Hydrometeorological Short-Range Ensemble Forecasts in Complex Terrain. Part I: Meteorological Evaluation

McCollor, Stull

August 2008

1. Objective:

Which is better for complex terrain numerical precipitation forecasts?

1. Using lower-resolution members of a nested suite to increase ensemble size
2. Using high-resolution members only.

Investigate the benefit of including multiple-resolution (Limited Area Model) LAMs to increase ensemble size

1. Introduction

* Multi-model system is used instead of multi-initial conditions. So, they are used in the research.
* High-resolution grids capture the orographic and variable surface flux components.
* Coarser resolution, many members ensemble can outperform finer resolution, few members ensemble
* Q. Synoptic-scale errors originating from analysis uncertainty?
* Q. multi-resolution nested systems?

1. Methodology
   1. *Location and Observation Data*

* Numerical results are verified with gridded analyses/point observations
* Better to verify forecast with observations rather than model analyses in complex terrain
* One constraint: observational data access was only for 75% of the watershed areas. There are no measuring stations above the height of 2000 m
* High elevation stations do not necessarily receive most rainfall because they could be under shadow of another hill in front of them or they are far away from the shoreline
  1. *Numerical models and study dates*
* GDCFDC (our group) runs real-time suite of three independent nested limited-area high-resolution mesoscale models.
* Outer nests (resol. 108 km) initialized using NCEP NAM model at 90 km grid spacing.
* Q. Mesoscale Compressible Community model (MC2): fully compressible, semi-implicit, semi-Lagrangian, non-hydrostatic mesoscale model. ??
* MC2 is run with horizontal grid spacing of 108, 36, 12, 4 and 2 km, with 1 way nesting.
* MM5 same grid, but with 2-way nesting.
* WRF: grid spacing of 108, 36 and 12 km, with 2-way nesting
* LAM archive across two consecutive wet-seasons. The value is interpolated to the weather station locations.
* Multi-model SREF outperforms varied-model SREF in representing model uncertainty.
* 24 h time window is used. 1. Because water managers usually ask for daily forecasts. 2. There is greater predictability of the model
* A good Ensemble Prediction System (EPS) should have enough members to have data whose probability distribution is, ideally, indistinguishable from the observations. Which means enough spread of observations
* Larger domain ensembles produce greater spread than the smaller-domain ensembles. Q. What does domain mean? More input/larger coverage?
* MC2, MM5 and WRF models are initialized at 0000 UTC and run for 60 hrs.
* The models run with dry conditions as initial conditions. So, to compensate for the bias, a day is taken from 1200 to 3600.
* 2 km of MC2 and MM5 were susceptible to missing forecast days. Why?
* Different configurations were used: i) all 11 MC2 all resolutions + MM5 all resolutions + WRF all resolutions ii) 8 ensembles excluding the 108 km resolution iii) 6 ensembles with mid resolutions (36 and 12 k only) iv) 5 members (12 and 4 k only)

1. Verification Procedures

* Degree of Mass Balance (DMB)
* Mean Error (ME)
* Mean Absolute Error (MAE)
* Mean Square Error (MSE)
* Root Mean Square Error (RMSE)
* Pearson correlation (r)
* Linear Error in Probability Space (LEPS). ?
* Brier Skill Score (BSS) incl. relative reliability and resolution term
* Relative Operating Characteristic (ROC) from hit rate and false alarm rate
* Rank Histograms
* There is no universal best way to verify the ensemble forecasts. It depends on the needs of an individual user.

1. Results:

a. Error performance of individual ensemble members

* Precipitation forecasts and observations should have similar distributions
* Precipitation forecasts are non-normal distribution
* Analysis should be both resistant (not influenced by outliers) and robust (independent of underlying distribution). Check it using box and whisker, and LEPS.
* Finding: Distribution of high-resolution models is more similar to the distribution of the observations.
* LEPS compares the cumulative distributions of each model ensemble member with the corresponding distribution of observations.
* LEPS score of 1 = best. 12 and 4 km models perform the best. 108 and 36 km models tend to under-forecast (i.e. DMB < 1). Even the highest resolution models tend to under-forecast in day-2.
* Mean error is drastically better for the high resolution forecasts with negligible for 4 km grid. Day 2 forecasts have higher error than day 1
* Reasons for systematic bias:
  + Model physics,
  + Initial conditions,
  + Imperfect boundary conditions,
  + Insufficient resolution to cover the complex terrain,
  + Point observation vs. Grid forecasts that represent a volume,
  + Model elevation is different than the station elevation
* Systematic Bias can be corrected using postprocessing
* Random error is included in mean error but not in mean absolute error. Because it has positive variance and zero mean. Random error is not resolution dependent and the only way to improve it is by improving the model.
* Ensemble averages minimizes random error.
* Better mass balance and improved mean error is found in finer-resolution models.

b. Brier skill score: Resolution and reliability

* compares BSS of ensembles of different sizes using a single metric with resolution and reliability components.
* Best model: resolution = 1 & reliability = 0
* Very fine resolution 5-member SREF performs the worst for both days and all threshold precipitation values.
* Reliability diagrams: ensemble with all 11 members outperformed the very high-resolution members.
* Reliability diagram = plot of observation probability vs. forecast probability.
* Q. Sharpness? In text: ‘the relative frequency of occurrence of the forecast probabilities.’

c. ROC curves

* graph of hit rate vs false alarm rate.
* Area under ROC curve = 1 for perfect discrimination
* ROC levels: >0.4 reasonable & useful, > 0.6 good and > 0.8 excellent
* General trend: increasing forecast threshold precipitation= better ROC skill score, except for Day 2 forecasts where it worsens during the 25 to 50 mm transition.
* Reason: as the days increase the forecast tends to be weak in predicting the rare events.

d. Equal Likelihood

* Rank histograms provide a measure of reliability
* Ideal case: forecast distribution matches with the observation distributions. Rank Histogram = flat
* Real life: Rank Histogram = U-shaped
* U-shaped reasons = model bias or lack of ensemble spread
* Delta >> 1: system does not reflect equal likelihood of ensemble member
* All 11 ensemble is the best performer, but it still does not give a flat histogram.
* Model diversity is important for surface sensible weather elements (wind speed and temperature)

1. Conclusion:

Limitation of very high resolution: inadequate subgrid-scale parameterization

High resolution = improved individual error and correlation metrics.

Low resolution = improved resolution and reliability components as per the Brier skill score.

Day 2 forecasts are worse than day 1, but they are still reliable with at least good ROC curve standing.