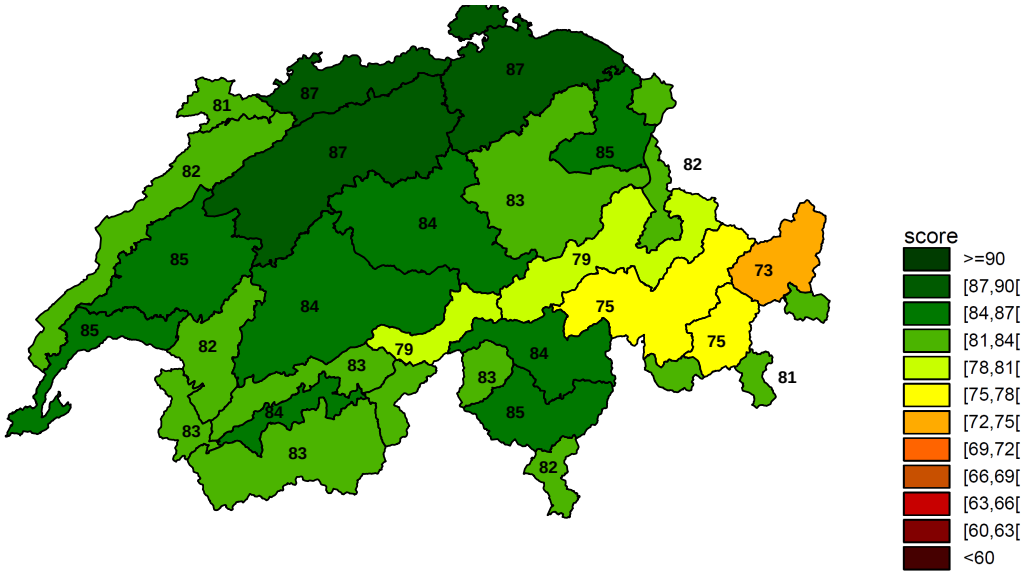

Figure 7: Annual partial scores for precipitation obtained by each forecast region for time-range D1, year 2014.

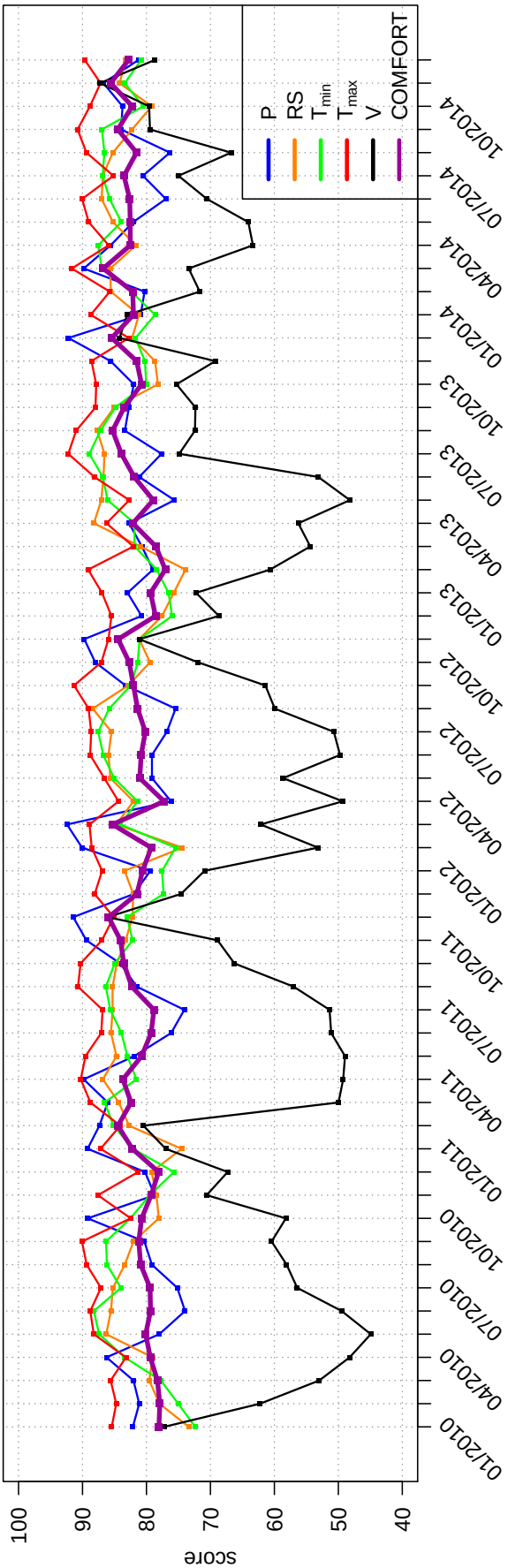

Figure 8: Monthly evolution of the COMFORT score and the partial scores composing it, averaged over Switzerland. The forecast time-range is D1.

Table 4: Annual partial scores and COMFORT score, averaged over Switzerland, for the period 2010-2014. The forecast time-ranges are D1, D3 and D5.

	D1					D3					D5				
	2010	2011	2012	2013	2014	2010	2011	2012	2013	2014	2010	2011	2012	2013	2014
S_P	80.6	84.4	82.5	82.1	82.3	75.8	79.8	75.9	76.4	74.3	68.7	73.3	69.1	71.1	67.4
S_{RS}	80.7	83.4	82.6	82.6	83.9	75.2	77.4	75.5	75.0	76.5	68.1	69.9	68.6	69.7	70.6
$S_{T_{\min}}$	81.6	83.5	82.1	82.9	84.2	76.6	77.3	76.7	77.6	81.7	71.5	71.2	68.5	70.7	77.5
$S_{T_{\max}}$	86.2	87.9	87.6	87.1	88.5	79.8	80.5	79.7	79.6	83.8	70.1	70.2	70.2	71.0	75.2
S_V	58.7	63.3	61.6	66.2	74.4	53.9	61.9	58.9	62.2	72.3	52.0	58.4	55.7	60.8	69.5
COMFORT	79.4	82.4	81.2	81.5	83.2	74.1	77.0	74.8	75.2	77.3	67.5	70.0	67.7	69.6	71.2

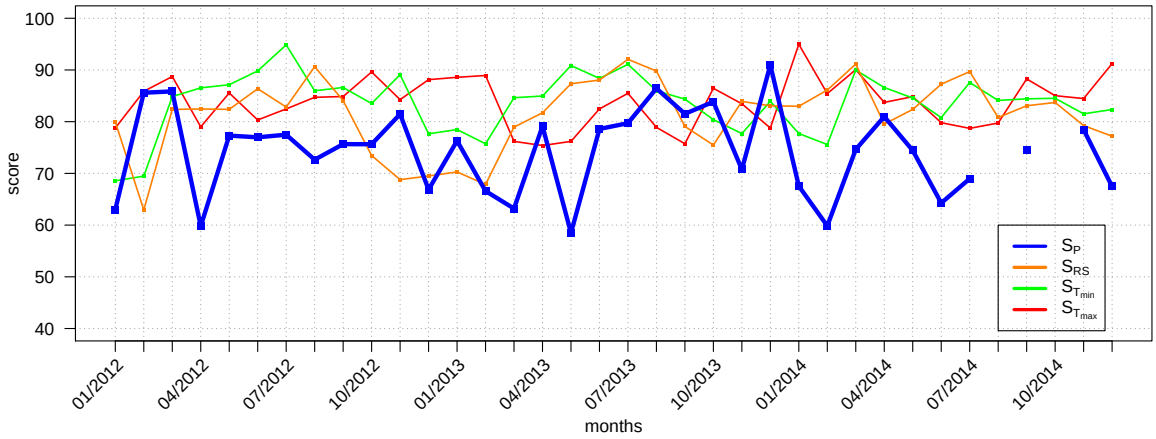

Figure 9: Monthly evolution of partial scores for precipitation, sunshine and min/max temperatures for the forecast region *Engadina bassa* (reference station *Scuol*). The time-range of the shown forecasts is D1. (The discontinuities in the blue line are because of too many missing observations during months of August an October 2014.)

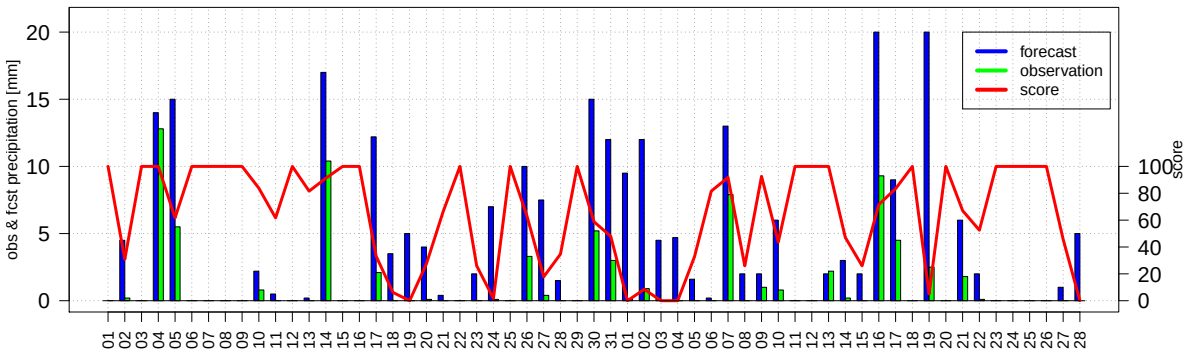

Figure 10: Daily evolution of the partial score for precipitation during January and February 2014, for the forecast region *Engadina bassa*; forecast time-range D1.

5.2 COMFORT's improvement potential

During the development of COMFORT, simulations were made with the aim to estimate COMFORT's sensitivity to different theoretical forecast enhancements. These simulations were largely motivated by administrative purposes. Both simulations were performed using forecasts validated by forecasters (denoted by VAL) and observation data from 2010 to 2012.

5.2.1 Simulation 1

The aim of this simulation was to quantify in some way the forecast enhancement needed to reach a COMFORT score of 85 points for the time-range D1. This was particularly to answer a question asked by the MeteoSwiss leadership when fixing quantitative medium and long-term objectives for the score. For each verification region, the forecasts VAL were modified as follows:

- precipitation: reduction of 50% of all forecast errors associated with moderately wet to very wet days (i.e., more than 10 [mm] of daily amount)
- sunshine: improvement of one sunshine class of all forecasts with an error larger than one class
- minimum/maximum temperatures: reduction of one [$^\circ$ C] of all errors larger than 2 [$^\circ$ C]
- wind speed: improvement of one editable[3] value of all forecasts with errors larger than 5 [kt].

The resulting enhanced forecast is denoted by COR. Table 5 contains the partial scores obtained by the forecast COR for each parameter, as well as the COMFORT score. Obviously, this simulation is quite demanding as it requires a systematic reduction of errors of a given type (greater than some magnitude or associated with some weather type), for each parameter. Thus, the values obtained via this simulation should rather constitute a long-term objective.

[3] Technically, the forecast for wind speed is edited in knots [kt] according to the following admitted values: 0, 3, 5, 7, 10, 12, 15, and then $15 + 5k$ for any positive integer k.

Table 5: Partial scores and COMFORT score for the validated forecast VAL and the corrected forecast COR, averaged over the period 2010-2012 and over Switzerland. The forecasts ranges are D1, D3 and D5.

	D1		D3		D5	
	VAL	COR	VAL	COR	VAL	COR
S_P	82.6	84.4	77.3	79.8	70.5	74.2
S_{RS}	82.5	87.0	76.2	83.8	69.0	79.8
$S_{T_{\min}}$	82.4	86.3	76.8	82.5	70.4	77.5
$S_{T_{\max}}$	87.2	91.8	80.0	86.9	70.2	78.9
S_V	61.0	68.9	58.5	66.6	55.5	63.8
COMFORT	81.1	85.1	75.4	81.2	68.5	76.0

5.2.2 Simulation 2

The second simulation aimed to show that the COMFORT score was able to capture and reward forecast adjustments based on new incoming weather information (new NWP output, new or additional observations, etc.). This should encourage forecasters to issue their forecasts according to their best available judgement. Thus, alternatively to a systematic reduction of errors, we have also simulated the enhancement of randomly selected forecasts. The corrections which are inserted below to the forecasts VAL are supposed to reflect adjustments that can be brought to forecasts by forecasters whose best judgement is modified. In order to stay as realistic as possible, we suppose that modifications brought by the forecaster can also sometimes impair the forecast.

The first part below simulates an adjustment of the forecasts while taking minimal risk: a forecast is adjusted only if the forecaster is confident about the change. There are less adjusted forecasts, but none of them is impaired. As we can see from Table 7, slightly improving only around 10% of the forecasts should be reflected in a significant way by the values of the score. In the second part below, one simulates an adjustment of the forecasts while taking additional risk: more forecasts are modified, some of which are negatively impacted. We see that on average, taking moderate risk is not unduly penalized by the COMFORT score (see Table 8). Indeed, the proportion of impaired forecasts ($\simeq 5\%$) is kept still largely smaller than the number of improved forecasts ($\simeq 20\%$).

- precipitation: for each verification region, we select uniformly at random a fixed number of days. For n_1 days, depending on whether the observation was higher/lower than the forecast, we increase/reduce the forecast value according to the scale defined in Table 6; for $n_1/2$ days, a double correction is applied. For n_2 days, the forecast is impaired according to the same rule (for $n_2/2$ days, a double deterioration is applied).

Table 6: Correction brought to a precipitation forecast for n_1 days (single correction) and for $n_1/2$ days (double correction).

forecast x [mm]	single correction [mm]	double correction [mm]
$x < 1$	± 0.2	± 0.4
$1 \leq x < 10$	± 1	± 2
$10 \leq x < 20$	± 2	± 4
$20 \leq x < 50$	± 5	± 10
$x \geq 50$	± 10	± 20

- sunshine: for each verification region, the forecasts for n_1 days (respectively n_2 days) selected uniformly at random are improved (respectively impaired) by one sunshine class (see Subsection 3.2)
- minimum/maximum temperatures: for each verification region, the forecasts for n_1 days (respectively n_2 days) whose error exceeds a given magnitude, selected uniformly at random, are improved (respectively impaired) by one [° C]
- wind speed: for each verification region, all forecasts whose errors are larger than 5 [kt] are improved by one editable class (see Subsection 3.3).

In this way, we construct a randomly (except for wind speed) modified forecast. After computing a large enough number of realisations, the mean score is calculated.

1. Table 7 contains the values for the following parameter choices:
   - precipitation: $n_1 = 110$ and $n_2 = 0$ (which corresponds to improve 15% of the forecasts)
   - sunshine: $n_1 = 110$ and $n_2 = 0$ (which corresponds to improve 10% of the forecasts)
   - minimum/maximum temperature: $n_1 = 55$ and $n_2 = 0$, for errors larger than 3 [° C] (which corresponds to improve 5% of the forecasts).

Table 7: Partial scores and COMFORT score for the validated forecast VAL and the corrected forecast RCOR (average of the scores of 50 realisations), for the period 2010-2012 over Switzerland. The forecast ranges are D1, D3 and D5.

	D1		D3		D5	
	VAL	RCOR	VAL	RCOR	VAL	RCOR
S_P	82.6	84.0	77.3	79.0	70.5	72.3
S_{RS}	82.5	85.2	76.2	79.2	69.0	72.2
$S_{T_{\min}}$	82.4	83.0	76.8	77.6	70.4	71.1
$S_{T_{\max}}$	87.2	88.0	80.0	80.8	70.2	70.9
S_V	61.0	68.9	58.5	66.6	55.5	63.8
COMFORT	81.1	83.3	75.4	77.9	68.5	71.0

2. Table 8 contains the values for the following choices of parameters:
   - precipitation: $n_1 = 220$ and $n_2 = 55$
   - sunshine: $n_1 = 220$ and $n_2 = 55$
   - minimum/maximum temperature: $n_1 = 110$ (for errors larger than 3 [° C]) and $n_2 = 55$ (for any error magnitude).

Table 8: Partial scores and COMFORT score for the validated forecast VAL and the corrected forecast RCOR (average of the scores of 50 realisations), for the period 2010-2012 over Switzerland. The forecast ranges are D1, D3 and D5.

	D1		D3		D5	
	VAL	RCOR	VAL	RCOR	VAL	RCOR
S_P	82.6	84.0	77.3	79.2	70.5	72.6
S_{RS}	82.5	85.2	76.2	79.6	69.0	72.8
$S_{T_{\min}}$	82.4	82.5	76.8	77.2	70.4	70.8
$S_{T_{\max}}$	87.2	87.6	80.0	80.6	70.2	70.6
S_V	61.0	68.9	58.5	66.6	55.5	63.8
COMFORT	81.1	83.2	75.4	78.0	68.5	71.2

5.3 Robustness against hedging

In our context, the term “hedging” should be understood as issuing a forecast which does not correspond to the forecaster’s best judgement, in order to obtain a better long-term mean score. We have

compared, for each verified quantity, scores obtained by forecasts edited by forecasters with scores obtained by different "no-skill" or "no-risk" forecasts. All results presented below were obtained by pooling together monthly scores over the period 2010-2012 and over all 27 forecast regions.

Precipitation amounts validated by forecasters, denoted by VAL, were compared to "no-skill" forecasts consisting of constantly forecasting "no rain" (i.e., 0 [mm] for all days) or "minimal rain" (i.e., 0.2 [mm] for all days). The corresponding forecasts are denoted by DRY and ALDRY[4]. Table 9 shows the average differences between monthly precipitation scores obtained by the forecasts VAL and DRY/ALDRY as well as the empirical probability of obtaining a better score when forecasting the scheme instead of the "best judgement", assuming that the forecast VAL always represented forecaster's best judgement. As we can see from Table 9, at least until three days ahead, emitting a "lazy" forecast had little chance to be rewarded better than the "best judgement"; for long-term forecasts, there was about 50% chance of being rewarded better. Figures 11 and 12 show the empirical distributions of the differences in monthly partial scores for precipitation S_P obtained by the forecasts VAL and DRY/ALDRY.

Table 9: Delta: average differences between monthly precipitation scores obtained by the forecasts VAL and DRY/ALDRY. Ratio: empirical probability of obtaining a better score when forecasting the scheme DRY, or ALDRY, instead of the "best judgement".

	Delta		Ratio	
	DRY	ALDRY	DRY	ALDRY
D1	16	18	0.1	0.1
D3	10	12	0.2	0.1
D6	1	3	0.5	0.4

Edited forecasts for relative sunshine were compared with "no-skill" forecasts obtained by constantly forecasting the "climatological" mean sunshine class. The forecast denoted CST consisted in forecasting the class $[20, 50[$ for all regions except those belonging to *Valais* and those from the administrative *South* region (regions denoted WS6 to WS9 and SS1 to SS7 on Figure 3). For these regions, the class $[50, 80[$ was forecast instead. Alternatively, forecasts for relative sunshine were compared with a modified "no-risk" forecast, denoted by NOR, obtained from VAL by avoiding to forecast the extreme sunshine classes $[0.5[$ and $[80, 100]$ (i.e., the class $[5, 20[$ was forecasted instead of $[0, 5[$ and the class $[50, 80[$ was forecasted instead of $[80, 100[$). The aim of this test was to check whether predicting "safe-looking" albeit not the climatological distribution average values for relative sunshine was not unduly favored. Results are summarized in Table 10. As expected, differences between NOR and VAL are much smaller than between CST and VAL, but remain in favour of VAL. There was almost no chance of obtaining a better score just while forecasting the climatological class. Also, the "safe-looking" prediction does not guarantees better scores than the "best judgement". This clearly follows from the fact that the climatological distribution for relative sunshine concentrates around the bounds of its support hence shooting at the middle is often inaccurate. Figures 13 and 14 show the empirical distribution of the differences in monthly partial scores for relative sunshine S_{RS} obtained by the forecasts VAL and CST/NOR.

[4] ALDRY stands for "almost dry".

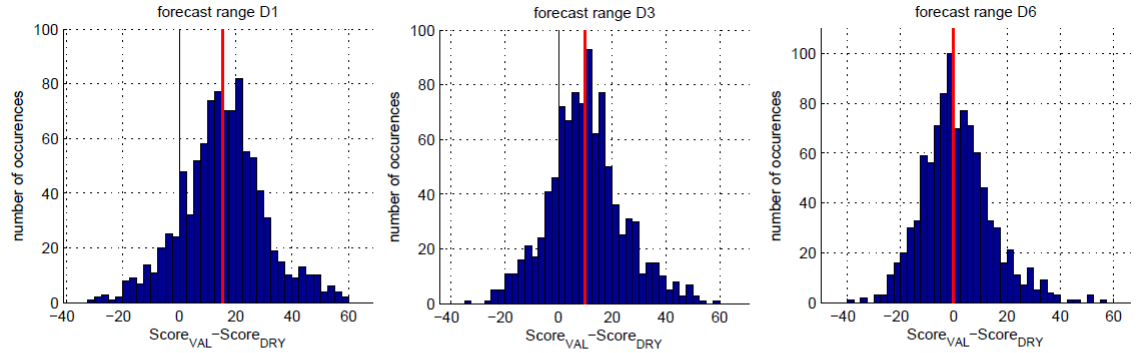

Figure 11: Empirical distribution of the differences in monthly scores S_P obtained by the forecasts VAL and DRY. The sample median is shown in red.

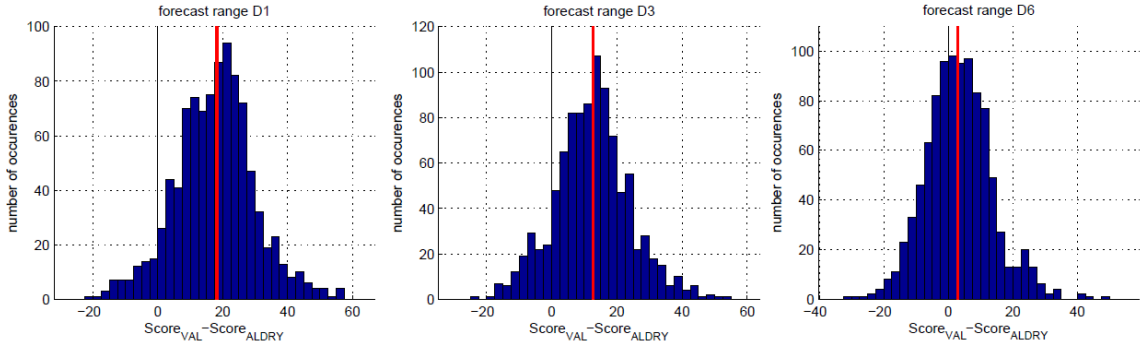

Figure 12: Empirical distribution of the differences in monthly scores S_P obtained by the forecasts VAL and ALDRY. The sample median is shown in red.

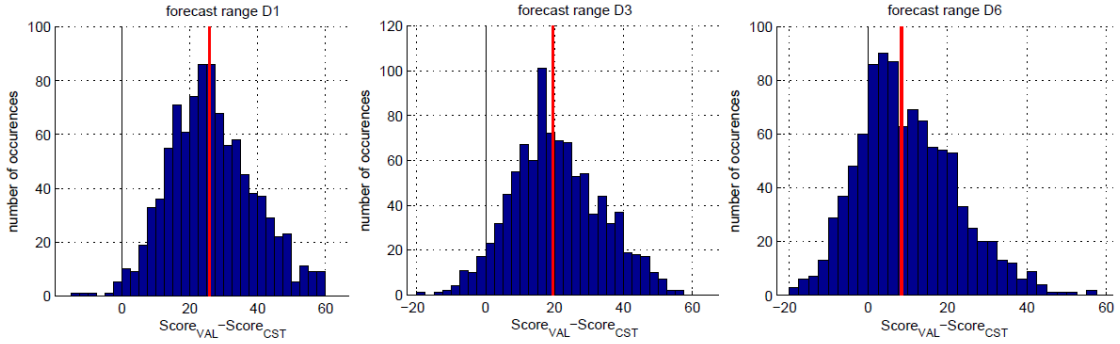

Figure 13: Empirical distribution of the differences in monthly scores S_{RS} obtained by the forecasts VAL and CST. The sample median is shown in red.

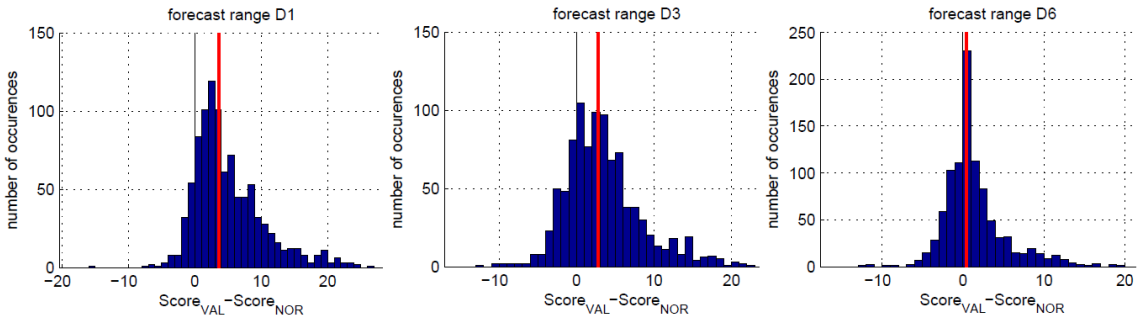

Figure 14: Empirical distribution of the differences in monthly scores S_{RS} obtained by the forecasts VAL and NOR. The sample median is shown in red.

Table 10: Delta: average differences between monthly sunshine scores obtained by the forecasts VAL and CST/NOR. Ratio: empirical probability of obtaining a better score when forecasting the scheme CST, or NOR, instead of the "best judgement".

	Delta		Ratio	
	CST	NOR	CST	NOR
D1	27	5	0.01	0.1
D3	20	3.5	0.05	0.25
D6	10	1.5	0.2	0.3

Edited forecasts VAL for minimum and maximum temperatures were compared with the persistence forecast, denoted by PER. Figure 15 shows the empirical distribution of the differences in monthly partial scores $S_{T_{\min}}$ and $S_{T_{\max}}$ for minimum and maximum temperatures obtained by the forecasts VAL and PER for time-range D1. For minimum temperature, the average difference between the monthly scores obtained by VAL and PER is 8 points in favor of VAL at range D1. The same difference grows to 18 points for maximal temperature. For longer-range forecasts, the previous values increase in favor of VAL as persistence becomes less reliable. When forecasting persistence for short-range forecasts (D1), regarding minimum temperature there was slightly less than a 10% chance to get a better score than if editing it according to the "best judgement", whereas this probability was only about 3% for maximum temperature.

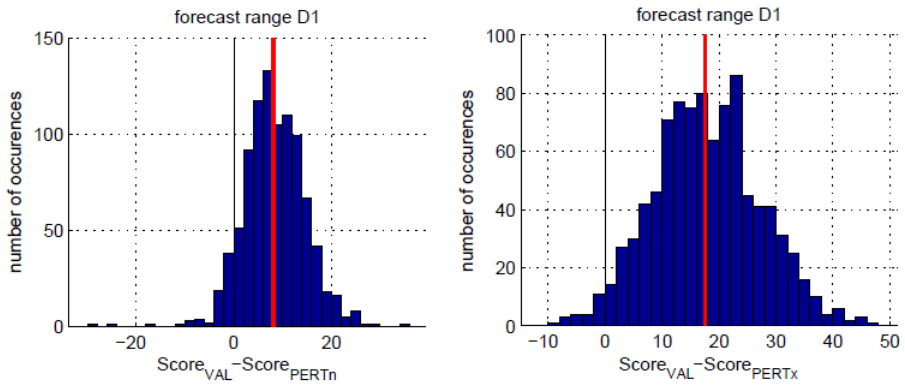

Figure 15: Empirical distribution of the differences in monthly scores $S_{T_{\min}}$ (left) and $S_{T_{\max}}$ (right) obtained by the forecasts VAL and PER for time-range D1. The sample median is shown in red.

5.4 Comparison with another administrative score

Partial scores composing COMFORT were compared with values provided by another administrative score, called the *two-alternative forced choice* (2AFC) score (*Mason and Weigel*, 2009). The 2AFC score is based on a discrimination test applied to all possible sets of two forecast-observation pairs, which assigns the values 0, 0.5 or 1 according to whether the forecasts allow to correctly distinguish the corresponding observations. Results of these elementary tests are then gathered into a single value for each verified quantity. A 2AFC score equal to 100 means perfect discrimination, whereas a score of 50 means the absence of discrimination skill.

Although the 2AFC score measures forecast quality from a different side than COMFORT by focusing on discrimination rather than on accuracy, it is nevertheless interesting to compare both scores' results; not their absolute values, but the value differences from one parameter to another. As it can be read from Table 11, the score differences from one parameter to another have almost all the same signs and similar magnitude when measured by COMFORT than when measured by 2AFC.

This comparison supports that the free parameters μ and α in each partial score forming COMFORT were chosen in a consistent way over all verified quantities, that is, each quantity is treated with a similar level of severity. This is particularly visible for wind: the gap in the mean score values between wind and the other parameters clearly appears in both scores. This gap measured by COMFORT is thus a real sign of poorer forecast quality for wind in comparison with the other parameters, and not an artefact created by setting for wind too severe tolerance and utility thresholds.

Table 11: For each verified parameter, the score 2AFC and the partial score forming COMFORT are shown. The values are averaged over Switzerland and over the period 2010-2012. The forecast time-ranges are D1, D3 and D5.

	D1		D3		D5	
	2AFC	COMFORT	2AFC	COMFORT	2AFC	COMFORT
P	82.3	82.5	78.1	77.2	70.1	70.4
RS	83.0	82.3	78.3	76.0	70.2	68.8
$T_{\min}$	82.9	82.4	79.0	76.9	72.5	70.4
$T_{\max}$	87.0	87.2	82.6	80.0	75.3	70.2
V	63.3	61.2	60.5	58.3	56.4	55.4

6 Appendix: Tables

Table 12: Reference and observation stations used for wind speed verification.

Long-range Region	Region Code	Reference and Observation Station
Jura + Plateau Ouest	WL1	PAY
Valais	WL2	SIO
Plateau Nord	EL1	BUS
Alpes du Nord	EL2	GLA
Tessin Sud	SL1	LUG
Alpes du Sud	SL2	COM

Table 13: Regions used for the verification of precipitation, together with the reference station and observation stations (the latter were used for testing COMFORT during the period 2010-2012, before *CombiPrecip* became operational).

Short-range Region	Region Code	Reference Station	Observation Stations
Pied Nord du Jura	WS1	DEM	DEM, FAH
Jura	WS2	CDF	CDF, CHA, DOL, FRE, BIE
Broye Neuchâtel	WS3	PAY	PAY, NEU, MAH
Bassin Lémanique	WS4	GVE	GVE, CGI, PUY
Préalpes Romandes	WS5	PLF	PLF, MLS, ORO, CHD
Chablais	WS6	AIG	AIG
Valais Plaine	WS7	SIO	SIO
Valais Mont. Nord	WS8	MVE	MVE
Valais Mont. Sud	WS9	ZER	ZER, EVO
Basel-Schaffhausen	ES1	BAS	BAS, RUE, SHA, LEI, HLL
Zentrales Mittelland	ES2	BUS	BUS, BER, WYN, THU
Oestliches Mittelland	ES3	KLO	KLO, GUT, HOE, REH, TAE, SMA
Berner Oberland	ES4	INT	INT, ABO
Zentralschweiz	ES5	LUZ	LUZ, ALT, ENG, NAP, PIL, LAG
Toggenburg-Alpstein	ES6	STG	STG, SAE, EBK
Schwyz-Glarus	ES7	GLA	GLA, WAE
Rheintal	ES8	CHU	CHU, VAD
Nordbünden	ES9	DIS	DIS
Mittelbünden	ES10	DAV	DAV, VAB, WFJ, PMA, AND
Goms	ES11	ULR	ULR, GRH, GUE
Sottoceneri	SS1	LUG	LUG, SBO
Ticino Centrale	SS2	OTL	OTL, CIM, MAG
Maggia-Sempione	SS3	CEV	CEV, ROE
Bleino-Lev.-Mesolcina	SS4	COM	COM, SBE, PIO
Bregaglia Poschiavo	SS5	ROB	ROB
Engadina Bassa	SS6	SCU	SCU, BUF
Engadina Alta	SS7	SAM	SAM, COV

Table 14: Regions used for the verification of relative sunshine, together with the reference station and the observation stations.

Short-range Region	Region Code	Reference Station	Observation Stations
Pied Nord du Jura	WS1	DEM	DEM, FAH
Jura	WS2	CDF	CDF, CHA, DOL, FRE, BIE
Broye Neuchâtel	WS3	PAY	PAY, NEU, MAH
Bassin Lémanique	WS4	GVE	GVE, CGI, PUY
Préalpes Romandes	WS5	PLF	PLF, MLS, ORO, CHD
Chablais	WS6	AIG	AIG
Valais Plaine	WS7	SIO	SIO
Valais Mont. Nord	WS8	MVE	MVE, EGH
Valais Mont. Sud	WS9	ZER	ZER, EVO, ATT, GOR
Basel-Schaffhausen	ES1	BAS	BAS, RUE, SHA, LEI, HLL
Zentrales Mittelland	ES2	BUS	BUS, BER, WYN, THU
Oestliches Mittelland	ES3	KLO	KLO, GUT, HOE, REH, TAE, SMA
Berner Oberland	ES4	INT	INT, ABO, JUN
Zentralschweiz	ES5	LUZ	LUZ, ALT, ENG, NAP, PIL, LAG
Toggenburg-Alpstein	ES6	STG	STG, SAE, EBK
Schwyz-Glarus	ES7	GLA	GLA, WAE
Rheintal	ES8	CHU	CHU, VAD
Nordbünden	ES9	DIS	DIS
Mittelbünden	ES10	DAV	DAV, VAB, WFJ, PMA, AND
Goms	ES11	ULR	ULR, GRH, GUE
Sottoceneri	SS1	LUG	LUG, SBO
Ticino Centrale	SS2	OTL	OTL, CIM, MAG
Maggia-Sempione	SS3	CEV	CEV, ROE
Bleino-Lev.-Mesolcina	SS4	COM	COM, SBE, PIO
Bregaglia Poschiavo	SS5	ROB	ROB
Engadina Bassa	SS6	SCU	SCU, BUF
Engadina Alta	SS7	SAM	SAM, COV

Table 15: Regions used for the verification of minimum and maximum temperatures, together with the reference and observation station.

Short-range Region	Region Code	Reference and Observation Station
Pied Nord du Jura	WS1	DEM
Jura	WS2	CDF
Broye Neuchâtel	WS3	PAY
Bassin Lémanique	WS4	GVE
Préalpes romandes	WS5	PLF
Chablais	WS6	AIG
Valais Plaine	WS7	SIO
Valais mont. Nord	WS8	MVE
Valais mont. Sud	WS9	ZER
Basel-Schaffhausen	ES1	BAS
Zentrales Mittelland	ES2	BUS
Oestliches Mittelland	ES3	KLO
Berner Oberland	ES4	INT
Zentralschweiz	ES5	LUZ
Toggenburg-Alpstein	ES6	STG
Schwyz-Glarus	ES7	GLA
Rheintal	ES8	CHU
Nordbünden	ES9	DIS
Mittelbünden	ES10	DAV
Goms	ES11	ULR
Sottoceneri	SS1	LUG
Ticino centrale	SS2	OTL
Maggia-Sempione	SS3	CEV
Bleino-Lev.-Mesolcina	SS4	COM
Bregaglia Poschiavo	SS5	ROB
Engadina bassa	SS6	SCU
Engadina alta	SS7	SAM

List of Figures

List of Tables

References

Cattani, D., A. Faes, M. G. Gaillard, and M. Matter (2015), Global forecast quality score for administrative purposes, *to appear in Special Issue of Mausam*.

Frei, C. (2013), Interpolation of temperature in a mountainous region using nonlinear profiles and non-euclidean distances, *International Journal of Climatology*, doi:10.1002/joc.3786.

Golding, B. W. (1998), Nimrod: A system for generating automated very short range forecasts, *Meteorological Applications*, *5*, 1–16.

Jolliffe, I. T., and D. B. Stephenson (Eds.) (2012), *Forecast verification: a practitioner's guide in atmospheric science*, 2nd ed., Wiley-Blackwell.

Mason, S. J., and A. P. Weigel (2009), A generic forecast verification framework for administrative purposes, *Monthly Weather Review*, *137*, 331–349.

MetOffice (2010), "Global NWP index documentation", *available online*.

Murphy, A. H. (1993), What is a good forecast? An essay on the nature of goodness in weather forecasting, *Weather and Forecasting*, *8*, 281–293.

Murphy, A. H., and R. L. Winkler (1987), A general framework for forecast verification, *Monthly Weather Review*, *115*, 1330–1338.

Sideris, I. V., M. Gabella, R. Erdin, and U. Germann (2011), Real-time radar-raingauge merging using spatiotemporal co-kriging with external drift in the alpine terrain of switzerland, *Quaterly Journal of the Royal Meteorological Society*, *00*, 1–22.

Stanski, H. R., L. J. Wilson, and W. R. Burrows (1989), Survey of common verification methods in meteorology, *Tech. rep.*, World Meteorological Organization.

Wilks, D. S. (2011), *Statistical methods in the atmospheric sciences*, *International geophysics series*, vol. 100, 3nd ed., Academic Press.

Acknowledgments

The authors wish to acknowledge all collegues from MeteoSwiss that have contributed either materially or conceptually to the development of COMFORT. We thank Pirmin Kaufmann for his numerous interesting remarks during the whole project, as well as for a careful review of the previous version of this work. We thank Lionel Peyraud for his attentive reading of a preliminary version of this text. We acknowledge Pertti Nurmi for having reviewed with a lot of attention the prefinal version of this manuscript. We are particularly grateful to the entire forecasting team from Locarno for their strong interest in this project. Their invitation to Locarno-Monti in June 2013 resulted in valuable discussions and contributed significantly to the development of COMFORT. We would like to thank Ioannis Sideris for introducing us to *CombiPrecip*, which proved to be a valuable tool in the precipitation verification process. We also thank Marc Musa and Markus Abbt for preparing, providing and supporting all necessary forecast data used in the simulations, as well as Joël Fisler and Erich Weber for providing *CombiPrecip* data which were used during the final stage of COMFORT development. Finally, we would like to thank Pierre Eckert for his continuous support as the Director of APW.

MeteoSchweiz
Operation Center 1
CH-8044 Zürich-Flughafen

T +41 58 460 91 11
www.meteoschweiz.ch

MeteoSvizzera
Via ai Monti 146
CH-6605 Locarno Monti

T +41 91 756 23 11
www.meteosvizzera.ch

MétéoSuisse
7bis, av. de la Paix
CH-1211 Genève 2

T +41 22 716 28 28
www.meteosuisse.ch

MétéoSuisse
Chemin de l'Aérologie
CH-1530 Payerne

T +41 26 662 62 11
www.meteosuisse.ch