30-day Forecast COMPS Experiments – What and Why?

This document is a summary of the information that can be found in quarterly progress reports to BC Hydro. For each forecast type (hourly temperature and precipitation and daily TMin/TMax and precipitation), it outlines what strategies were tried, what worked, what didn't, and why certain strategies were kept or discarded. For further details (e.g., graphs), see past quarterly progress reports.

Unless otherwise stated, all of the bias-correction, probability modelling, and probabilistic calibration schemes used here were run using the COMPS framework. As you will see, the adaptive training method that COMPS uses (in contrast with a moving window training method), while being proven for short- and medium-range forecasts of many variables, may not be entirely suitable for long-range forecasting (i.e., beyond 15-16 days).

General Ensemble Information

Most testing was done on a 102-member ensemble containing the following models:

- RDPS
    - 48-hour (2-day) forecast
    - Hourly forecasts use the 12 UTC RDPS run (which is most up-to-date)
    - Daily forecasts use the 06 UTC RDPS run (which is a bit older, but because TMin/TMax/P24 forecasts are calculated on 06 – 06 UTC windows, the 12 UTC gives missing Day 1 forecasts)
- GDPS (12 UTC)
    - 240-hour (9-day) forecast
- GFS (6 UTC)
    - 256-hour (10-day) forecast
- 42-member NAEFS ensemble (00 UTC)
    - 384-hour (16-day) forecast
- 34-member UBC 00 UTC SREF ensemble
    - Various lengths up to 7 days
- Climatology
    - 30-day guidance
- 20 Historic Traces (1996 – 2015)
    - 30-day guidance
- CFS 00 UTC Control Member
    - 30-day guidance
    - CFS generates forecasts over the globe with lead-times of several months, but the archive we maintain is sliced and diced to cover Canada with a lead-time of 31 days to save storage
- CFS 06 UTC Control Member
    - As for 00 UTC CFS

The operational forecasts also include the following models, which were tested briefly due to constraints on hindcast availability and COMPS parameter spin-up time:

- 3 additional CFS 00 UTC members
- 3 additional CFS 06 UTC members
- 4 CFS 12 UTC members
- 4 CFS 18 UTC members

The operational ensemble thus has 116 members with varying forecast horizons.

For more information on the individual processing of these models, see the *30-day_superensPrep.docx* document or the README files in model@levant:~/HydroMet30/scripts/.

Hourly Temperature

- COMPS schemes selected for operational use:
    - Bias correction of individual ensemble members: MeanBias correction scheme with e-folding time of τ=30
    - Ensemble deterministic forecast: Mean
    - Probability modelling: Method of moments ("MM2"), normal distribution with spread parameter = *a* + *b***ensVar*, where *a* and *b* are regression parameters trained based on past values of *ensVar* (ensemble variance) and the errors of the ensemble mean (i.e., they describe the spread-skill relationship); fitted to full ensemble mean
- Other schemes tested:
    - Bias correction using Kalman Filter – not appropriate for correcting historic traces because the nature of their biases does not match the assumptions of the KF (according to Henryk); tested MeanBias scheme with 30-day e-folding time, which was found to be better for all ensemble members, so all were switched to MeanBias (Sept 2017 progress report)
    - MM0 (probability spread parameter = *a*) and MM1 (probability spread parameter = *a* **ensVar*) probability models – MM1 ensemble spread drops dramatically when NAEFS ends after Day 16, tended to be underdispersive; MM0 spread evolves nice and smoothly; otherwise very similar to MM2 results, but MM2 tended to have slightly lower ignorance so it was selected (Dec 2017 progress report)
    - Ensembles composed of n-best ensemble members (based on recent MAE) – reduced ensembles beneficial only in short-range due to exclusion of historic traces, but difference is small; full ensemble tends to be better at longer lead-times; probability forecasts with reduced ensemble struggle because they are trying to fit parameters to an ensemble whose membership could be changing on a daily basis (result is spiky probability spread in meteograms); for this reason, n-best ensemble configurations were not tested for other forecast variables

- Other “restricted” ensembles (Dec 2016 progress report) – excluding climatological/historic members at shorter lead-times helps in interior only (away from moderating influence of ocean where deviations from climatology are small), but this project is for coastal region
- Weighted ensemble means (done in R; Dec 2016 progress repor) – extremely small improvement in MAE for coastal stations, some improvement in interior, not worth extra computational expense
- BMA – originally planned to test, but tossed that idea because weighted means didn’t do well and experiments on hydrologic forecasts (Appendix C of my thesis) showed that because there are so many parameters to fit, lengthy training periods are required and weights are extremely slow to respond to changes in model performance (at least in COMPS)
- Calibration – because characteristics of probability forecasts change seasonally, the calibrator will tend to use the wrong correction for much of a season (because it has been trained on data with different calibration requirements) if long e-folding times are used; ignorance increases after calibration, with larger increases for smaller e-folding times, which would tend to not have the seasonal training problem described above (Dec 2017 progress report)

Hourly Precipitation

- COMPS schemes selected for operational use:
  - Bias correction of individual ensemble members: dmbMitDet (regular hourly dmb with mitigation turned on to avoid blowing up bias-correction factors due to missed Tstorm events) and dmb24 (uses Precip24 dmb factors to correct hourlies)
  - Ensemble deterministic forecast: Mean and Median still being tested
  - Probability modelling: MM2 fitted to ensemble mean trained using e-folding time of 30, then calibrated using e-folding time of 60 to transform the normal distribution into a more Gamma shape (“NormCal”). The operational setup will run NormCal with deterministic forecast = ensemble mean AND median, with individual ensemble members corrected using dmbMitDet. Text output sent to BC Hydro will be the version that uses the ensemble mean. If dmb24 is better than dmbMitDet, then the winning probability model will have to be spun up with dmb24 inputs for the next water year.
  - Discrete POP: Consensus (percentage of ensemble members predicting >0 precipitation)
- Other schemes tested:
  - Bias correction – correction of hourly precip with past hourly precip forecasts/observations was problematic for large forecast busts, which can be fairly common for hourly precipitation (e.g., missed thunderstorm); these forecast busts/misses can result in sudden, large inflation of the

bias-correction parameter for a particular forecast hour, which persists for over a week

- Probability modelling: tested various gamma distributions (June 2018 progress report). Gamma fitted using MM1/MM2 results in failure to converge and eventual crash (same for Precip24). Gamma fitted with maximum likelihood is stable but not as good as NormCal, which fits a normal distribution then calibrates it to produce more of a gamma shape.

Daily Temperature (TMin/TMax)

- COMPS schemes selected for operational use:
  - Bias correction of individual ensemble members: MeanBias scheme with e-folding time of 30 for model guidance and 90 for observation-based (climatological/historic) guidance
  - Ensemble deterministic forecast: Mean
  - Probability modelling: MM2 trained with e-folding time of both 30 and 90 (both being run for evaluation in fall 2018; addition of full CFS ensemble changed ensemble characteristics enough that it was not possible to select between the two, so a longer verification period is needed before the winner can be selected)
- Other schemes tested:
  - Different training lengths for bias-correction of long-range, observation-based forecast guidance – poor long-range TMin/TMax reliability suggested that MeanBias with the standard e-folding time of 30 was not appropriate for climatology and historic traces (Dec 2017 progress report); e-folding times of 90 -180 were tested on observation-based guidance, with longer training lengths producing better results for everything except bias; biggest improvement was in jump from e-folding time of 30 to 90, with very small further improvement for longer e-fold, so 90 was retained for further testing in probability forecasts and was found to be better than 30
  - Different training lengths for probability model – e-folding time of 90 produced slightly flatter PIT histograms than e-fold=30, but differences were very small (March 2018 progress report); different bias-correction and probability training lengths were not found to have significant impact on long-range reliability, which is not great; suggests that COMPS is not suitable for long-range forecasting due to adaptive nature of parameter updating; addition of full CFS ensemble changed probability forecast characteristics enough that it was not possible to choose between the two tested schemes
  - Different probability models – (MM0, MM1) were not tested on TMin/TMax because they were not found to be beneficial for hourly temperature, and because MM2 is used for the operational 15-day system developed by Greg West

- Calibration – found to deteriorate all measures of forecast quality (March 2018 progress report)

Daily Precipitation (Precip_24)

- COMPS schemes selected for operational use:
    - Bias correction of individual ensemble members: Degree-of-Mass-Balance (DMB) scheme with e-folding time of 30 for model guidance and 180 for observation-based (climatological/historic) guidance
    - Ensemble deterministic forecast: Mean
    - Probability modelling: MM2 fitted to ensemble mean trained using e-folding time of 30, then calibrated using e-folding time of 60 to transform the normal distribution into a more Gamma shape
    - Discrete POP: Consensus (percentage of ensemble members predicting >0 precipitation)
- Other schemes tested:
    - Different training lengths for bias-correction of long-range, observation-based forecast guidance – some historic/climatological precipitation errors worsened following standard DMB correction with e-folding time of 30 (Dec 2017 progress report); note that this behaviour was not observed for hourly precipitation forecasts; DMB with e-folding time of 180 was best (or the least of all evils, anyway) for observation-based guidance, and standard e-folding time of 30 was still best for model guidance
    - Different probability models and training lengths – other methods produced forecasts with extreme bias (March 2018 progress report has table showing methods tested)
    - Different discrete probability forecast methods: diLogit (logistic regression) was selected as a candidate scheme; probability is of form logit(P) = a + b * m + c * f, where a, b, c are constants, m is the ensemble mean, and f is the fraction of ensemble members lying on the discrete point; this scheme may not be fully coded in COMPS, as it produces the same POP for every lead time on a particular forecast day (June 2018 progress report)

Operational Runs

Handled by comps_driver, which Henryk has set up on the cron for the daily forecasts. He can do the same for the hourlies once those are spun up and ready to go live.

Hindcast Runs

Handled by two scripts run monthly (on the 7th of the month) by run_30d_hindcast:

- run_30d_hourlies {date1} {date2}
- run_30d_dailies {date1}{date2}

With {date1} ~ 2 months ago and {date2} ~ 1 month ago. This hindcast is run to populate the observation and verification fields in the verif files. This doesn't need to be done every month (it could be done for a full year at the end of each water year, for example), but the hourly forecasts can take a very long time to run, so monthly hindcasts are a less painful way to build up the water-year verification. Because the forecast is 30 days long, the most recent forecast initialization date we can fully verify will be just over a month in the past (to account for potential lag-times in getting observations).