A Global Forecast Quality Score for Administrative Purposes

D. Cattani, M. Matter, MeteoSwiss

Introduction

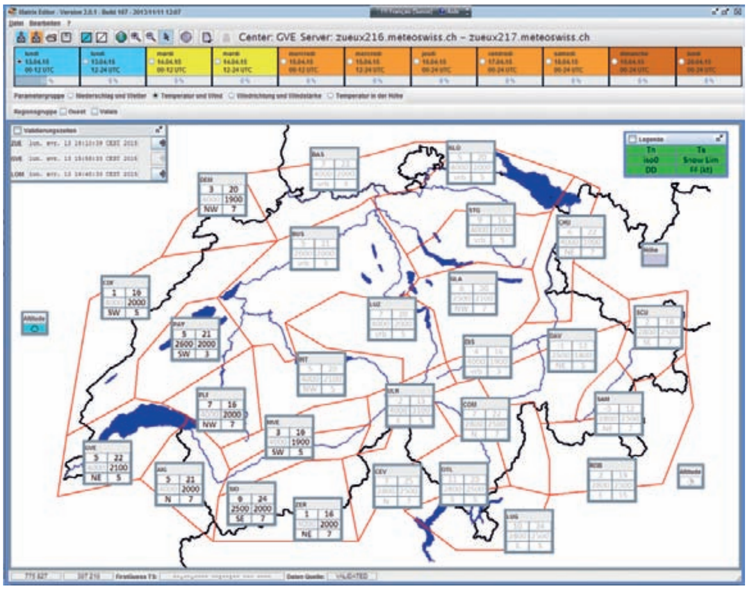
Since 1985, MeteoSwiss has used a global score for systematically assessing the basic weather forecasts issued by the regional forecasting centres. This assessment is done for two main reasons. Firstly, it is used for administrative purposes as the weather centres are expected to communicate in a simple way to the general public and to the government the evolution of the quality of their forecasts. On the other hand, the forecasters need to know the performance of their predictions in order to improve them. In 2013, we developed a new verification scheme, called COMFORT (for COntinuous MeteoSwiss FORecast qualiTy), which also accounts for benefits from the evolution of the forecasting system as well as of the present automated observation networks.

COMFORT is a global measure of accuracy which verifies deterministic forecasts of quantities representing sensible weather in Switzerland, namely: precipitation (without distinction of its type), sunshine duration, minimum and maximum temperatures, and wind speed. Specifically, COMFORT assesses the generic forecasts which are edited numerically by the forecasters. These forecasts serve then as a basis for generating a variety of products, ranging from web apps to agriculture or support for TV broadcasts.

A requirement that COMFORT had ideally to fulfil was to encode in a single value the general forecast quality, together with the capability to provide intelligible explanation for a high/low global score, typically computed over a long period and over a large territory, to people that are neither experts in verification, nor forecasters. A way of conciliating these conflicting requirements was to make it possible to focus on specific periods and/or geographical areas in order to detect and analyse forecasts whose accuracy deviates from the average. Also, forecasts for all time-ranges are verified using the same spatial and temporal resolutions, which allows comparison across different lead times. In parallel, COMFORT can be applied to NWP forecasts, typically the “First Guess” predictions which initialize the forecast editing tool used by MeteoSwiss bench forecasters, making it possible to measure forecasters’ added value with respect to NWP.

Data Used in the Verification

Bench forecasters working at MeteoSwiss edit their predictions with a graphical interface named the *Matrix Editor*. These are either numerical values or categories (the latter only for relative sunshine which is edited according to five classes) and represent 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. The Swiss territory is partitioned into 27 regions for short-range forecasts (time-ranges D1 and D2), into 11 regions for medium-range fore-


Center: GVE Server: zueux216.meteoswiss.ch – zueux217.meteoswiss.ch
Legende
Tn
Tx
iso0
Snow Lim
DD
FF (kt)

▲ *Figure 1: Matrix tool*
Tool used by the forecaster by which they modify a first guess, with a station corresponding to each region. Sunshine duration, precipitation, temperature minimum and maximum and wind are forecast.

casts (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, as well as a number of observation stations, each reference station being an observation station itself.

The verified quantities are of two types: temperature and wind speed are defined in the Matrix Editor as local quantities, which means that the predicted values attributed by forecasters hold for the reference stations only; they are thus verified using observations from 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. Regional quantities are verified using averaged observations over the corresponding region. 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 from a multi-sensor observation scheme, *called CombiPrecip* [Sideris et al., 2011]. 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. 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.

Verification Principles

As mentioned in the introduction, we consider deterministic forecasts only. For any verified quantity, the forecast's accuracy is split into three qualifications : *correct, useful* and *useless*. These categories are defined by two thresholds that should be seen as tunable parameters which depend on the verification context. Both thresholds are defined once for all when setting up the verification framework. The first threshold defines a tolerance interval μ around the forecast value. This threshold should be seen as an estimation of the maximum error below which a forecast is assumed as completely correct. The second threshold is the maximum error beyond which the forecast is considered too erroneous to be of any value, and defines the utility interval α around the forecast value. Deciding whether a forecast is correct or useless remains largely subjective and depends on the verification context. For instance, the threshold values that we have defined for the maximum temperature are: μ = 1 °C and α = 6 °C. Between these thresholds, the accuracy of the forecast is measured as for a continuous quantity (for instance using mean absolute error).

This approach explained above can be applied independently to each verified quantity, which in our case are:

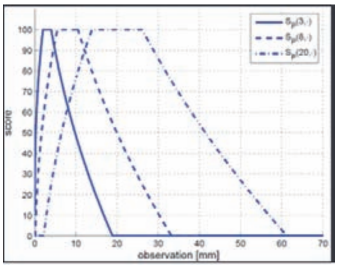
1. **precipitation** (denoted by P): daily amount [mm]
2. **relative sunshine** (denoted by RS) with respect to the maximum daily sunshine duration in [%]
3. **minimum daily temperature** (denoted by Tmin) in [° C]
4. **maximum daily temperature** (denoted by Tmax) in [° C]
5. **wind speed** at 10m above ground level (denoted by V): maximum hourly average between 6am and 6pm in [kt].

A partial score Si is defined for each of the previous parameters. For all quantities but precipitation, the score between the tolerance and the utility thresholds decreases linearly. Also, for precipitation the tolerance and utility intervals depend on precipitation intensity, reflecting the assumption 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 (Fig 2).

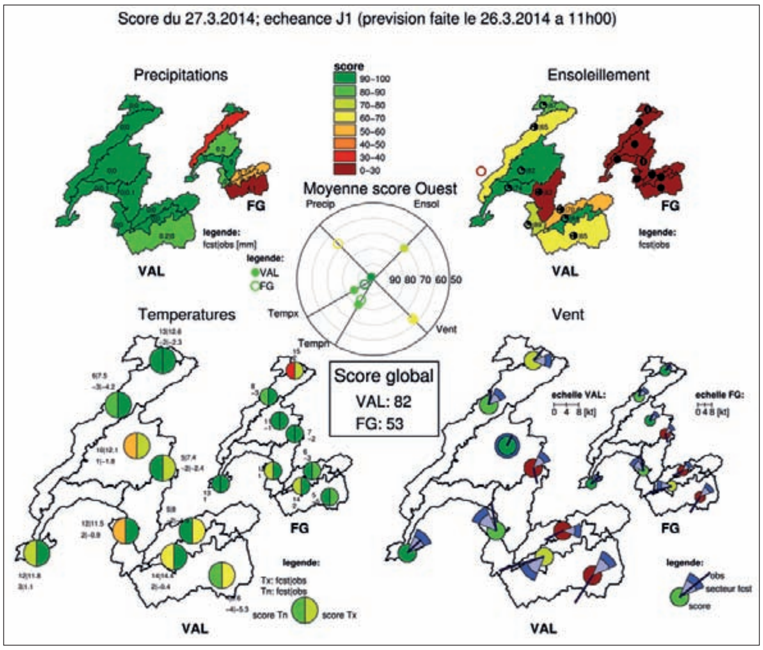
Before combining scores for different quantities into a single value, since errors might be of different magnitudes depending on the verified quantity, it is necessary to rescale them on a common scale. A score valued between 0 and 100 with higher values corresponding to better forecast accuracy is intuitive. The global COMFORT score is then obtained as a weighted sum of partial scores computed for each verified quantity:

COMFORT=ρpSp+ρRSSRS+ρTminSTmin+ρTmaxSTmax+ρVSV

where ρp, ρRS, ρTmin, ρTmax, ρV are the weight of the partial scores for respectively precipitation, relative sunshine, minimum and maximum temperatures and wind speed.



▶ *Figure 2: Behaviour of the partial score for precipitation with respect to the observation, for three different values of the forecast; 3, 8 and 20 mm.*



▲ *Figure 3: Example of daily analysis with the scores for precipitation, sunshine duration, temperature minimum and maximum and wind for the forecaster (VAL) and first guess (FG). In the centre is the score for each parameter and underneath the global score.*

The tuning of the parameters μ and in each partial score can be made 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 their requirements. 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.

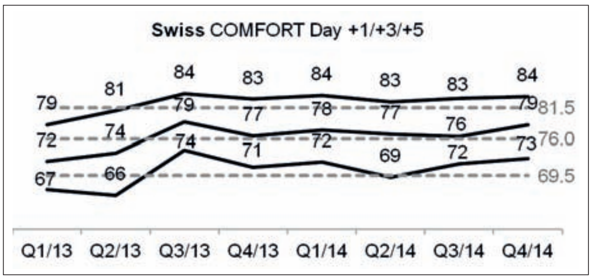
The weights i, which should always sum to 1, represent the relative importance of each verified quantity in the global score and can also be adjusted according to the verification context. We give a similar weight to all verified parameters except for wind: 0.3 for precipitation, 0.3 for sunshine, 0.15 for Tmin as well as for Tmax, and 0.1 for wind. 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 which capture the dominant winds blowing in Switzerland. For countries with larger flatlands or coastlines, where measures might be more representative of the regional weather conditions, more importance might be given to this parameter.

Tests

A significant piece of the work related to the development of the COMFORT score was devoted to simulations with the aim of testing with real data 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.

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 best-judgement.

Simulations were also made with the aim of estimating COMFORT's sensitivity to different theoretical forecast enhancements. On one hand, this was useful to answer a question asked by the MeteoSwiss leadership when fixing quantitative



▲ *Figure 4: Quarterly score communicated to the government. Solid lines represents the global score performed by the forecaster within the 3 weather centres in Switzerland, and dashed line represent the global to achieve for 3 forecast ranges.*

medium and long-term objectives for the score. On the other hand, we 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.

Communication

Every quarter, the global score COMFORT obtained by three administrative regions for the elapsed period is communicated to the leadership and to the government, allowing them to monitor the overall evolution of the forecast quality. Fig 4 shows the quarterly evolution of the score for day +1, day +3 and day +5; the dashed lines represent the goal fixed by the government.

In parallel, regular feedback in the form of daily bulletins verifying in greater detail individual forecasts is automatically delivered to forecasters (see Fig 3). These bulletins show the partial scores of the forecasters' predictions (VAL) as well as scores achieved by the NWP "First Guess" forecast (FG) for all forecast regions and for a given day. In particular, this allows forecasters to see what value they added to NWP on a concrete occasion. For each verified quantity, mean values for the whole responsibility region are provided, as well as a global score for that day.

From this daily feedback, results over a given period (a season or a year) can be gathered together under the form of periodic analysis to find out the strengths and weaknesses of the forecasts and, whenever possible, to provide guidelines to forecasters.

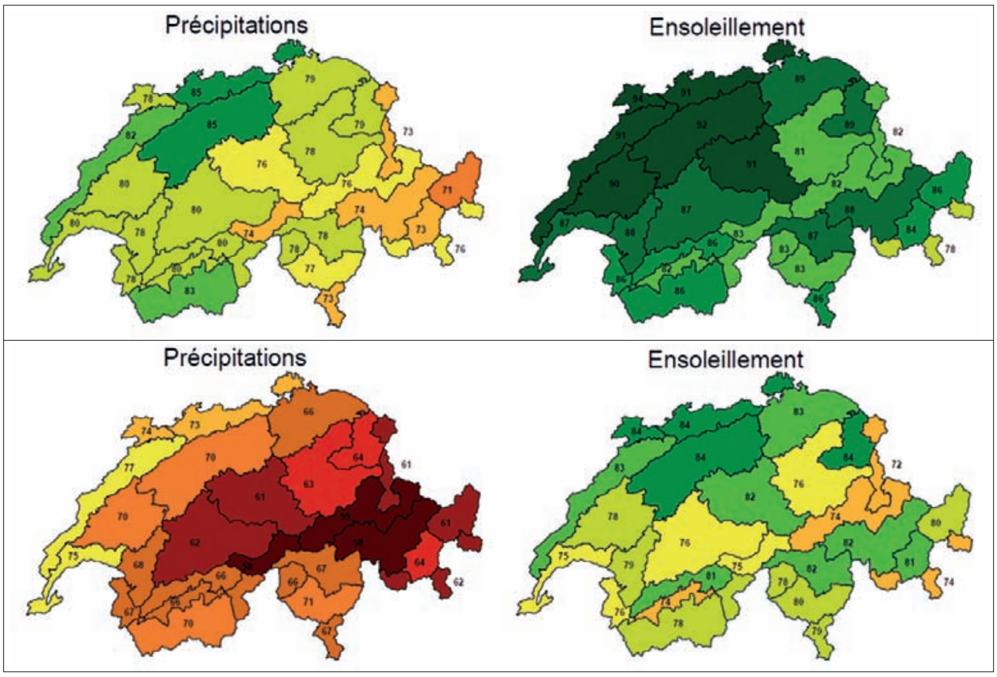
Fig 5a,5b show the scores of precipitation and sunshine duration for the summer 2014 for respectively day +1 and day +3.

Reference

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

Letestu A.-C. "The New Production Process at MeteoSwiss". The European forecaster, 15, may 2010.

Cattani, D., Faes A., Giroud Gaillard M., Matter M., "Global Forecast Quality Score for Administrative Purposes". To appear in MAUSAM (July 2015)



◀ *Figure 5a, 5b: Partial COMFORT score for precipitation and sunshine duration calculated for summer 2014 (june, july, august) for respectively D1 and D3*