Model Verification and Its Relevance for Forecasters

André-Charles Letestu, MeteoSwiss

Introduction

Whilst numerical models are becoming more accurate and numerous, weather forecasters are becoming increasingly confused faced with so much information. Which model to trust in which meteorological situation?

It is difficult for the forecaster to know the strengths and weaknesses of all the models. The interaction between forecaster and modeller could still be improved. A good model skill doesn't have the same meaning for a forecaster or a modeller; the actual model verification is performed over long periods or averaged over large areas. This only provides general global information, but doesn't meet the specific needs of the forecaster. The forecaster would like to know specifically when to trust the accuracy of the model.

As a part of the project COST733, an evaluation of the different weather types was made by looking at the ability to represent different precipitation patterns in the Alpine domain. For verification purposes we are more interested in differentiating weather classes where the models have difficulties from those where it performs well.

History

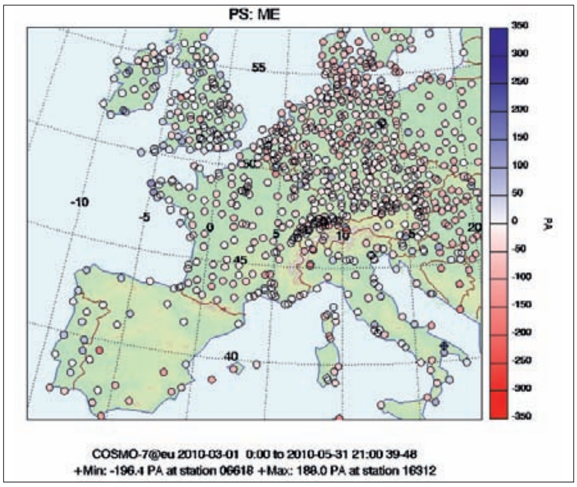
Back in 1987, the sources of information were much fewer. In Switzerland, the forecasters had access to only one model (ECMWF). In those days, the horizontal resolution was about 200 km. The representation for the Alps within the model was very rough. A large bell-shaped orography contained the Alps, Jura and Massif Central without differentiation. The forecaster needed a good knowledge of the climatology of the region as well as the understanding of the weaknesses of the model. Nowadays, the output of the model is so detailed that it is difficult for the end user to form his opinion objectively.

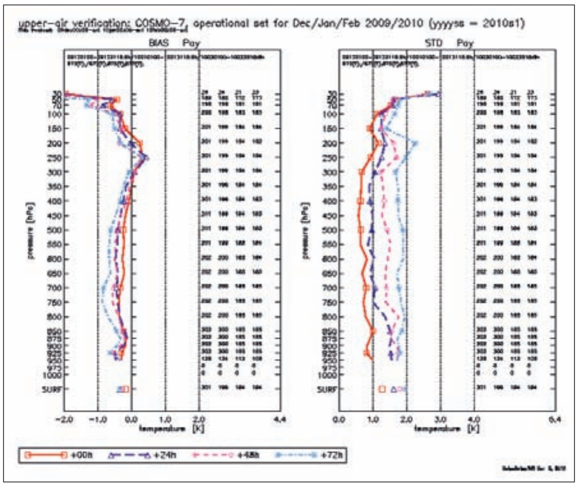
Available verification

Fig 1 shows examples of model verification available in Switzerland, issued on a regular basis. Fig1a represents the mean error over Europe of the surface pressure between the model COSMO 7 and the SYNOP observation for the Spring season. Fig 1b shows the bias and the standard deviation of the model COSMO 7 compared to the Payerne ascent and averaged over one season (Spring).

Figure 1

Systematic verification of the COSMO7 model, 1a: Pressure surface verification (mean error) comparison with SYNOP data for Spring 2010. 1b: Upper air verification (temperature), comparison with Payerne ascent (00z and 12z) for Spring 2010.





Although the verification is very useful for the monitoring of model skill, for the forecaster, very little relevant information can be extracted. In the case of the surface pressure verification, the verification period is far too long and the information is difficult to extract. For the upper-air verification, only the systematic errors can be pointed out.

Stratification by weather type

Following the above remarks, the forecaster needs a type of verification which allows him to criticise objectively the model output. Stratification by weather type is one way to achieve this.

Models

The verification presented below is based on the COSMO-7 local model. Switzerland is a participating member of the COSMO (Consortium for Small scale Modelling) and has developed two fine-mesh models, COSMO-7 and COSMO-2. Along with the IFS model, they are the most widely used by forecasters in Switzerland.

Briefly, the main features of both models are:

COSMO-7:
Grid length : 6.6 km
Vertical levels : 60
Outputs : 3 daily : 00z, 06z and 12z.
Forecast duration : 72 h
Boundary conditions : ECMWF
Own assimilation : Nudging

COSMO 2:
Grid length : 2.2 km
Vertical levels :
Outputs : 8 daily : 00z, 03z, 06z, 09z, 12z, 15z, 18z, and 21z.
Forecast duration : 24 h except for the 03z run up to 36h.
Boundary conditions : COSMO 7
Own assimilation : Nudging, radar assimilation

Stratification by weather type

Different methods of classification are currently used at MeteoSwiss, many of which are used to recover old weather situations in order to perform a similarity analysis. For the purpose of analysing the model's performance, a simpler and automatic classification must be used. Since 2002, weather situations have been analysed according to the main flow. Two classes are defined; advective and convective based on the 500 hPa wind. The advective cases are classified according to the direction of the flux in steps of 45 degrees; the convective cases are catalogued according to 3 classes - high, low and flat corresponding to the surface pressure pattern. Until 2008, the classification was performed manually; since 2009, an automatic method has been used. The results of the two methods are very consistent even if the automatic method shows more northerly flow and fewer high pressure situations.

Fuzzy verification method

Very often, verification methods spoil the skill of fine mesh models; the precipitation could be shifted either in time or in space. Globally, the rain could be well forecast but may arise too early or too late.

The idea of the fuzzy verification method (E. Ebert, 2008) is to compare forecast and observations over a square (window) containing a defined number of grid points and a chosen threshold. This window is displaced over the model area and a skill score is calculated. By varying the size of the window and the threshold, a neighbourhood plot containing the skill can be drawn with the threshold in abscises and spatial scale in ordinate (fig. 2). It shows if the model is better for high or low thresholds, respectively small or larger scale. The same plot can be used by subtracting the skill of two models (i.e COSMO-7 and COSMO-2). The

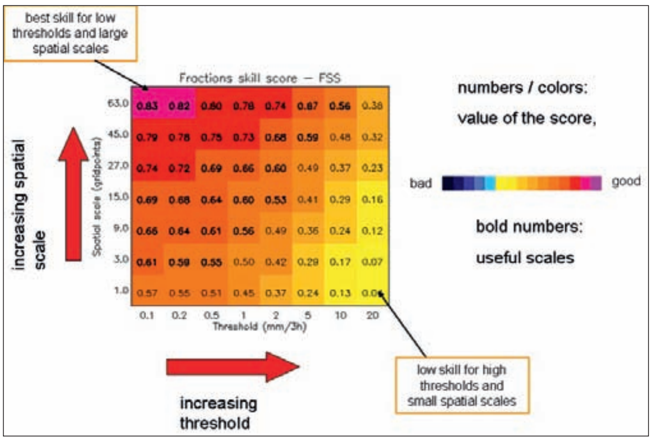


Figure 2
How to read a neighborhood plot

result will point out where a model is better than another. In case of different grid lengths, the spatial scale has to be adjusted between the two (Weusthoff et al. 2010).

Scores

Two skill scores are generally used in fuzzy verification; Upscaling (Zepeda-Arce et al. 2000) and the Fraction Skill Score (Roberts & Lean 2005). The results presented here are performed using the fraction skill score.

This score is defined thus: within the window defined above, the fraction of the observation and the model output data exceeding the threshold value over all the grid points within the window are computed. The Fraction Skill Score (FSS) is defined as the ratio between the two. The FSS is ideally suited for considering the overall performance of a model. The zero value corresponds to a mismatch and 1 to a perfect match.

Results for 2009

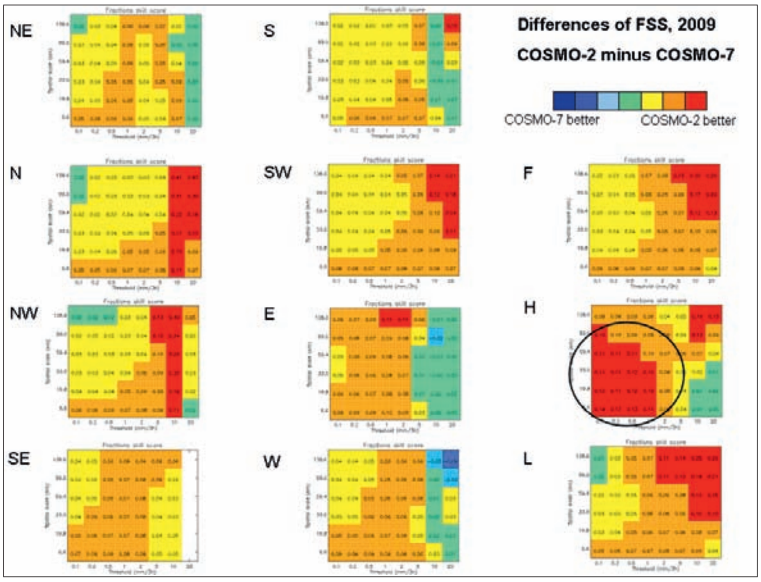
The stratified verification of the model was performed for precipitation by comparing model data and radar accumulated data. The data set has been chosen as follows:

For the COSMO data, the 3 hour accumulated precipitation for the lead-times +04h and +07h: 00z and 12z runs for COSMO7 and the 00z, 03z, 06z, 09z, 12z, 15z, 18z, 21z runs for COSMO2 have been used.

Three hour accumulation data from the Swiss Composite radar (NASS) have been used as precipitation observations. In the case of missing data, the whole day was omitted (26 days).

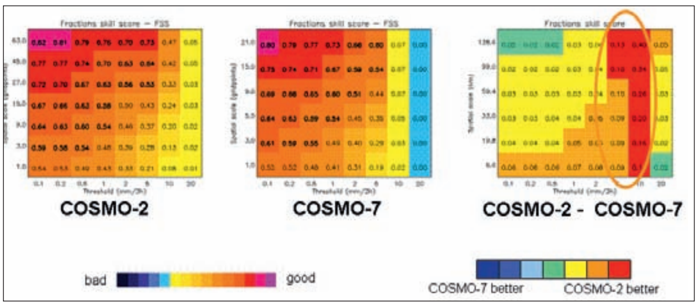
Fig 3 shows the result of the stratified verification for the year 2009. The difference between COSMO7 and COSMO2 has been displayed in order to show in which situation, for which threshold or for which scale one model is better than another. One grid point of COSMO7 corresponds to three grid points of COSMO2. Focusing on high pressure situations (H), COSMO2 is clearly better than COSMO7 especially for low threshold and small scales. The details for a north-westerly situation (fig 4) show that for the two models, the skill is better for low threshold and large scale, but the difference between the two shows that COSMO2 is better for the 10 mm/3h threshold.

Switzerland being a mountainous country, it is also interesting to know spatially where a model is good and in which situation. Fig 5 shows for one year, the average distribution of precipitation predicted by the model (COSMO7) for a given situation. By comparing with the radar data reported on the same grid, the differences point out the area where the skills of the model are the least accurate.



▲ *Figure 3*

Results of the stratification verification for 2009



▶ *Figure 4*

Example of verification for a northwesterly situation with the scores of COSMO7 and COSMO2 and the difference between the two

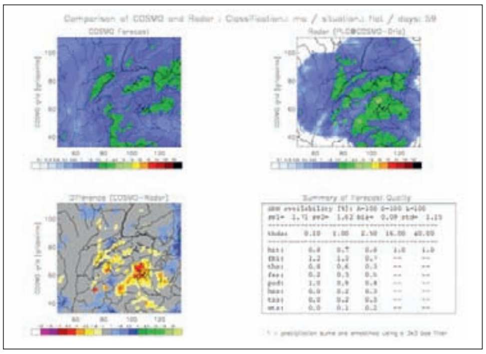


Figure 5

Radar verification of the COSMO 7 model for 2009 and for flat pressure patterns

Guidelines

Regularly, modellers issue fact sheets in which they describe the weaknesses and the strengths of the models for the last season. They also issue guidelines in order to help forecasters to use the model appropriately. Although the reading is very enlightening, many soft links are included in the text, and it is also very detailed. The forecaster has to search for the information. During a shift, this information should be easy to access.

Following a meeting of COST in Geneva, June 2010, a list of recommendations concerning the guidelines was issued:

- The guidelines should be self-contained (without links).
- They could look like a manual, for example for the use of a parameter or for the treatment of a specific situation.
- A 'light' version could be at the disposal of the forecaster on duty (usually under time constraints), while a longer version could be studied offline.
- This longer version could be used as an education tool for newcomers.
- The shorter version could be implemented as a seasonal factsheet.
- The seasonal factsheets should include (if possible) the expected changes of the current model version with respect to the version which was running the previous season.
- Generally speaking the guidelines should be short, attractive and meaningful.

The forecaster should also be part of the process by:

- Providing forecaster feedback organized either by mailbox, a forum or regular discussions.
- At the end of each season a debriefing could be organized and a synopsis written. This could form a good base for the following corresponding season.

Some suggestions that would be an aid to forecasters

Day to day verification

At the end of each day, plots similar to fig 5 could be produced for rainfall and sunshine comparing the different runs of the models with respective rainfall accumulated data from the radar, and the sunshine data derived from satellites.

Synthetic map

The map (fig 6) represents another way to summarize the skill of a model especially in mountainous regions. The patchwork of colors show the climatological regions, the letters the direction of the flow and the + or – signs the indication whether the model has over or under estimated the rainfall for the given situation. A similar chart could be produced for sunshine or other parameters.

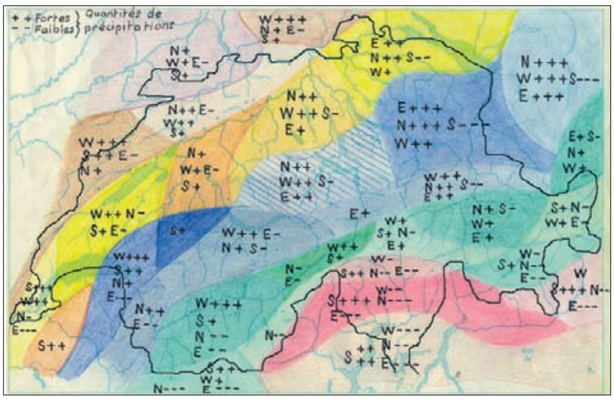


Figure 6

Synthetic map of verification, details in text

Advice

To avoid having to extract relevant information from amongst comprehensive guidelines, a short text could be inserted next to each forecast parameter.

This information would describe the skill of the model for the specific parameter, the present weather situation and the actual season. For example: “northerly flow, the model tends to overestimate precipitation along the northerly side of the Alps”.

Conclusions

With the multiplicity of models and their fast-growing evolution, it is sometimes difficult for the forecaster to have an objective opinion regarding quality and choice. A good or a bad experience concerning the model can influence one’s choice (Gaia, 2007).

Verification issued by the modelers are very often not precise enough to be used as guidelines. The verification is carried out to analyze the general skill of the model over a large area or a very long time. The crucial question for the forecaster is this: where and under which conditions the chosen model is most effective, and which corrections need to be applied? The verification is not easy to access especially when under stressful shift conditions. Updates of models are also frequent; minor updates occur a few times a year often confusing the user. For instance a model could, over a short interval, overestimate and/or underestimate the convection.

The forecaster is also interested in how different models evolve synoptic patterns. For instance, are there some models which are better than others in forecasting the end of an omega block? Do some models produce more cut-offs than others?

In this article some verification ideas have been suggested:

• A day to day verification, in order to see the skill of the model at a glance whilst the recollection of the situation is still fresh.

• The guidance should be easily accessed, such as short text on the side of the browser.

• A synthetic map should summarize the skill of a parameter by region and by season.

References

Elbert E. 2008. Fuzzy verification of high-resolution gridded forecast: e review and proposes framework. Meteorol. Appl. 15: 51-64.

Gaia M. 2008. The human side of weather forecasting. The European forecaster 13: 17-20.

Roberts NM, Lean HW. 2008. Scale selective verification rainfall accumulation from high-resolution forecasts of convective events. Monthly Weather Review 136: 78-97.

Weusthoff T., Ament F., Arpagaus M., Rotach M. 2010, Assessing the benefit of convection-permitting models by neighborhood verification : Examples from MAP D-PHASE. 138: 3418:3433

Zepeda-Arce J, Foufoula-Georgiou E, Droegemeier KK. 2000. Space-time rainfall organization and its role in validating quantitative precipitation forecasts. Journal of Geophysical Reasearch. 105 (D8):10,129-10,146.