Australia's National Science Agency



How can we check the reliability of ensemble flood forecasts?

James Bennett & David Robertson | 19 Nov 2020



The problem



- Low streamflows are not like floods!
- How to verify the bits we want?



The problem



- Low streamflows are not like floods!
- How to verify the bits we want?



Some context: Aus forecasting services

Flood forecasting



7-Day Ensemble service





occur"

Why care about reliability?

Practical

- Forecast probabilities reflect flood risk
- Uncertainty can be propagated downstream
- Outputs can be used directly in decision models

Theoretical

- "Maximise sharpness subject to reliability" *Gneiting et al.*
- No hedging needed





Reliability and forecast selection

 Overall reliability does not guarantee conditional reliability (Hamill 2001)

'Observation-based' selection

- Bellier et al. -> Can't measure reliability
- Lerch et al. -> CRPS not strictly proper

'Forecast-based' selection

- Reliability is measurable
- CRPS strictly proper
- False alarms measurable (!)

Forecaster's Dilem Forecast Evaluation

Sebastian Lerch, Thordis L. Thoraring

Abstruct. In public disc

cally focuses on the predict

ever, the restriction of conv of extreme observations has to discredit skillful forecasts

erating process is low. Con

theoretical assumptions of confronting forecasters with

probabilistic forecasts, pro-

lecision-theoretically just emphasis on extreme eve iments and a real data stu

gross domestic product (caster's dilemma along wi

Key words and phrases:

ratio test, Neyman-Pearso

forecast, proper weighted s

NOTE Interpretation of Rank has sheen

The chaotic nature of the atmosphere 1969, 1982) ensures that errors will grou-ministic numerical weather forecast, even ing that forecast no better than climate ropriate, then, to envision the goa her prediction as providing inform tive likelihood of possible weather sce tical way of doing this is through ease F), whereby a set of numerical foreca in different initial conditions (Toth an 997: Molteni et al. 1995: Houtekamer et rkamer and Lefaivre 1997), different i al perturbations (e.g., Stensrud et a et al. 1999), different models (Evans et mann 2000; Richardson 2000), and/or i fixed fields and constants (Houtekamer et tekamer and Lefaivre 1997). The ensu

The National Center for Atmospheric Reserve

Universitatsplatz 1, 39100 Bozen-Bolzano, Baly francescoratvazzolor@gmail.com). Tilmann Gnev Group Leader, Heidelberg Institute for Theoret. (HITS), and Professor of Computational Statis for Stochastics, Karlsruhe Institute of Technolog gGmbH, Schloss-Wolfsbrannenweg 35, 69118 H. Germany (e-mail: tilmann.gneiting@h-its.org).

Sebastian Lerch is PostDoc, Heidelberg Institu

Theoretical Studies (HITS), and Institute for Sto

Karlsruhe Institute of Technology, HITS gGmbH,

Schloss-Wolfsbrunnenweg 35, 69118 Heidelber

(e-mail: sebastian.lerch@h-its.org). Thordis L.

Thorarinsdottir is Chief Research Scientist, Nor

Computing Center, P.O. Box 114, Blindern, 031

Norway (e-mail: thordiv@nr.no). Francesco Rav

Associate Professor, Free University of Bogen/Bo

The focus of this note is on one of these verification tools, the rank histogram. The rank histogram was de-Dr. Thomas M. Hamill, NOAA-RCDC1, 325 Broadway, Boul-

er, CO \$0309-0216. © 2001 American Meteorological Society

Probabilistic forecasts are nowadays widely used in

the meteorological community, since they provide a

useful estimate of the predictive uncertainty. In an op-

erational context, these forecasts are generally in the

form of ensembles representing possible scenarios. De-

assessing the quality of the forecasts by comparing them

CONTRACTOR NO. 201

1. Introduction

Sample Stratification in Verification of Ensemble Forecasts of Continuous Scalar Variables: Potential Benefits and Pitfalls

BELLIER ET AL

JOSEPH BELLIER AND ISABELLA ZIN Université Grenoble Alpes, Grenoble INP, CNRS, IRD, IGE, Grenoble, France

GUILLAUME BONTRON Compagnie Nationale du Rhône, Lyon, France

(Manuscript received 29 December 2006, in final form 23 May 2017)

ABSTRACT

In the verification field, stratification is the process of dividing the sample of forecast-In the ventration field stratification is the process of dividing the sample of forecast-observation pars raise quark desargements subsets, in order to learn more on hose forecasts below used repetitive conditions. A generate funsaments for artification to presented to the case of rememble forecasts of continuous scalar variables. Distinction is made between forecasts based, observation based, and external-based stratification, variables. visibles. Distinction is made breaves finanzia based, diversities based, and extensiza based attraifications depending on the citetricometismous ratio-fip-should be stratified. The fitter based of the strategiest and the strategiest are breaked based on the strategiest and the strategiest are breaked based on the strategiest are breaked based based based based on the strategiest and the strategiest are breaked based b observation based stratification leads to significantly souffat histogram for calibrated forecasts. Neverthermost, an previous studies have warned, statistical artifacts created by a forecast-based stratification may all occur, thus a graphical test to describe them is suggested. To finitatate potential insights about forecast behaviorthat can be placed from stratification, a sumerical example with two different datasets of mean areal behaviorthat can be also for the strategisted. precipitation forecasts is presented.

a large number of forecast-observation pairs in the verification sample for being statistically robust. To help increase the sample size, various forecasts may be pooled together (in the same sample), for example, for different locations, for various ranges of predictands, or from different model versions. However, computing a verification measure over an inhomogeneous sample faces the risk of having different forecast behaviors that spite progress in the verification field since their emeraverage out. Stratification, as the process of partitioning gence, the complexity of their behavior still represents a the verification sample into different subsets, aims a great challenge for verification practitioners (Casati et al. 2008). In few words, verification is the action of

conditioning the verification measure to specific conditions, so as to minimize this risk and lead to more in sightful verification case studies. It is difficult to trace back the origin

to their corresponding observations (Jolliffe and Stephenson 2003). Since a complete picture of the stratification, since the concept has probably emerged forecast quality cannot be obtained from a single measoon after first meteorological forecasts were verified. sure, different verification measures have been pro-Indeed, authors very often present performance meaposed, which evaluate different attributes (i.e., aspects) sures for different locations or seasons, which is an imof the forecast quality (Murphy 1973). All measures, plicit way of stratifying the complete verification sample though, have in common the fact that they require Such an approach aims at making measures of forecast skill independent from the climatological frequency of events that have to be verified, which varies both in

Corresponding nuthor: Isabella Zin, isabella zin@univ-grenoble-

0.2015 American Mercenslepical Society. For information regarding result of this content and general copyright information, someth the AMS Copyright Didge (new antition, cog/T/DMR end. (creme)). DOI: 10.1175/MWR-D-16-0487.1



Forecast-based selection



- Heightened probability of flood is forecast -> forecast selected
- Analogous to a 'flood watch'



Verifying forecast selection





Reliability – probability integral transform







Also need

- Sharpness (AWPI)
- Accuracy (CRPS)



Case study: 2 forecasting methods



Verifying forecast selection



- Flood threshold exceeded in 26 forecasts
- ERRIS-A up to ~2000 forecasts selected
- ERRIS-B up to ~120 forecasts selected



Reliability of flood forecasts





Accuracy and sharpness of flood forecasts

Reliability of flood peak magnitude and timing









Reliability of flood peak magnitude and timing







Recommendations

- Forecast-based selection crucial to measure reliability
 - (and false alarms!)
 - But... forecast-based selection means comparing different numbers of forecasts of different events
- Obs-based selection still useful for communicating performance



References

Forecast verification

Bellier J, Zin I, Bontron G. 2017. Sample Stratification in Verification of Ensemble Forecasts of Continuous Scalar Variables: Potential Benefits and Pitfalls. Monthly Weather Review 145: 3529-3544. DOI: 10.1175/mwr-d-16-0487.1

Gneiting T, Balabdaoui F, Raftery AE. 2007. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69: 243-268. DOI: 10.1111/j.1467-9868.2007.00587.x.

Hamill TM. 2001. Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *Monthly Weather Review* **129**: 550-560. DOI: 10.1175/1520-0493(2001)129<0550:Iorhfv>2.0.Co;2.

Lerch S, Thorarinsdottir TL, Ravazzolo F, Gneiting T. 2017. Forecaster's Dilemma: Extreme Events and Forecast Evaluation. Statist. Sci. 32: 106-127. DOI: 10.1214/16-sts588.

7-day ensemble streamflow forecasting service

Bennett JC, Robertson DE, Shrestha DL, Wang QJ, Enever D, Hapuarachchi P, Tuteja NK. 2014. A system for continuous hydrological ensemble forecasting (SCHEF) to lead times of 9 days. *Journal of Hydrology* **519**: 2832-2846. DOI: 10.1016/j.jhydrol.2014.08.010.

Bennett JC, Robertson DE, Wang QJ, Li M, Perraud J-M. (In review). Propagating reliable estimates of hydrological forecast uncertainty to many lead times. Submitted to Journal of Hydrology.

Li M, Robertson DE, Wang QJ, Bennett JC, Perraud J-M. 2020. Reliable hourly streamflow forecasting with emphasis on ephemeral rivers. Journal of Hydrology 125739. DOI: 10.1016/j.jhydrol.2020.125739.

Robertson DE, Shrestha DL, Wang QJ. 2013. Post-processing rainfall forecasts from numerical weather prediction models for short-term streamflow forecasting. Hydrology and Earth System Sciences 17: 3587–3603. DOI: 10.5194/hess-17-3587-2013.

Shrestha DL, Robertson DE, Bennett JC, Wang QJ. 2015. Improving precipitation forecasts by generating ensembles through postprocessing. Monthly Weather Review 143: 3642-3663. DOI: 10.1175/mwr-d-14-00329.1.





Thank you

Land & Water James Bennett Senior Research Scientist

+61 2 9545 2462 james.bennett@csiro.au https://people.csiro.au/B/J/James-Bennett

Australia's National Science Agency



Reliability and forecast selection



Synthetic example

- Obs and forecasts drawn from the same normal distributions
- 'Flood threshold' based on 99.5% quantile of 'observations'



Some context: Aus forecasting services





