Hydrol. Earth Syst. Sci. Discuss., 6, C1294–C1297, 2009
www.hydrol-earth-syst-sci-discuss.net/6/C1294/2009/

Hydrology and Earth System Sciences Discussions

Interactive comment on "Dynamically vs. empirically downscaled medium-range precipitation forecasts" *by* G. Bürger

Anonymous Referee #2

Received and published: 21 June 2009

Review of "Dynamically vs. empirically downscaled medium-range precipitation forecasts" by G. Bürger

General comments

This paper compares the accuracy of catchment-scale precipitation forecasts from two downscaled NWP models: the DWD GME model dynamically downscaled using the LM, and the ECMWF IFS model empirically downscaled using EOFs. The period of comparison was 2002-2005. For the three catchments investigated the empirically downscaled forecasts outperformed the dynamically downscaled forecasts according to the Gilbert Skill Score using thresholds of 0.1 mm and the 95th and 99th percentiles. In the discussion the author suggests that the improvement of the IFS/EDS over the

GME/LM is due primarily to the method of downscaling.

It bothers me a bit that the comparison is not a clean one, in that the parent model (GME or ECMWF) differs between the two experiments. Although it may not be possible, it would have been better to use the same parent model with two different downscaling approaches. It is generally acknowledged that the ECMWF IFS model is the best overall performer for atmospheric prediction over Europe (in fact the world) and I would be surprised if the GME made better predictions over Germany than the IFS did during 2002-2005. You should contact Uli Damrath or Detlev Majewski to get their comparisons of NWP model performance over Germany that they do for WGNE. The point I am trying to make is that the improvement of IFS/EDS over GME/LM is partly related to the use of a better parent model. You argue on p.3528 that this difference is not likely to be much, but I suspect it may be more than you think. The COSMO-SREPS uses different combinations of parent models with child models to make a short-range ensemble – this work shows that the parent model matters a lot, at least for dynamical downscaling.

Regarding binary verification scores, the GSS by itself tells only part of the story. It is customary to also show the frequency bias index (FBI), since for forecasts that are horizontally offset from the observations, one can increase the GSS by over-predicting the rain area (Baldwin and Kain, WAF 2006). The contingency tables indicate that for at least the extreme amounts the FBI for IFS/EDS is >1 and for GME/LM it is <1. It would be useful to see a version of Fig. 5 for FBI. Note that over-prediction may be a reasonable strategy if the loss associated with missing a heavy rain event is much greater than the cost of false alarms.

Specific comments

p.3518 line 16: change "most" to "much", as a lot of the damage is unavoidable even when the event is predicted.

p.3519, lines 10-11: These papers compared a large number of NWP models for their

ability to accurately predict precipitation. They did not specifically address the issue of downscaling, although regional models (which would be dynamically downscaled from global models) were included in the comparison.

p.3519, line 23: I agree with your use of the word "I" and wish more authors would be this clear in their writing (but admit I have not been courageous enough to use it myself yet. . .).

p.3522, line 8: 81 EOFs seems like a very large number. I am less familiar with use of EOFs in NWP downscaling than for climate variability studies, which typically use many fewer EOFs (less than 10). Are you sure you are not over-fitting? In line 19 you have n=85 – why not n=81?

p.3525, line 6 and Fig. 3: It would be easier to see the under-prediction of strong events by GME/LM using a pdf rather than a cdf. This maybe does not matter too much since the contingency tables also show this effect (in which case Fig. 3 may be unnecessary).

p.3525, line 9: it would be useful to know what the values of the upper 5% and 1% quantiles were in mm/d.

p.3525, line 18: It would be useful to mention that the GSS is the same as the equitable threat score (ETS), which is probably more familiar to most readers.

p.3526: The numbers C=10 k and L=100 k appears to be arbitrary – are they somehow reasonable according to some scenario. Since different users have different cost/loss ratios it would make more sense to compute the expected expense e for a range of C/L between 0 and 1.

p.3527, end of section 3: The GSS=0.54 is very likely due to the effect of a small sample. In fact, the Q99 statistics have a very small number of samples (16 for Alb, fewer for Upper Danube), so the true value of the statistic has a large amount of uncertainty. This could be easily estimated using a bootstrap approach (Wilks 1995). This would

be a good idea for all of your GSS values, but especially the ones for rare events.

p.3528, line 26: change "wishful" to "desirable". All forecasts should ideally be verified; even if that is not possible you want your forecasts to be as accurate as possible. Recommend taking out the phrase, "at least if they are going to be verified."

p.3529, line 1: Were the EOFs chosen by cross-checking within the training data (1997-2001), or also including the test data (2002-2005)? It should be the former, but you should clarify this.

Technical comments

p.3521, line 3: change "nested to" to "nested in"

p.3522, line 1: add degree symbols to 1x1

p.3522, lines 16, 19: You use the symbol "n" to mean two different things. One of them should be changed.

p.3522, line 24: EDS has not yet been defined.

p.3525, line 12: change "less" to "fewer"

p.3528, line 21: don't need "probably".

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., 6, 3517, 2009.