MAUSAM, **66**, 3 (July 2015), 645-656

551.509.5

Global forecast quality score for administrative purposes

DANIEL CATTANI, ANNA FAES, MARIANNE GIROUD GAILLARD and MICHEL MATTER

MeteoSwiss, Forecast Division, 7bis av. de la Paix, CH-1211 Geneva 2

e mail : michel.matter@meteoswiss.ch

सार – मीटिओस्विस 1985 से प्रादेशिक पूर्वानुमान केंद्रों द्वारा जारी सामान्य मौसम पूर्वानुमान का व्यवस्थित रूप से आकलन करने के लिए वैश्विक स्कोर का उपयोग कर रहा है। यह आकलन निम्नलिखित दो मुख्य कारणों से किया जाता है: इसका उपयोग प्रशासनिक उद्देश्य के लिए किया जाता है क्योंकि मौसम केंद्रों से आम जनता के साथ संचार करने तथा सरकार के साथ उनके पूर्वानुमान की गुणवत्ता में सुधार लाने की अपेक्षा की जाती है। दूसरी तरफ, पूर्वानुमानकर्ताओं को उनके पूर्वानुमान के बारे में जानना आवश्यक होता है जिससे कि वे उसमें सुधार ला सकें। वर्ष 2013 में हमने एक नई सत्यापन स्कीम का विकास किया जिससे पूर्वानुमान प्रणाली के विकास तथा वर्तमान स्वचालित प्रेक्षण संजाल का लाभ लिया जा सकता है। यह सत्यापन प्रणाली, जिसे COMFORT (सतत मिटिओस्विस फोरकास्ट क्वालिटी) कहा जाता है, का डिजाइन संचार व्यवस्था और प्रबंधन उपलब्ध कराने के लिए किया गया। इसके साथ ही मिटिओस्विस द्वारा उपलब्ध कराए गए सामान्य पूर्वानुमानों की गुणवत्ता की जाँच के लिए बाहर के लोगों जैसे नीति निर्माता, मिडिया आदि के लिए भी इसे डिज़ाइन किया गया है। स्कोर COMFORT का विकास मिटिओस्विस प्रचालनात्मक पूर्वानुमान प्रणाली के अंदर किया गया। इसकी स्पष्ट विशिष्टता के बावजूद, COMFORT सामान्य विचारों पर आधारित है जिसे अन्य मौसम पूर्वानुमानन सेवाओं पर प्रत्यक्ष रूप से अंतरित किया जा सकता है। COMFORT को विकसित करने का मुख्य भाग अनुकरणों (सिमुलेशन) का निष्पादन करना था ताकि बचाव करने हेतु उसका संतुलन अथवा पूर्वानुमानों के वैश्विक सुधार को दर्शाने की उसकी क्षमता जैसे स्कोरों के विभिन्न गुणों वाले वास्तविक आँकड़ों की जाँच की जा सके।

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 can 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 system, 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 measurement of the quality of general forecasts provided by MeteoSwiss. The score COMFORT was developed within the MeteoSwiss operational forecasting system. In spite of its apparent specificity, COMFORT is based on general ideas that might be directly transposable to other weather forecasting services. An important part in the development of COMFORT was to perform simulations in order to test with real data different properties of the score such as its robustness against hedging or its ability to reflect a global improvement of the forecasts.

Key words – Forecast verification, Accuracy, Global score, Administrative score, Sensible weather, Communication.

1. Introduction

The COMFORT forecast verification method was developed at the Swiss Federal Office of Meteorology and Climatology in order to provide a simple measurement of some attributes of the quality of the forecast sources used to produce general forecasts. An important issue was to be able to explain in a simple way the variations of a global score to different entities not specialized in forecast verification, such as hierarchy, policy makers, press, etc.

A variety of verification methodologies were developed in the last few decades, resulting in a profusion of scores assessing various characteristics of the forecasts [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. Also, we follow a measure-oriented approach rather than a distribution-oriented approach [Murphy and Winkler, 1987; Stanski *et al*., 1989] 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 it preserves temporal and spatial dependency. Two requirements that COMFORT should ideally fulfill are that it should encode in a single value the general forecast quality with the capacity to provide intuitive and intelligible explanation for a good/bad 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.

In Section 2, we define a Global Continuous Accuracy Score (GCAS). A GCAS is a linear combination of partial scores defined for each verified quantity. Each quantity is assumed to be continuous. 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 Section 3, we briefly present the operational forecasting system of MeteoSwiss. As we shall see, bench forecasters perform deterministic forecasts by quantitatively editing, for a number of regions, a series of parameters describing sensible weather. Then, in Section 4, we define the score COMFORT by applying the principles introduced in Section 2.

A significant part of the work related to the development of COMFORT was devoted to simulations with the aim to test with real data during a period of three years running from 2010 to 2012 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. A selection of results is presented in Section 5.

2 Verification principles

In this work, only deterministic forecasts are considered. 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. A score valued between 0 and 100 with higher values corresponding to better forecast accuracy, *i.e.*, 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 $|f - o|$ μ then the score obtained by the pair (f, o) is maximized (= 100). The tolerance threshold reflects the principle that an error which is small enough does not affect the quality of the forecast.

- **a utility threshold α** : if $|f - o| > \alpha$ then the score obtained by the pair (f, o) is minimized (= 0). The utility threshold reflects the principle that an error which is too large makes 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

$$\mathrm{CAS}(f,o) = \max\left\{0,100\cdot\left[1-\frac{\mathrm{ERR}_{\mu}(f,o)}{d}\right]\right\} \qquad (1)$$

where, $\mathrm{ERR}\mu(f, o) = \min \{\mathrm{ERR}(o, z); z \in [f - \mu, f + \mu]\}$ 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):

$$\mathrm{GCAS} = \sum_{i=1}^{n} \rho_i \mathrm{CAS}_i \qquad (2)$$

where n is the number of quantities involved.

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.

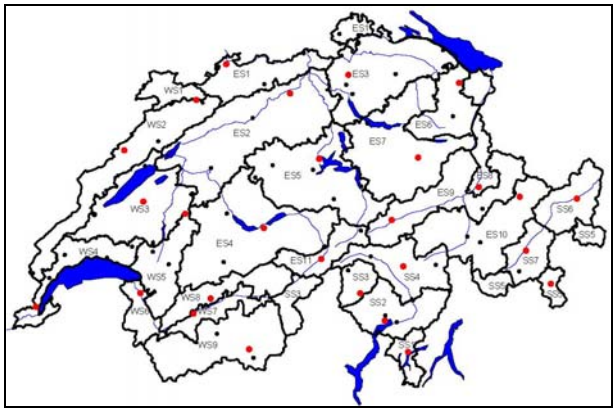

Fig. 1. The 27 forecast regions of the Matrix Editor used for short-range forecasts. These regions are used for the forecast verification at all time-ranges

The weights ρ_i in equation (2), which should always sum to 1, represent the relative importance of each verified quantity in the global score and can be adjusted according to the verification context. As previously argued, it is desirable for communication purposes to perform a verification of parameters representing sensible weather. Thus, 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 explain below.

It should be noted that 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 have left it out from our verification framework. However, for countries in which relative humidity is a relevant feature, *e.g.*, in India, this quantity can be added to the list of verified parameters. Following the principles presented in this section, one can define a partial score for relative humidity in an analogous way than for relative sunshine (see Section 4.2).

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 (2) according to its climatological and administrative specificities. Obviously, if for a given parameter different tolerance and utility thresholds are

TABLE 1

Partition of relative sunshine into classes and corresponding cloud coverage in okta

RS [%]	$0 \leq RS < 5$	$5 \leq RS < 20$	$20 \leq RS < 50$	$50 \leq RS < 80$	$80 \leq RS \leq 100$
okta	●	◐	◓	◔	○
Description	"Cloudy"	"Mostly cloudly"	"Partly sunny"	"Mostly sunny"	"Sunny"

used, then direct comparison is trickier. Common thresholds can always be set, keeping in mind that comparison is made between the absolute accuracy of the forecasts rather than between their respective skill.

3. Data

Bench forecasters working at MeteoSwiss are editing through a graphical interface named the Matrix Editor either numerical values or categories (the latter for relative sunshine) representing deterministic forecasts for a number of regions. The spatial resolution of a forecast edited in the Matrix Editor depends on the forecast's time-range. As shown on Fig. 1, 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). Each region is assigned a reference station (indicated by the red dots on the map), as well as a number of observation stations (indicated by the black dots on the map), each reference station being an observation station itself.

The verified quantities are of two types. Temperature and wind speed are local quantities: predicted values are attributed by forecasters at the reference stations and they are verified using observations from the reference stations only. Precipitation and relative sunshine are regional quantities: predicted values attributed by forecasters represent averages over the forecasted region and they are verified using average observations over the region. Relative daily sunshine duration RS is edited by forecasters using five sunshine classes according to the partition shown in Table 1.

For the verification of relative sunshine, mean observations for a given region are obtained by averaging measures from a number of representative stations situated in the region. For the verification of precipitation, we benefit of a multi-sensor observation scheme, called CombiPrecip [Sideris *et al*., 2014]. This tool provides precipitation estimates at a very high spatial and temporal resolution using a combination of a continuous field of precipitation provided by radar images and of sparser measurements provided by the automatic rain gauge network. Geostatistical techniques such as kriging with external drift are generalized and used in order to perform a smart calibration of the radar estimates (Fig. 2). Regional mean amounts used for the verification are then obtained from the high resolution grids by taking the average of the values at the grid-points belonging to a given region.

We have performed all tests during the development of COMFORT using the forecasts issued by forecasters during a period of three years running from 1 January, 2010 to 31 December, 2012.

4. Application

We apply now the principles of Section 2 to define the global score COMFORT. The verified quantities represent sensible daily weather and are edited by MeteoSwiss bench forecasters (Section 3). As already discussed in Section 2, the choice of the free parameters μ and α in each partial score is based on empirical estimation of what a correct/useless forecast is, and is supported by a series of tests (Section 5).

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 correct whereas a score of 0 means that the forecast is useless. The following quantities are verified by the COMFORT score:

(*i*) **precipitation** (denoted by P) : daily amount [mm]

(*ii*) **relative sunshine** (denoted by RS) to the maximum daily sunshine duration in [%]

(*iii*) **minimum daily temperature** (denoted by T_{min}) in [°C]

(*iv*) **maximum daily temperature** (denoted by T_{max}) in [°C]

(*v*) **wind speed** at 10 m above ground level (denoted by *V*) : maximum hourly average between 6 am and 6 pm in [kt].

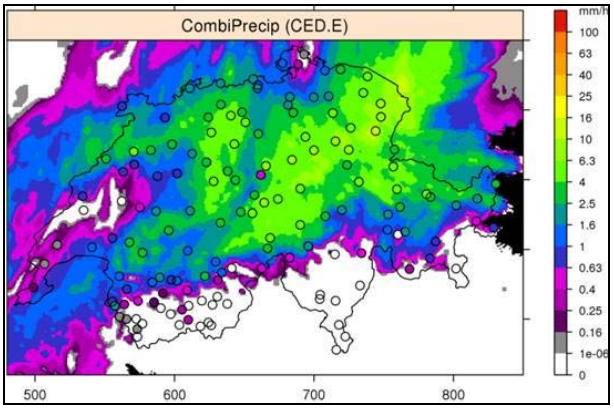

Fig. 2. Example of an observation grid generated by CombiPrecip: a combination of radar images and measurements from the automatic rain gauge network (circles) provides regional observations used for the verification of precipitation forecasts

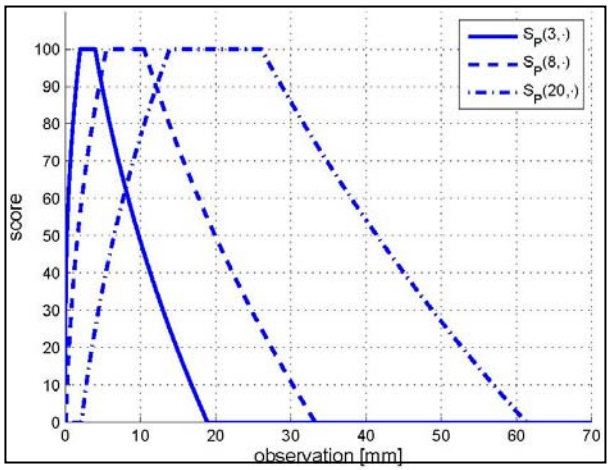

Fig. 3. Behaviour of the partial score SP (f, o) with respect to the observation o, for three different values of the forecast : f = 3, f = 8 and f = 20 [mm]

For each of the previous quantities, a partial score is defined according to (1) (see Subsections 4.1 to 4.3). The global COMFORT score is then a particular case of (2):

$$\text{COMFORT} = \rho p S_p + \rho_{RS} S_{RS} + \rho T_{\min}\, ST_{\min} + \rho T_{\max}\, ST_{\max} + \rho v S v \qquad (3)$$

where S_P is the partial score for precipitation, S_{RS} is the partial score for relative sunshine, $ST_{\min}$ and are partial scores for minimum and maximum daily temperatures respectively and S_V is the partial $T_{\max}$ 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 [Met. Office, 2010]:

$$\rho_P = \rho_{RS} = 0.3;\ \rho T_{\min} = \rho T_{\max} = 0.15;\ \rho_V = 0.1 \qquad (4)$$

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. Also, as already discussed in Section 2, the list of parameters verified by COMFORT was established in order to cover the main features of sensible weather in Switzerland. This list can be adapted to countries with different climatologies by including whenever necessary other parameters, such as relative humidity.

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

4.1. *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 :

$$S_p(f,o) = \begin{cases} 100 & \text{if } |f-o| \le \mu(f), \\ 100.\left[1 - \dfrac{o^p - [f+\mu(f)]^p}{d}\right] & \text{if } 0 < o^p - [f+\mu(f)]^p < d, \\ 100.\left[1 - \dfrac{[f-\mu(f)]^p - o^p}{d}\right] & \text{if } 0 < [f-\mu(f)]^p - o^p < d, \\ 0 & \text{otherwise} \end{cases} \qquad (5)$$

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 μ (f) = 0.3f + 0.1, so that the score granted to the pair (f, o) is maximized

TABLE 2

Annual partial scores and COMFORT score, averaged over Switzerland, for the period 20102013. The forecast time-ranges are D1, D3 and D5

	D1				D3				D5			
	2010	2011	2012	2013	2010	2011	2012	2013	2010	2011	2012	2013
S_p	80.6	84.4	82.5	82.1	75.8	79.8	75.9	76.4	68.7	73.3	69.1	71.1
S_{RS}	80.7	83.4	82.6	82.6	75.2	77.4	75.5	75.0	68.1	69.9	68.6	69.7
$ST_{\min}$	81.6	83.5	82.1	82.9	76.6	77.3	76.7	77.6	71.5	71.2	68.5	70.7
$ST_{\max}$	86.2	87.9	87.6	87.1	79.8	80.5	79.7	79.6	70.1	70.2	70.2	71.0
S_v	58.7	63.3	61.6	66.2	53.9	61.9	58.9	62.2	52.0	58.4	55.7	60.8
COMFORT	79.4	82.4	81.2	81.5	74.1	77.0	74.8	75.2	67.5	70.0	67.7	69.6

whenever the observation o falls inside a neighborhood of 30% of the magnitude of the forecast *f*. The choice of letting μ depend on the forecast is deliberate. By slightly favorizing humid forecasts, we encourage 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 $d = 3/2$. Fig. 3 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 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 [(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).

4.2. *Partial score for relative sunshine*

The forecast for relative daily sunshine duration RS is edited by forecasters using sunshine classes (Section 3). Observations are continuous values bounded by 0 and 100. The classical approach would be to partition the observations into the same categories as forecasts before comparing them, using the multi-categorical verification framework. Instead of this, we shall avoid reducing observations into classes and use formula (1) with a tolerance threshold corresponding to the forecasted class: the score is maximized whenever the observation falls into the same class than the forecast and the score decreases continuously (and linearly) from the bounds of the forecasted class. More precisely, denoting by [a(f), b(f)] the class of the forecast f (for instance, if f = 35 then [a(f),b(f)[= [20, 50[), one defines

$$S_{RS}(f,o)=\begin{cases}100 & \text{if } \alpha(f)\le o\le b(f),\\ 100.\left[1-\dfrac{o-b(f)}{d}\right] & \text{if } 0<o-b(f)<d,\\ 100.\left[1-\dfrac{a(f)-o}{d}\right] & \text{if } 0<\alpha(f)-o<d,\\ 0 & \text{otherwise}\end{cases} \tag{6}$$

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 classes are not constant (Table 1). The free parameter d defines the utility threshold: $\alpha = \mu + d$; if the observation falls further away than d% from the upper/lower bound of the forecasted sunshine class, then SRS (*f*, *o*) = 0. For our verification purposes, we have retained the value $d = 40$.

4.3. *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

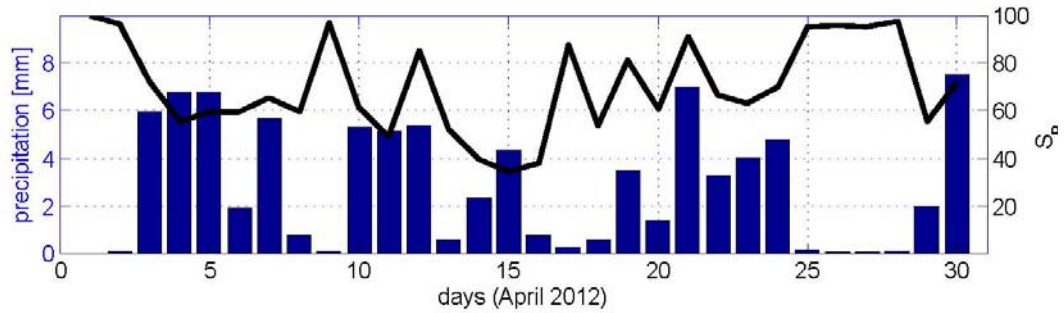

Fig. 4. The lines show the daily evolution of the partial scores for precipitation during April 2012 for the "administrative" region West. The bars show the corresponding observations averaged over the same region

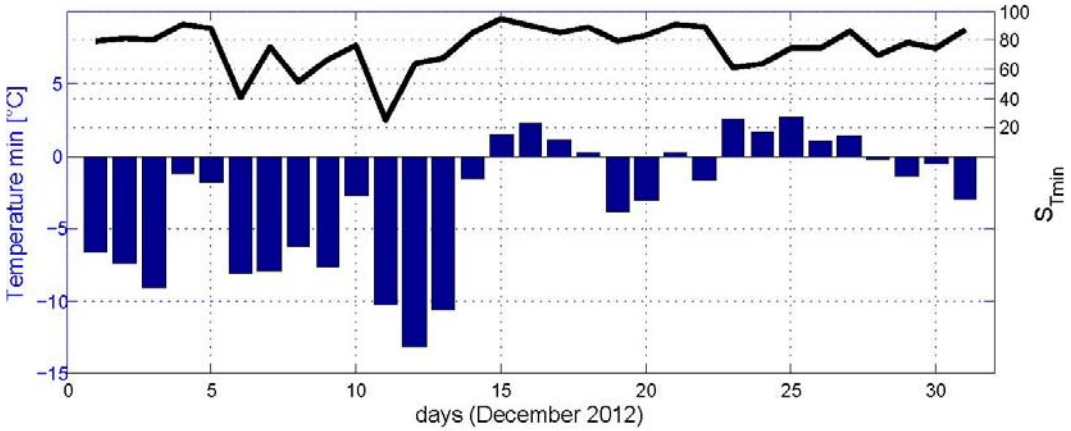

Fig. 5. The lines show the daily evolution of the partial scores for minimum temperatures during December 2012 for the "administrative" region West. The bars show the corresponding observations averaged over the same region

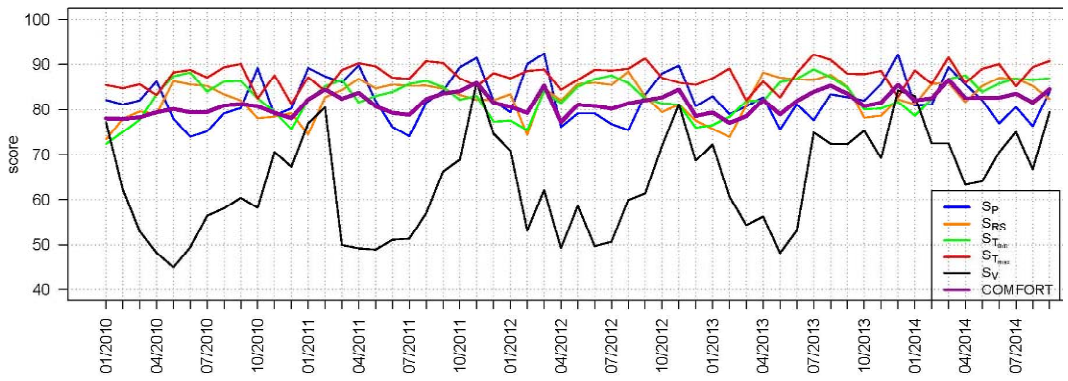

Fig. 6. Monthly evolution from January 2010 to September 2014 of the COMFORT score and the partial scores composing it, averaged over Switzerland. The time-range of the presented forecast is D1

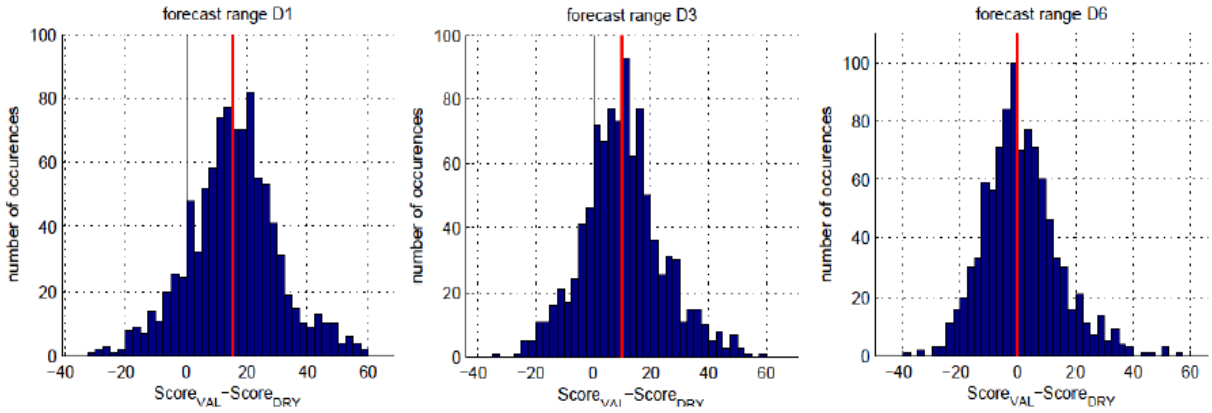

Fig. 7. Empirical distribution of the differences in monthly scores S_P obtained by the forecasts VAL and DRY. The sample median is shown in red

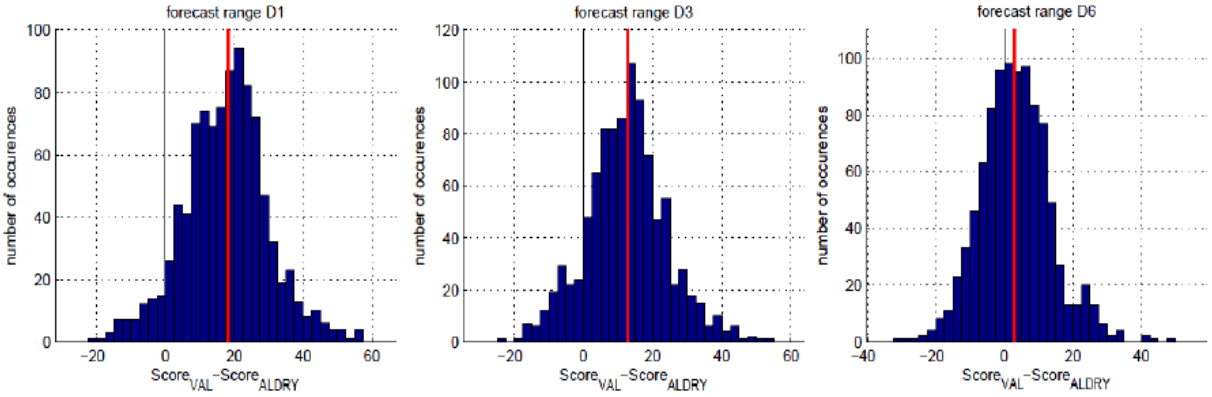

Fig. 8. Empirical distribution of the differences in monthly scores S_P obtained by the forecasts VAL and ALDRY. The sample median is shown in red

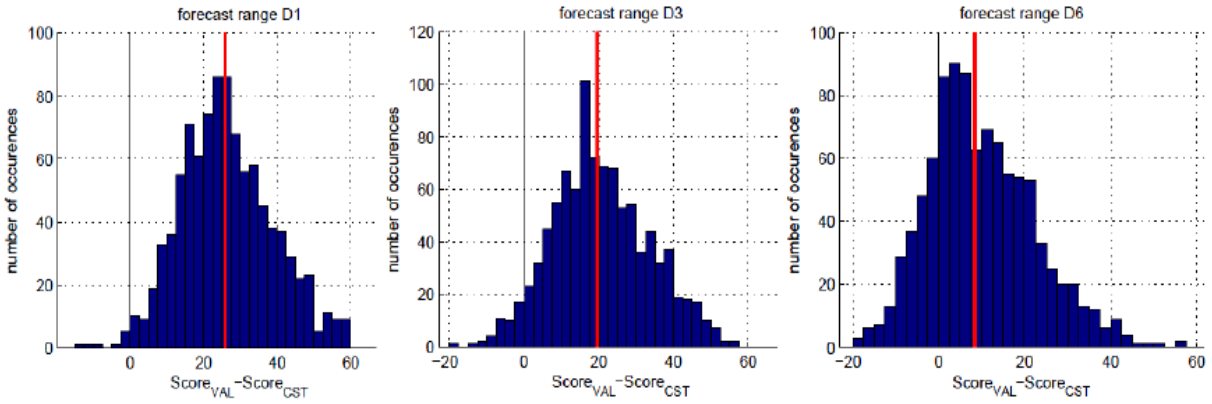

Fig. 9. Empirical distribution of the differences in monthly scores S_{RS} obtained by the forecasts VAL and CST. The sample median is shown in red

TABLE 3

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	76.89	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

TABLE 4

Delta : average differences between monthly precipitation scores obtained by the forecasts VAL and DRY (resp. ALDRY). Ratio: empirical probability of obtaining a better score when forecasting the scheme DRY (resp. 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

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

$$S_{\text{Param}}(f,o)=\begin{cases}100 & \text{if } |f-o|\le\mu,\\ 100.\left[1-\dfrac{|f-o|-\mu}{\alpha-\mu}\right] & \text{if } \mu<|f-o|\le\alpha,\\ 0 & \text{otherwise}\end{cases} \tag{7}$$

where, Param ε[T_{min}, T_{max}, V]. For our verification purposes, we have fixed the following thresholds:

- maximum temperature: μ =1 [°C] and α = 6 [°C]
- minimum temperature: μ =0.5 [°C] and α = 6 [°C]
- wind speed: μ = 2.5 [kt] and α = 5 [kt].

5. Results and discussion

This section contains a selection of the results obtained during the development of the COMFORT score and discusses some of its main features. Results presented in this section (except Fig. 6 and Table 2 which also contain latest data) are based on forecast and observation data issued from 2010 to 2012.

5.1. *From rough values to a finer analysis*

As already argued in the introduction, a convenient feature of COMFORT is that the global score obtained for a given period can be easily decomposed parameter by parameter, both spatially and temporally, making easier the interpretation of its values. The Table 2 contains annual (partial and global) scores, averaged over the entire of Switzerland. Three time-ranges are shown: D1, D3 and D5. The Fig. 6 shows the temporal evolution of monthly scores, again averaged over the entire country, for time-range D1.

Different observations can be drawn out from this plot, 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, 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 notices a clear improvement in accuracy since July 2013, which coincides with the operationalization of a new model output statistics.

When analyzing COMFORT's results for specific periods and regions, it is convenient to "zoom" on them. The daily scores obtained for a given region can then be put into relation with the meteorological context, for further analysis. For instance, Figs. 4 and 5 show the daily evolution of partial scores obtained for precipitation and minimum temperature forecasts, for months during which the scores were below the average. In both cases, the focus is on the Western administrative region of Switzerland. As we can see from the precipitation plot, dry periods of several days usually give good scores. On the opposite, successions of wet days but not necessarily with very large amounts of precipitation (as around the 15. of April) are often trickier to forecast accurately. When having a look on the temperature plot, we see that the first half of the month was much colder (actually much below the December norm) than the second half. The scores during the first period are clearly worse than during the second one, reflecting probably difficulties to forecast "extreme" values.

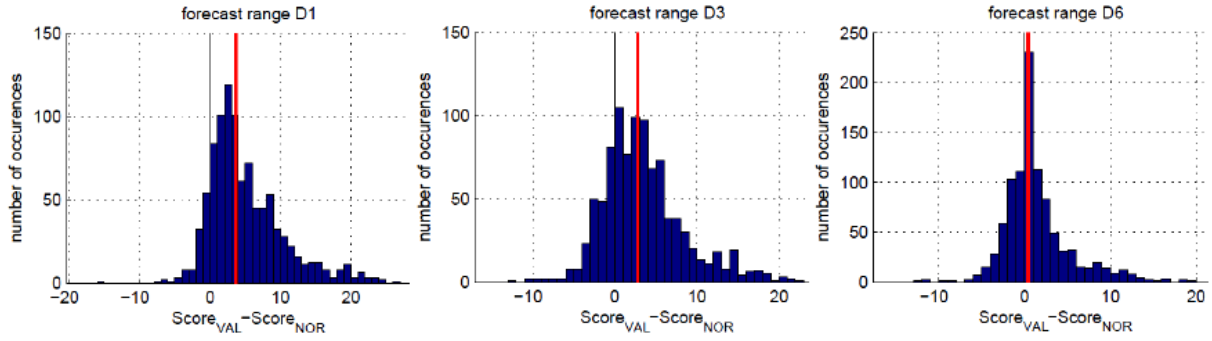

Fig. 10. Empirical distribution of the differences in monthly scores S_{RS} obtained by the forecasts VAL and NOR. The sample median is shown in red

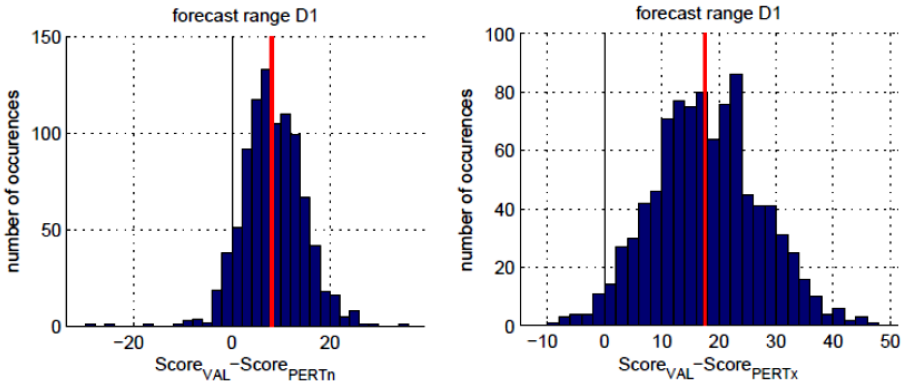

Fig. 11. Empirical distribution of the differences in monthly scores ST_{min} (left) and ST_{max} (right) obtained by the forecasts VAL and PER for time-range D1. The sample median is shown in red

5.2. *Robustness against hedging*

We have tested the robustness of the score COMFORT against hedging by considering different "no-skill" or "no-risk" forecasts in order to ensure that there is no obvious systematic way of obtaining better scores over a long period of time, especially for short-range predictions, by forecasting some predefined scheme. This should encourage forecasters to issue their forecasts according to their best-judgement. All results presented below as well as in the Tables were obtained by pooling together monthly scores over the test-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 any day) or "minimal rain" (*i.e.*, 0.2 [mm] for any day). The corresponding forecasts are denoted by DRY and ALDRY, respectively. Table 4 shows the average differences between monthly precipitation scores obtained by the forecasts VAL and DRY (respectively 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 represents forecaster's best judgement. As we can see from Table 4, at least until three days ahead, emitting a "lazy" forecast has little chance to be rewarded better than the "best judgement"; for long-term forecasts, there is about 50% chance of being rewarded better. Figs. 7 and 8 show the empirical distributions of the differences in monthly partial scores for precipitation SP obtained by the forecasts VAL and DRY (respectively ALDRY).

Edited forecasts for relative sunshine were compared with "no-skill" forecasts obtained by constantly forecasting the "climatological" mean sunshine class. The forecast denoted CST consists 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 Fig. 1). For these regions, the class [50, 80] is forecast instead.

TABLE 5

Delta : average differences between monthly sunshine scores obtained by the forecasts VAL and CST (resp. NOR). Ratio: empirical probability of obtaining a better score when forecasting the scheme CST (resp. 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

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] is forecasted instead of [0, 5] and the class [50, 80] is forecasted instead of [80, 100[). The aim of this test was to check whether predicting "safelooking" albeit not the climatological distribution average values for relative sunshine was not unduly favored. Results are summarized in Table 5. As expected, differences between NOR and VAL are much smaller than between CST and VAL, but remain in favour of VAL. There is 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. Figs. 9 and 10 show the empirical distribution of the differences in monthly partial scores for relative sunshine SRS obtained by the forecasts VAL and CST (respectively NOR).

Edited forecasts VAL for minimum and maximum temperatures were compared with the persistence forecast, denoted by PER. Fig. 11 shows the empirical distribution of the differences in monthly partial scores $ST_{\min}$ and $ST_{\max}$ 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 is slightly less than a 10% chance to get a better score than if editing it according to the "best judgement", whereas this probability is only about 3% for maximum temperature.

5.3. *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. Although the 2AFC score measures forecast quality from a different side than COMFORT, it is nevertheless interesting to compare both scores' results. The point here is not to show that both scores yield similar absolute values for a given parameter, as they do not measure the same quantities, but that across all verified parameters, the differences between them have almost all same signs (both scores are by definition positively orientated) and magnitude; see Table 3. This contributes to show 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.

6. Conclusion

We have developed a new score to measure the overall weather forecast accuracy. This score was recently introduced by our administration to report on the forecast quality as well as to fix objectives to the weather centers for the next years. From communication side, factsheets introducing the new verification scheme were produced for distribution among administration and media. As a perspective, some of our final forecast products may be accompanied with the corresponding global COMFORT score, or with partial scores composing it.

Based on the partial scores defined in Section 4, a bulletin containing a detailed verification of the very last forecasts is generated and sent every day to the forecasters. It is challenging to forecasters to improve their forecasts having in mind that COMFORT is more severe than the former scores used at MeteoSwiss but that it correctly rewards "best judgement" forecasts, as there is no way to hedge the score with a given strategy (Subsection 5.2).

Acknowledgements

The authors wish to acknowledge all collegues from MeteoSwiss that have contributed either materially or conceptually to the development of COMFORT. In particular, we would like to thank Pirmin Kaufmann for

his numerous interesting remarks during the whole project. We are especially grateful to the entire forecasting team from Locarno for their strong interest in this project; this resulted in valuable discussions and contributed significantly to the development of COMFORT. Finally, we would like to thank Pierre Eckert for his continuous support as the Director of the forecasting center located in Geneva.

References

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

Jolliffe, I. T. and Stephenson, D. B., editors, 2012, "Forecast verification: a practitioner's guide in atmospheric science", Wiley-Blackwell, 2nd edition.

Mason, S. J. and Weigel, A. P., 2009, "A generic forecast verification framework for administrative purposes", *Mon. Wea. Rev.*, **137**, 331-349.

Met Office, 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 Winkler, R. L., 1987, "A general framework for forecast verification", *Mon. Wea. Rev.*, **115**, 1330-1338.

Sideris, I. V., Gabella, M., Erdin, R. and Germann, U., 2014, "Real-time radar-raingauge merging using spatiotemporal co-kriging with external drift in the alpine terrain of Switzerland", *Q. J. Roy. Meteor. Soc*., **140**, 680, 1097-1111.

Stanski, H. R., Wilson, L. J. and Burrows, W. R., 1989, "Survey of common verification methods in meteorology", Technical report, World Meteorological Organization.

Wilks, D. S., 2011, "Statistical methods in the atmospheric sciences", volume 100 of International geophysics series. Academic Press, 3rd edition.