NEW STATISTICAL POST-PROCESSING RESULTS WITH THE GLOBAL ENSEMBEL SYSTEM

Bo Cui*, Zoltan Toth, Yuejian Zhu, Dingchen Hou*

Environmental Modeling Center, NCEP/NWS
*SAIC at Environmental Modeling Center, NCEP/NWS

Acknowledgements: Jeff Whitaker, Tom Hamill

STATISTICAL POST-PROCESSING OF ENSEMBLES

- MOTIVATION:
  - First phase of NAEFS to be operationally implemented in 2006
    - Develop and implement a statistical post-processing scheme to reduce the biases in ensemble forecasts w.r. t the verifying analysis fields on the model grid
      - Correct both the 1st and 2nd moments

- LIMITATIONS:
  - Preliminary study based on:
    - 500 mb height and 2m temperature
    - Four seasons (2004)

METHOD / APPLICATION – 1

- Adaptive (Kalman Filter type) Bias-Correction Algorithm

Implementation of decaying averaging for 1st moment bias

decaying averaging mean error = (1-*w*) * prior t.m.e + *w* * (f – a)

For each lead time separately, tme = time mean error

- Application to NCEP Operational Ensemble

- OPR_RAW: NCEP T00Z 10 ensemble forecasts
- OPR_DAV2%: *w* = 2% (most recent ~30 days used)
- OPR_OPT: 31-day centered running mean forecast error is removed, operationally not feasible, used as “optimal” benchmark

METHOD / APPLICATION – 2

- CDC GFS Reforecast Data Set (Hamill & Whitaker)
  - Model: T62L28 MRF, circa 1998
  - Initial states: NCEP Reanalysis
  - Duration: 15 days runs at 00Z from 19781101 to now
  - Ensemble: Breeding, 10 members used from 15
- Bias correction

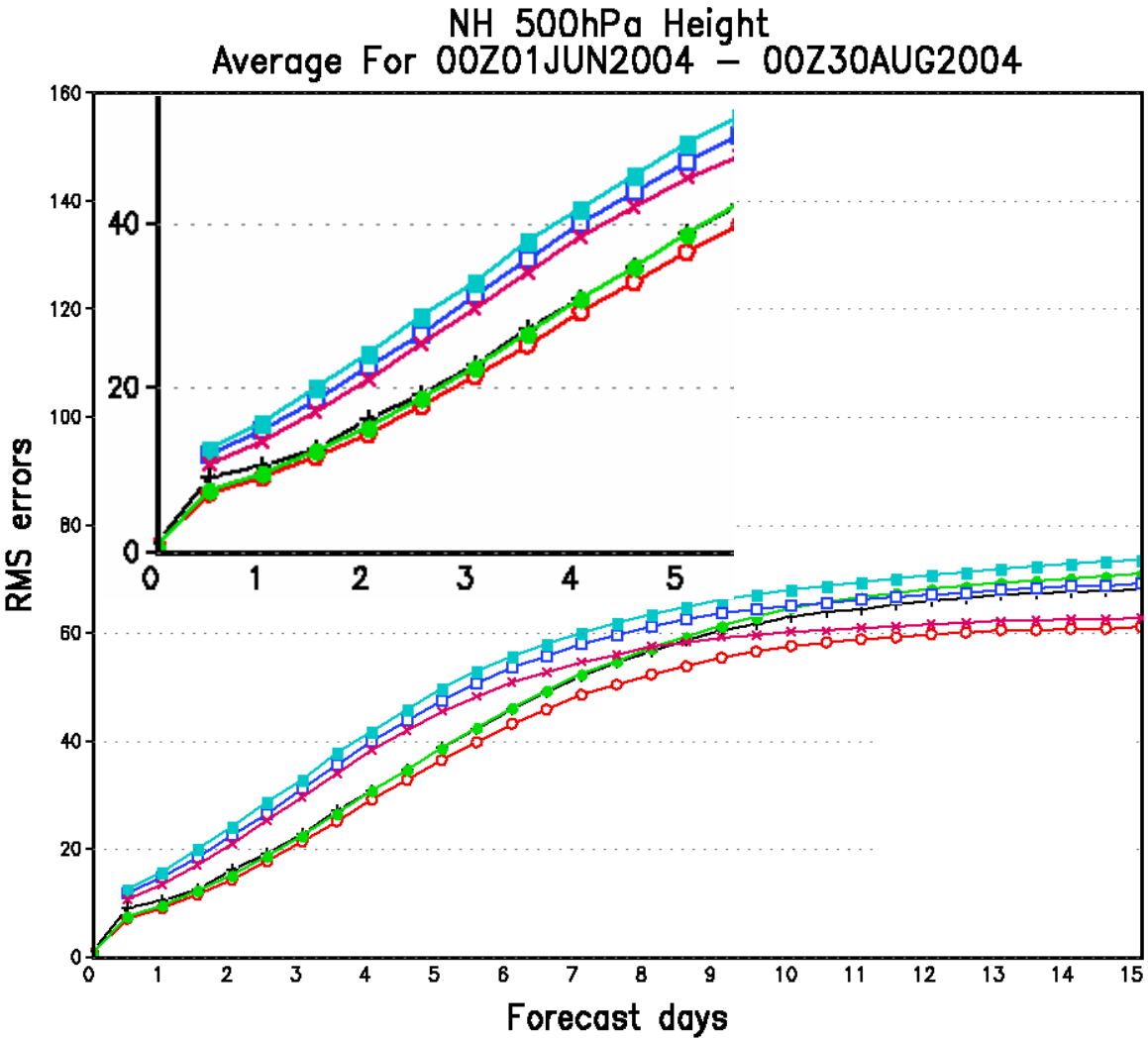
  Climatological (out of sample) mean forecast error (25 yrs) removed (1979-2003, 1st moment)
- Experiments
  - RFC_RAW

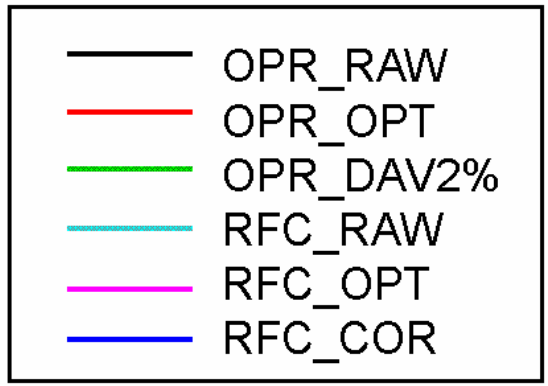
    CDC reforecast ensemble forecasts (no bias correction)
  - RFC_COR

    Calibrated CDC reforecast
  - RFC_OPT

    31-day centered running mean forecast error is removed, operationally not feasible, used as “optimal” benchmark

RMS: 500 mb Height, 2004 Summer
Northern Hemisphere





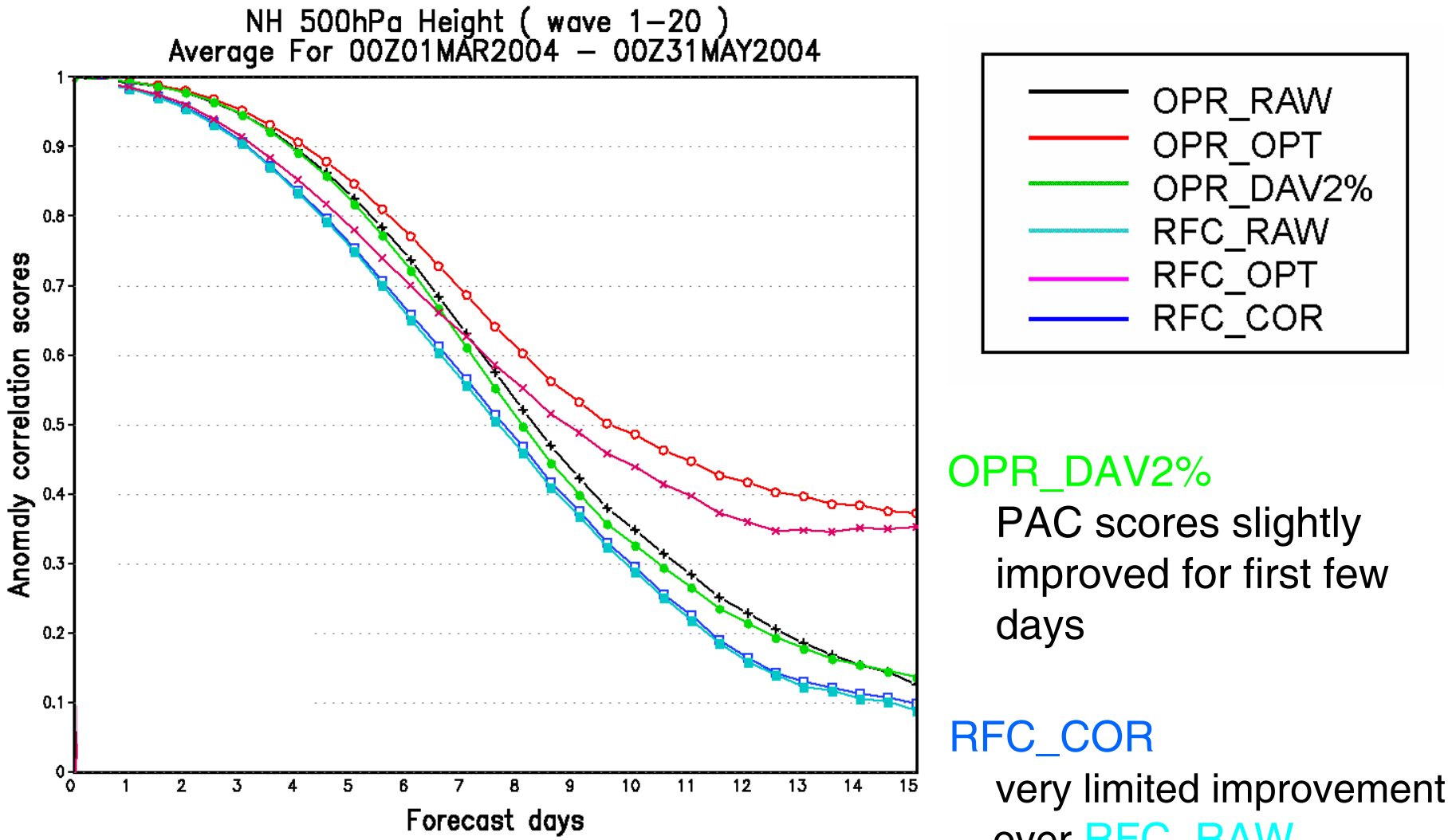
OPR_DAV2%

RMS error reduced for first week

RFC_COR

improvement for all lead times wrt RFC_RAW

PAC: 500 mb Height, 2004 Summer
Northern Hemisphere



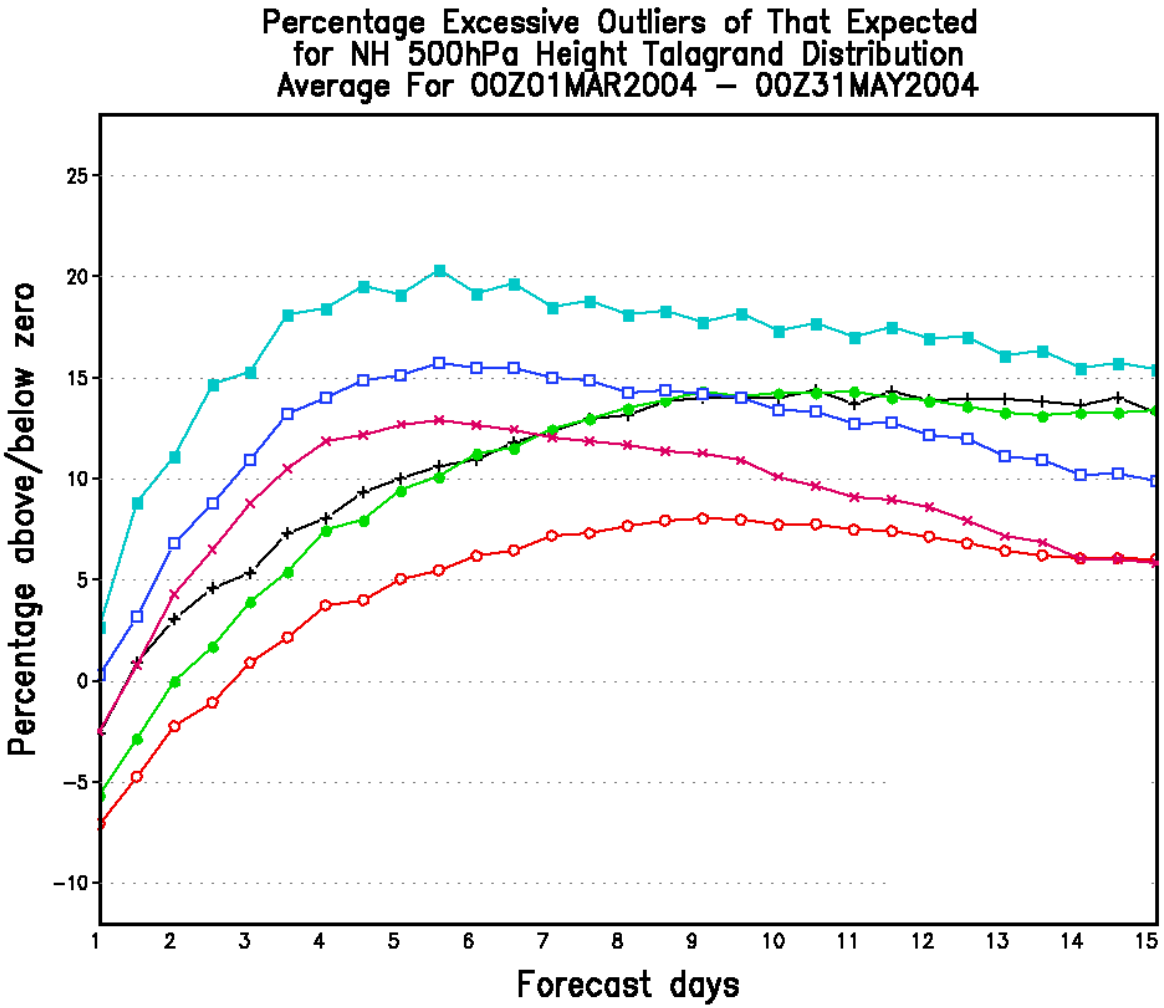
OPR_DAV2%

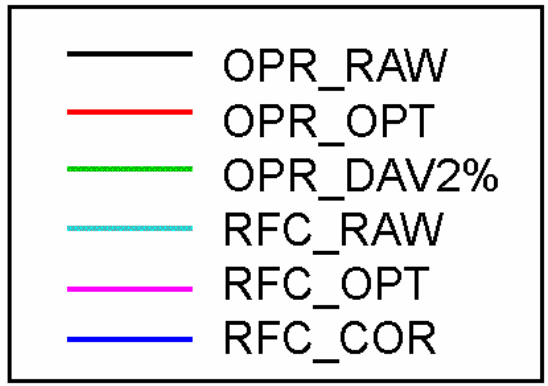
PAC scores slightly improved for first few days

RFC_COR

very limited improvement over RFC_RAW

Excessive Outliers: 500 mb Height, 2004 Summer Northern Hemisphere





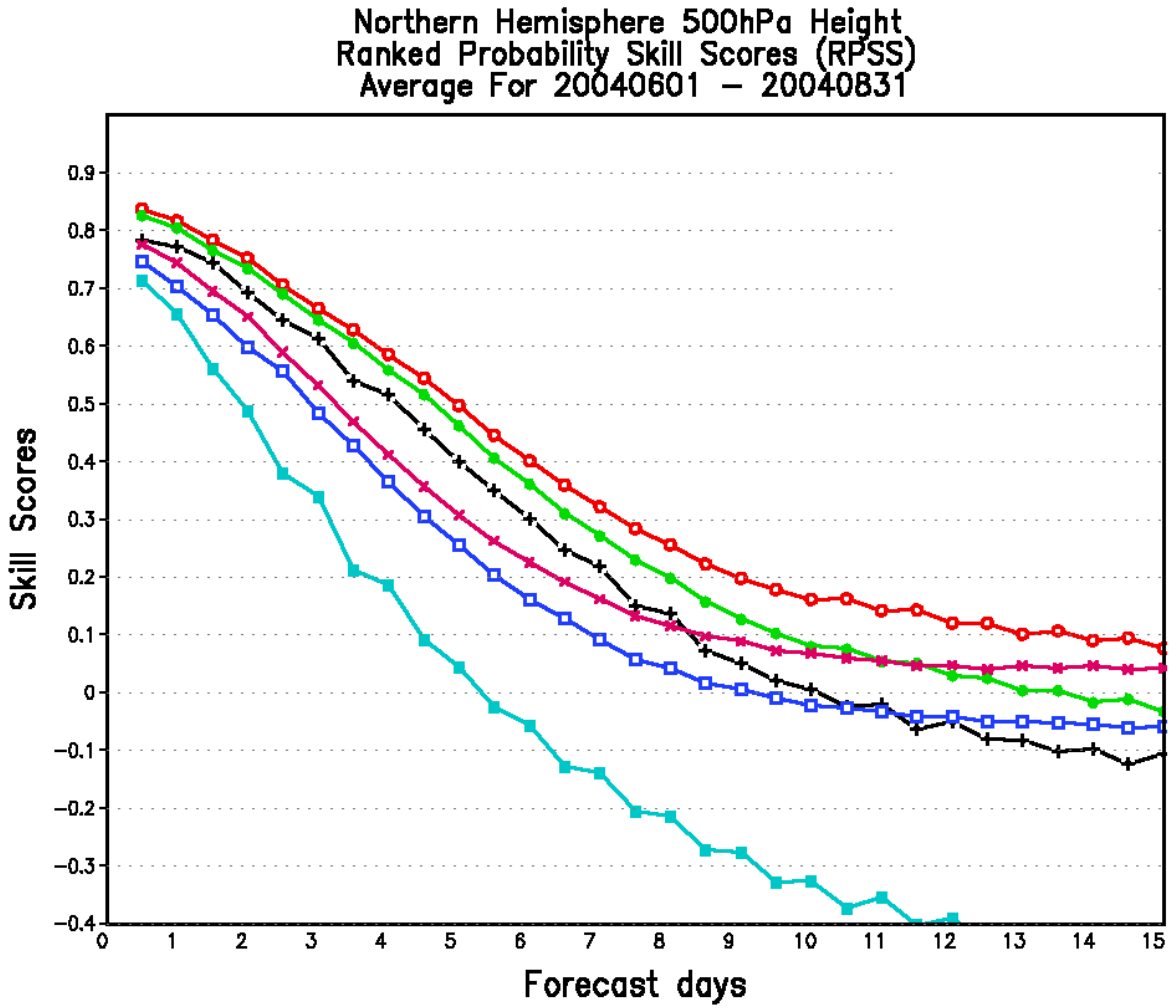
OPR_DAV2%

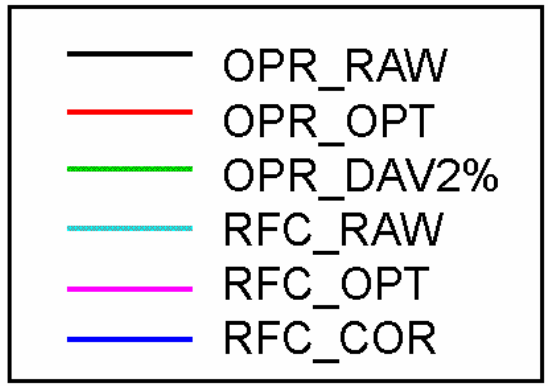
improved performance for up to 5-7 days

RFC_COR

improvement for all lead time vs. RFC_RAW

RPSS: 500 mb Height, 2004 Summer Northern Hemisphere





OPR_DAV2%

RPSS improved for all lead time

RFC_COR

significant improvement for all lead time vs. RFC_RAW

PRELIMINARY RESULTS (1)

1. Decaying averaging (2% weight, ~30-day oper. training data):
   - Short range: Works very well, all measures improved (~Day 5)
   - Week 2: Limited success
     - Degrades ensemble mean (rms, PAC)
     - Improves probabilistic performance (ie, outlier stats, RPSS)

2. Climatological mean error removed (25-yr CDC training data):
   - RMS and PAC: Very limited improvement
   - Probabilistic measures (RPSS, etc): significant gain

3. Operational (raw or bias-corrected) vs. CDC bias-corrected ens:
   - Ensemble mean: Operational much better than CDC hindcast
     - CDC has ~50% larger initial error
   - Probabilistic scores: Operational much better for out to day 10
     - For some measures, CDC hindcasts better beyond day 10

TENTATIVE CONCLUSIONS (1)

- Adaptive, regime dependent bias correction works well for first few days (almost as good as “optimal”)

- Climate mean bias correction can add value, especially for wk2 prob. fcsts

METHOD / APPLICATION – 3, 4

- Use large hindcast data set for correcting NCEP operational forecast

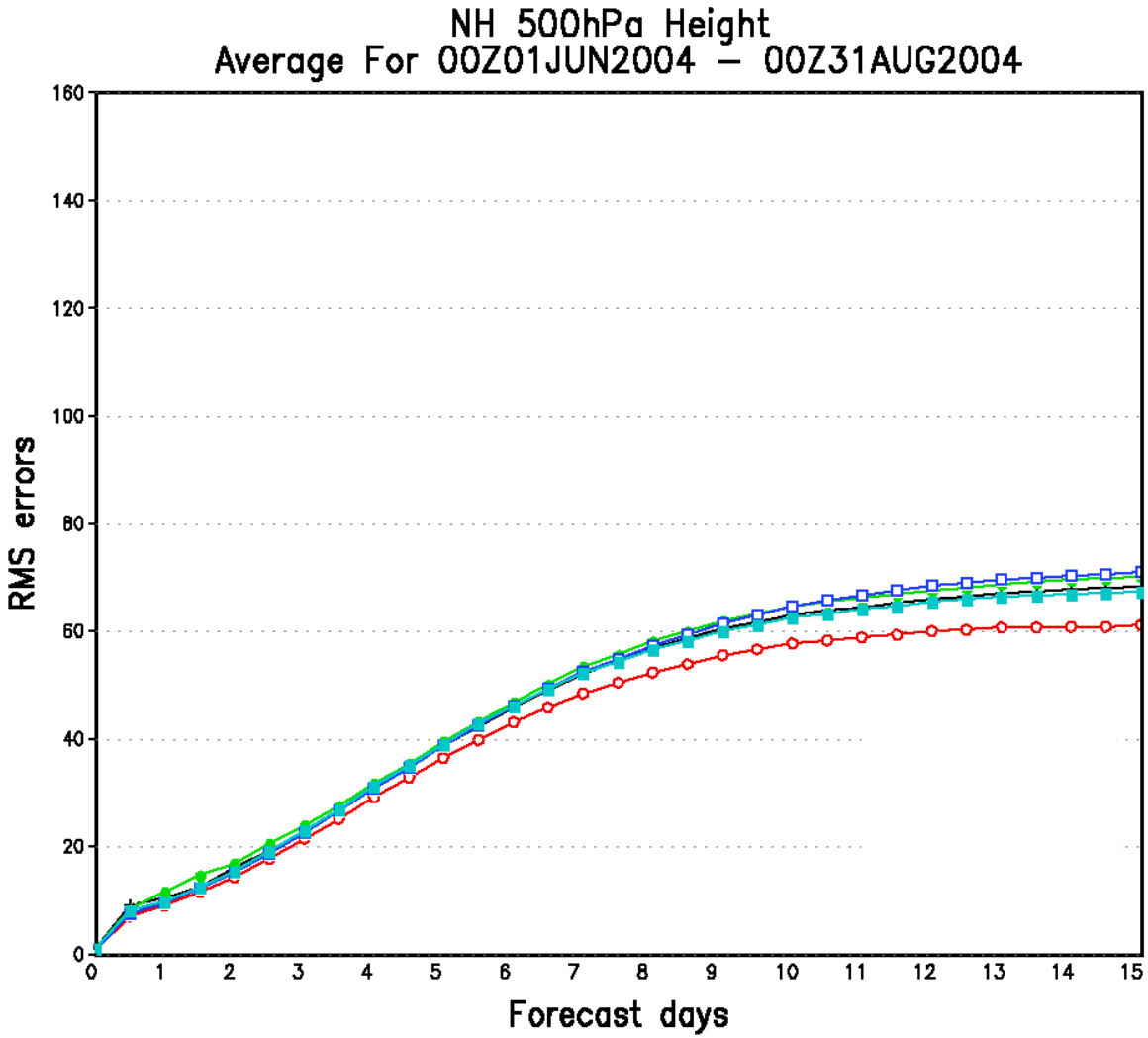
$$FCST_{clibriated} = FCST_{OPR} - BIAS_{25yr\_clim}$$

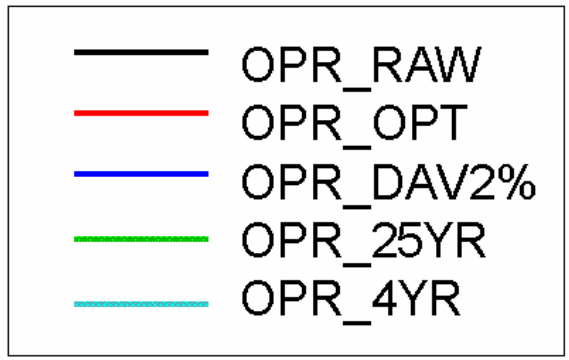
- Use 4-year average operational forecast error for correcting NCEP operational forecast

$$FCST_{clibriated} = FCST_{OPR} - BIAS_{4yr\_error}$$

- Two Experiments
  - OPR_25YR: 25-year climatological mean fcst. errors removed
  - OPR_4YR: 4-year average optimal bias (defined as 31-day centered running mean error, 2000-2003) removed

RMS: 500 mb Height, 2004 Summer
Northern Hemisphere

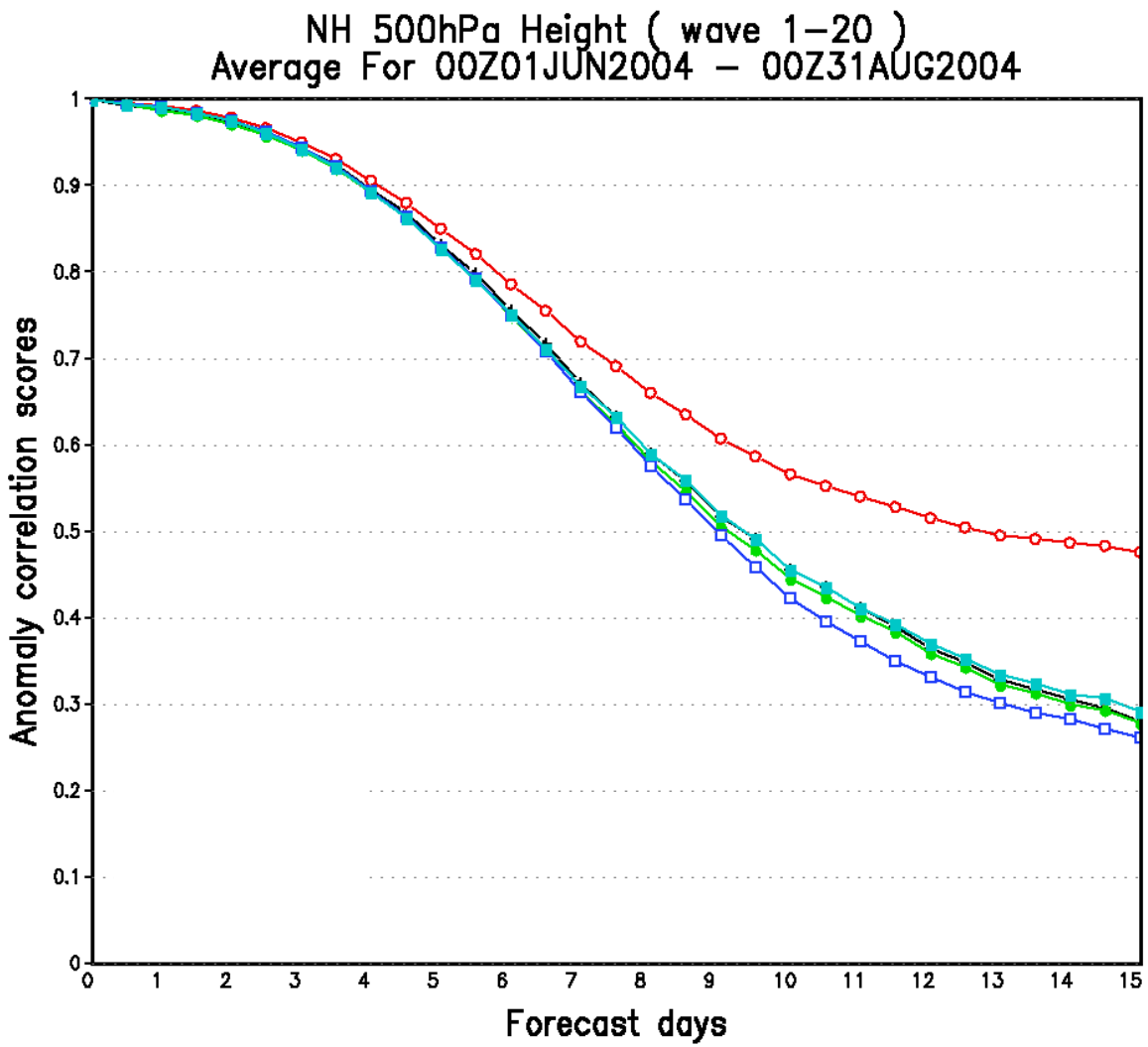


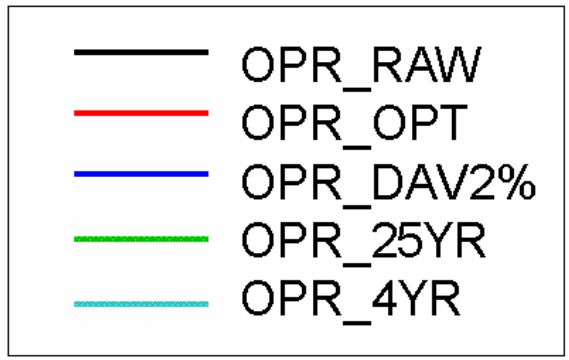


OPR_25YR
degrades ensemble mean for week-1 vs. OPR_DAV2%

OPR_4YR
limited improvement for week-2 vs. OPR_DAV2% and OPR_RAW

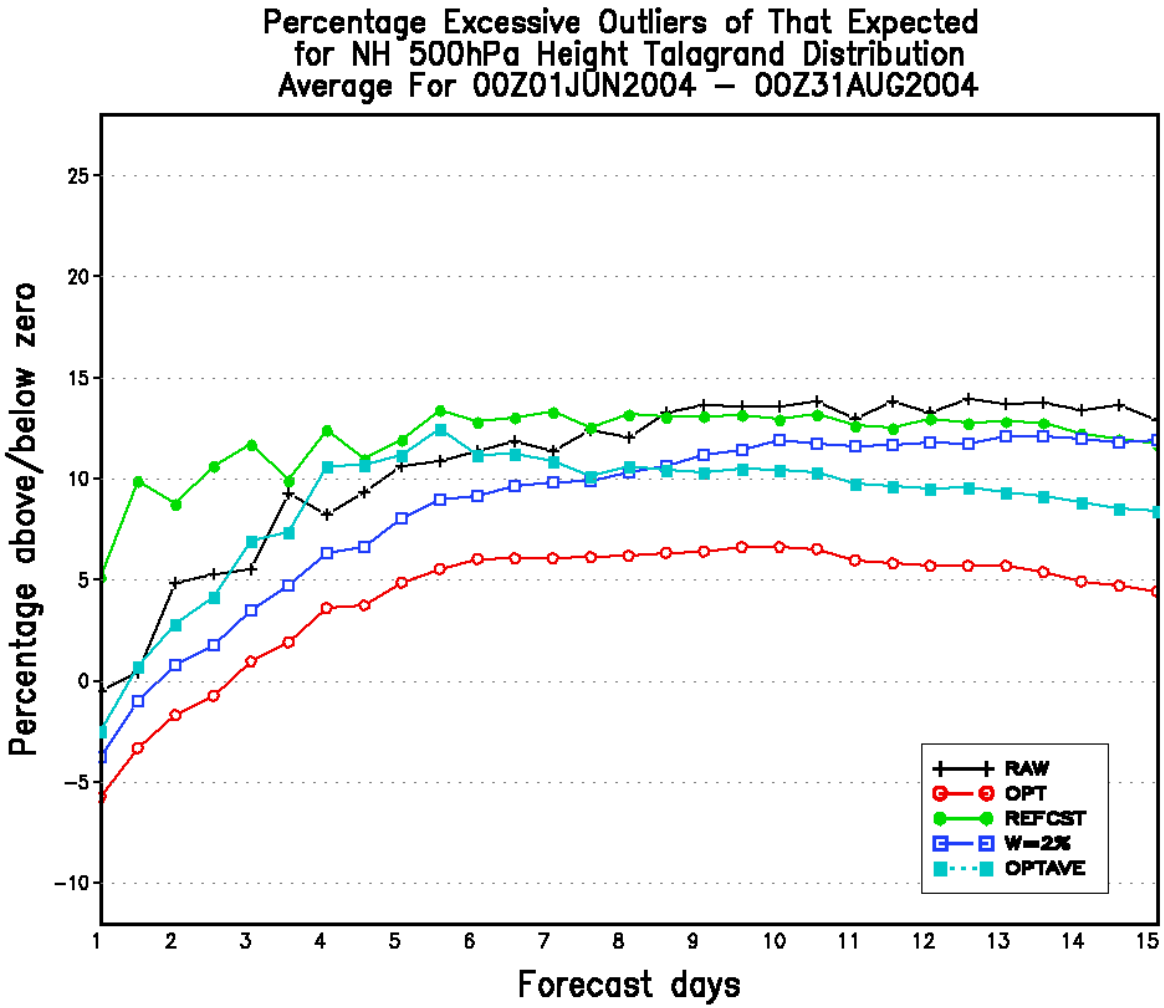
PAC: 500 mb Height, 2004 Summer
Northern Hemisphere

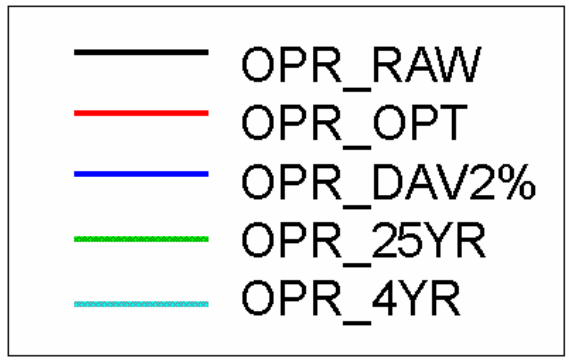




OPR_25YR & OPR_4YR improvement for week-2 vs. OPR_DAV2%

Excessive Outliers: 500 mb Height, 2004 Summer Northern Hemisphere

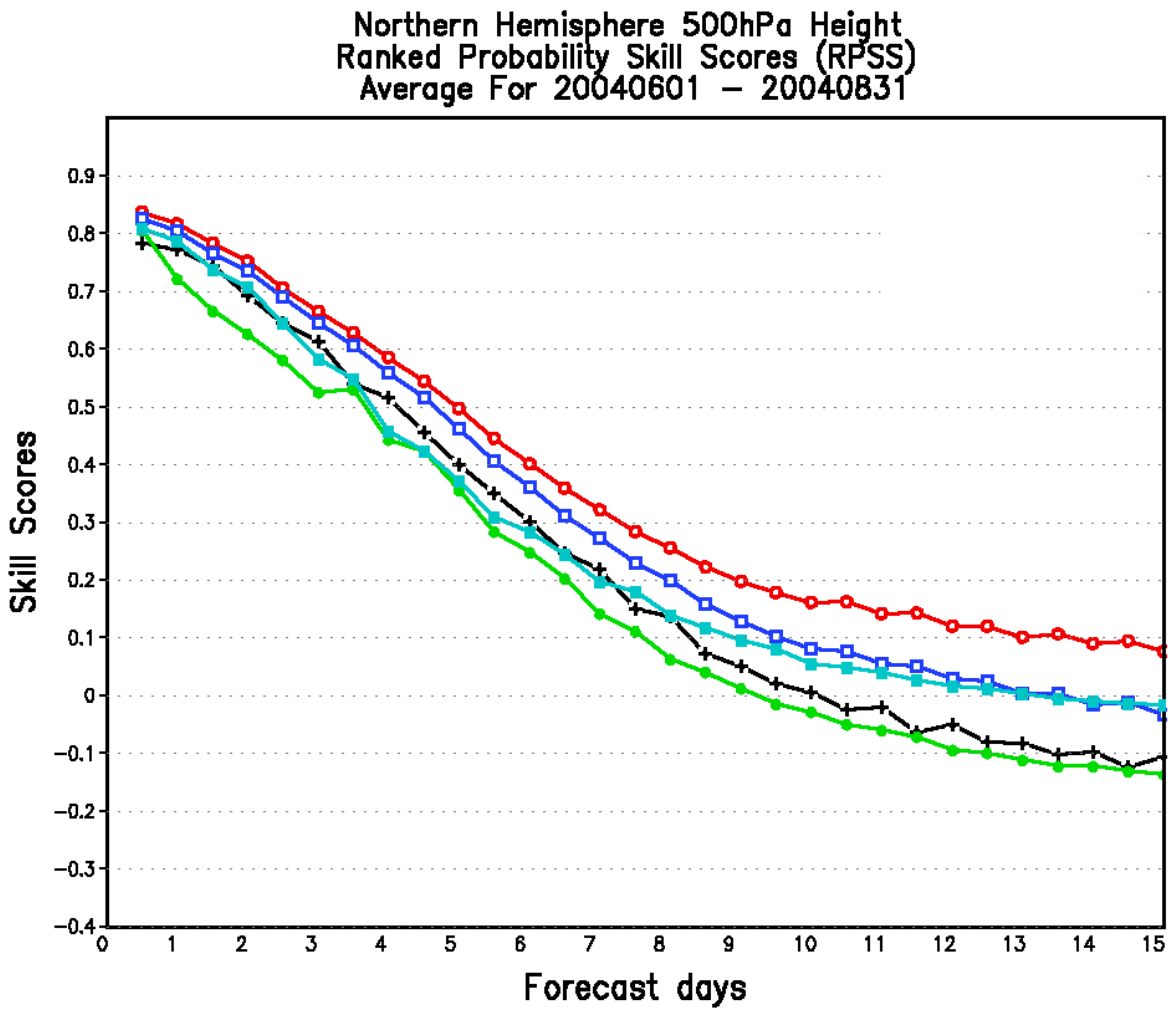


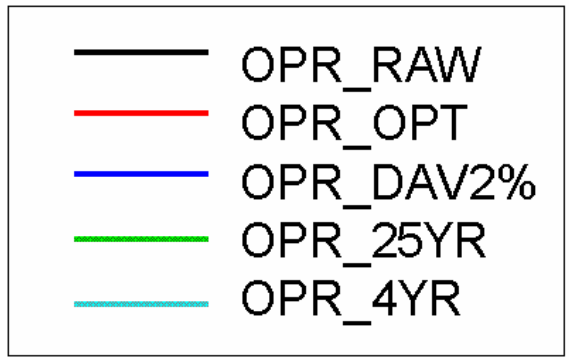


OPR_25YR
no improvement for all lead time vs. OPR_DAV2%

OPR_4YR
improvement after day 8 vs. OPR_DAV2%

RPSS: 500 mb Height, 2004 Summer Northern Hemisphere





OPR_4YR & OPR_25YR

no improvement for all lead time vs. OPR_DAV2%

PRELIMINARY RESULTS (2)

4. 25-yr climate mean error (CDC training data)

- The value added by climate mean bias correction is dependent on the NWP modeling system.
- The reforecast climatological mean error can't be used directly to calibrate the NCEP current operational ensemble

5. 4-yr NCEP operational forecast mean error:

- For some measures such as PAC, RMS and outlier stats, the bias-corrected ensemble can get improvement for week-2
- The choice of the length of training data remains a open question
- Generation of large hind-cast ensemble is expensive but can be helpful

METHOD / APPLICATION – 5

- Use large hindcast data set for correcting operational fcst. by using decaying average difference between operational and reforecast

$$FCST_{clibriated} = FCST_{OPR} - BIAS_{25yr\_clim} - BIAS_{OPR\text{-}RFC}$$

- Application to NCEP Operational Ensemble

- OPR_RFC_DAV2%: 25-year climatological mean fcst. errors and decaying averaging mean error (w=2%) between NCEP operational and CDC refcst. removed

METHOD / APPLICATION – 6

- Second Moment Bias-Correction Algorithm

ratio = r.m.s of ensemble mean / standard deviation

decaying averaging mean ratio = (1-*w*) * prior time mean ratio + *w* * ratio

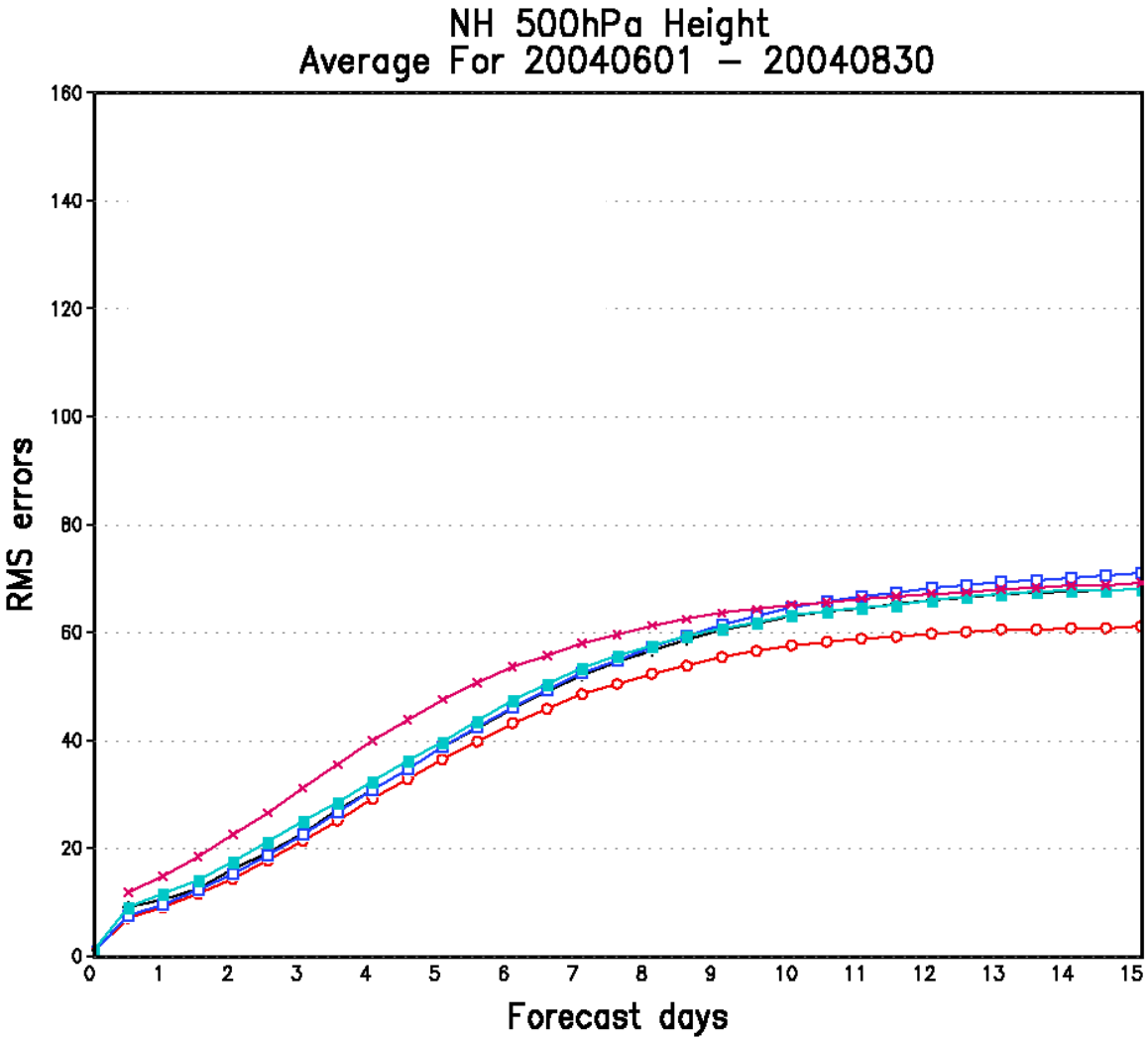
$$FCST_{clibriated} = FCST_{mean} + Ratio * (FCST_{m} - FCST_{mean})$$

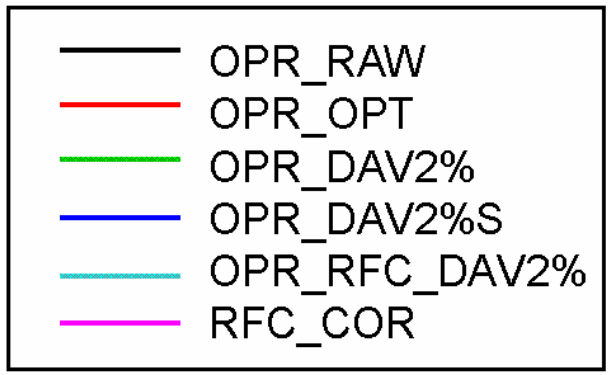
For each lead time separately

- Application to NCEP Operational Ensemble

OPR_DAV2%S: *w* = 2% (most recent ~30 days used)

RMS: 500 mb Height, 2004 Summer Northern Hemisphere

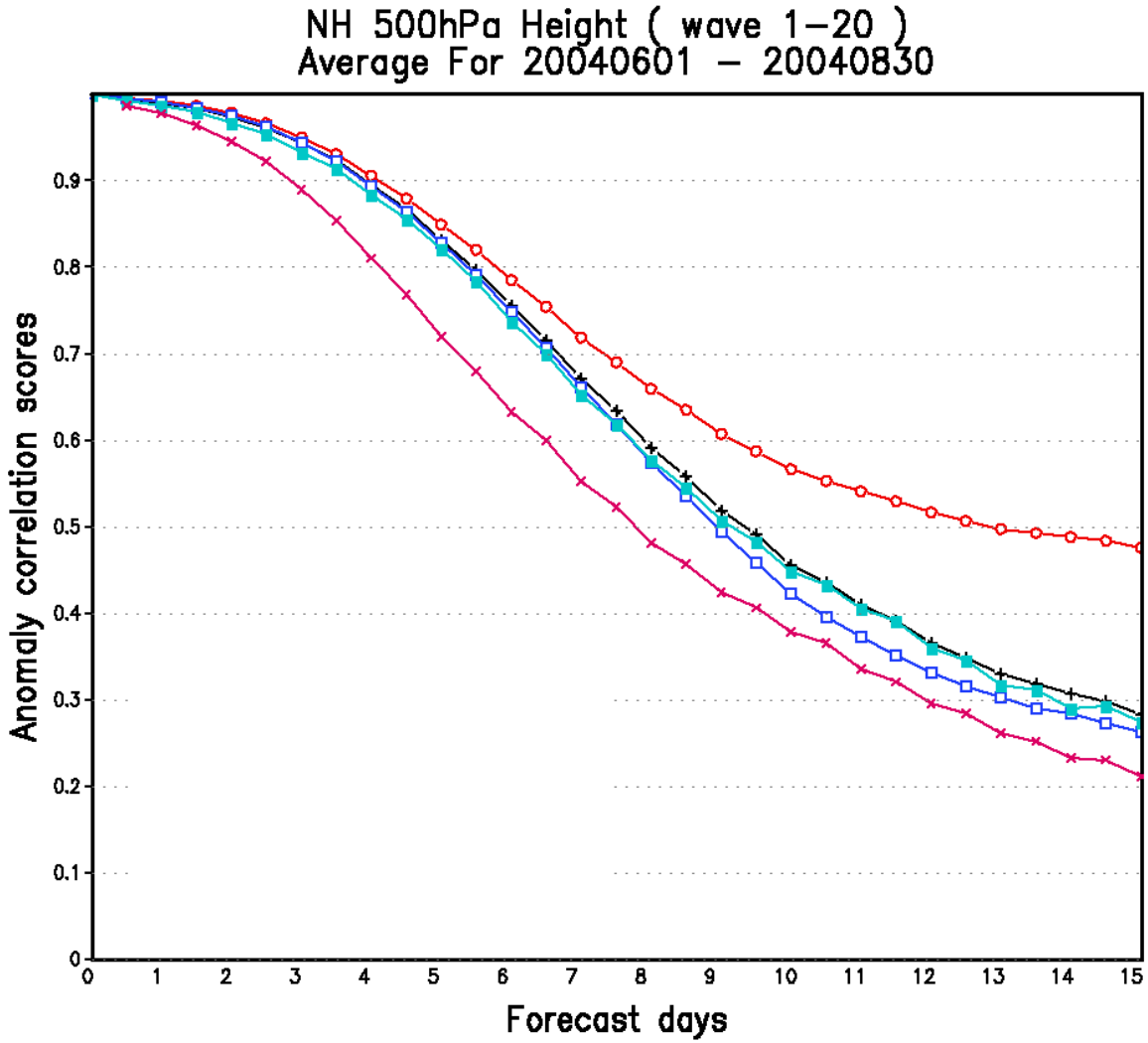


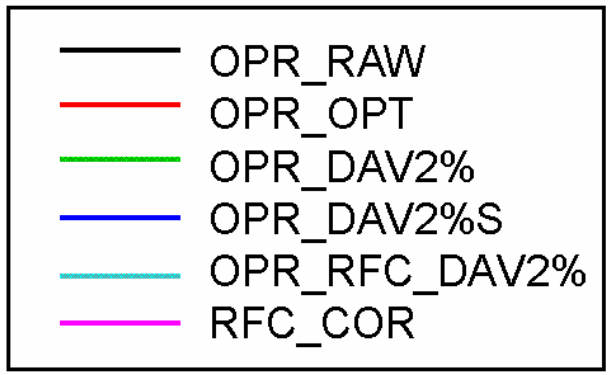


OPR_DAV2% & OPR_DAV2%S overlap

OPR_RFC_DAV2%
degrades ensemble mean for week-1, but improves it for week-2 vs. OPR_RAW
better than OPR_DAV2% and OPR_DAV2%S

PAC: 500 mb Height, 2004 Summer
Northern Hemisphere

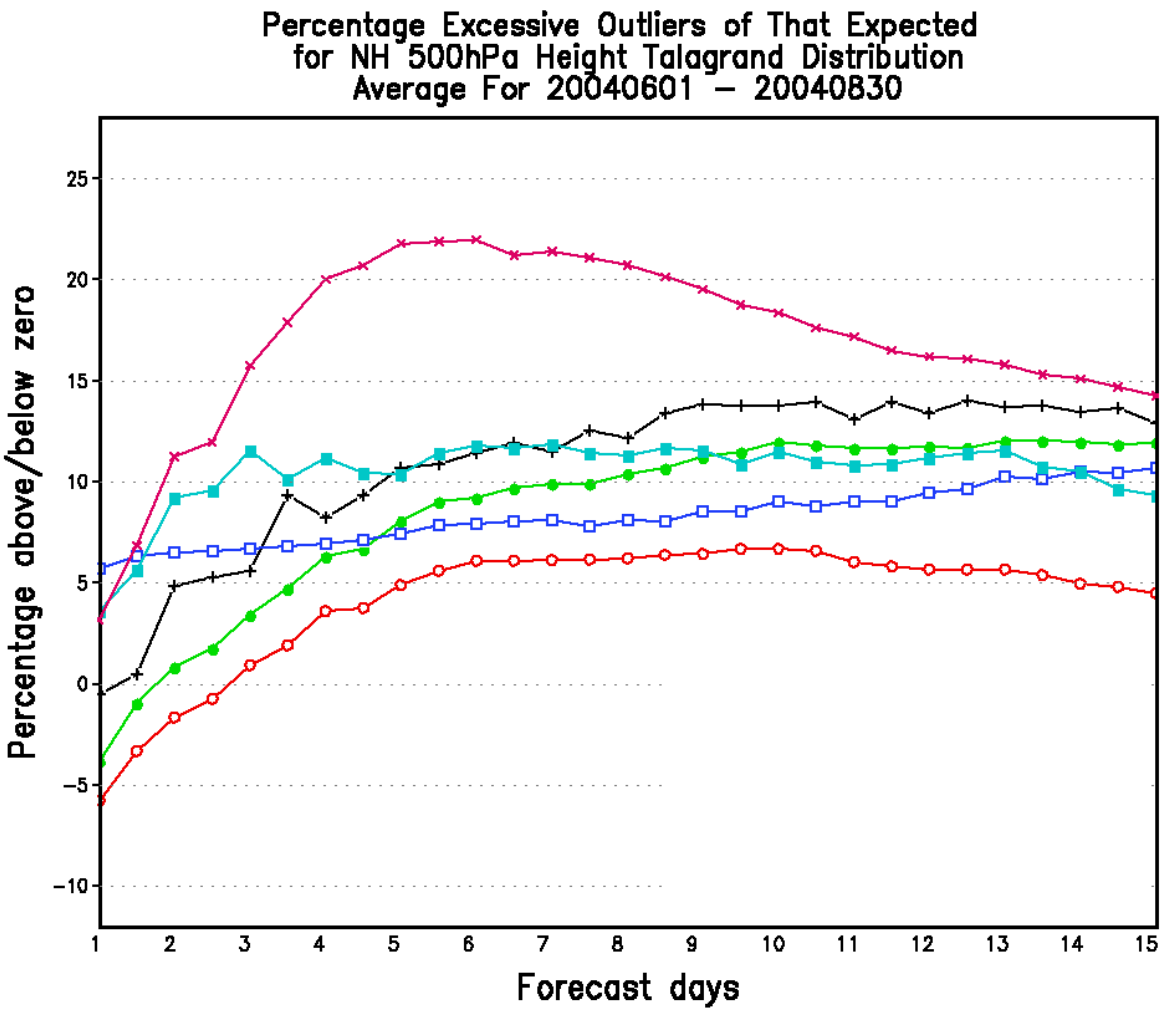


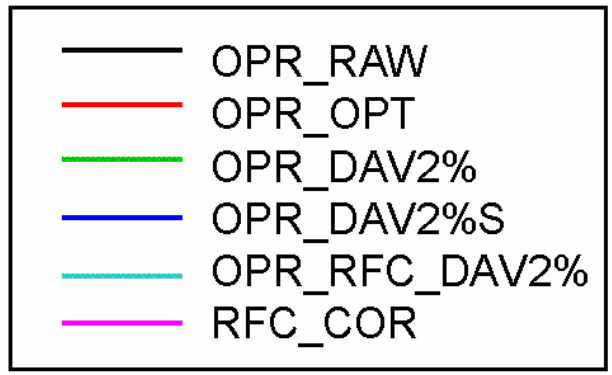


OPR_DAV2% & OPR_DAV2%S overlap

OPR_RFC_DAV2% improvement for week-2 vs. OPR_DAV2% and OPR_DAV2%S

Excessive Outliers: 500 mb Height, 2004 Summer Northern Hemisphere

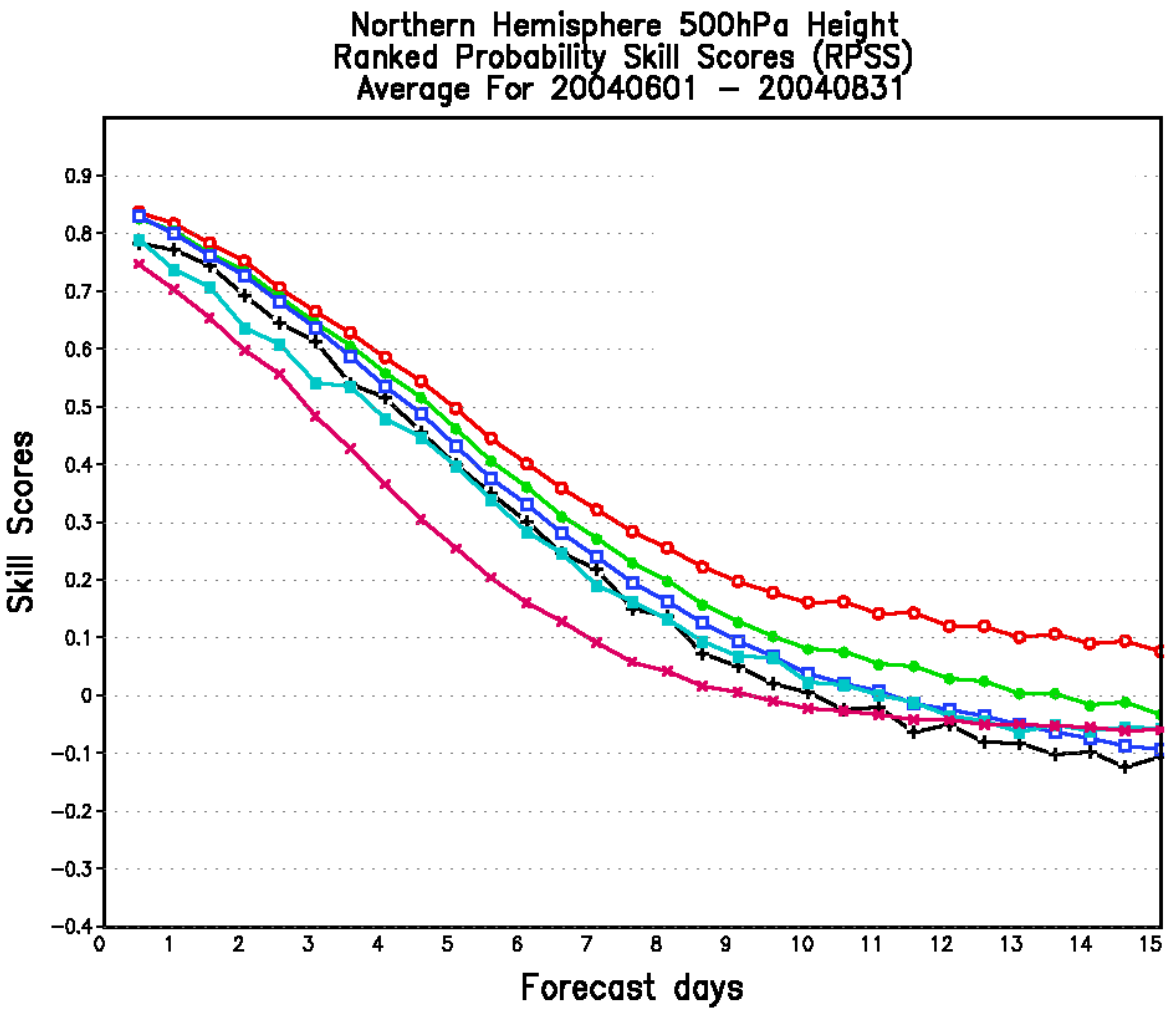


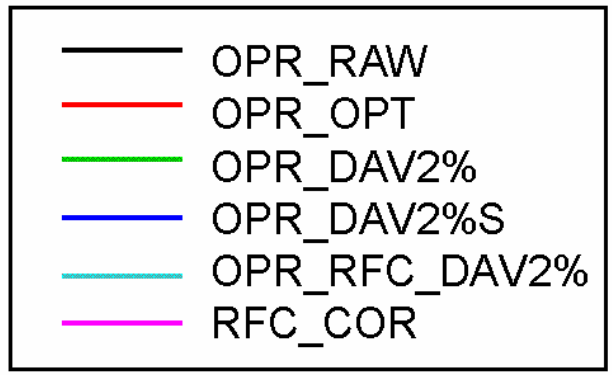


OPR_DAV2%S
Improvement vs.
OPR_DAC2% after day 5

OPR_RFC_DAV2%
better than OPR_DAV2%
after day 9 and better than
OPR_DAV2%S after day 14

RPSS: 500 mb Height, 2004 Summer Northern Hemisphere

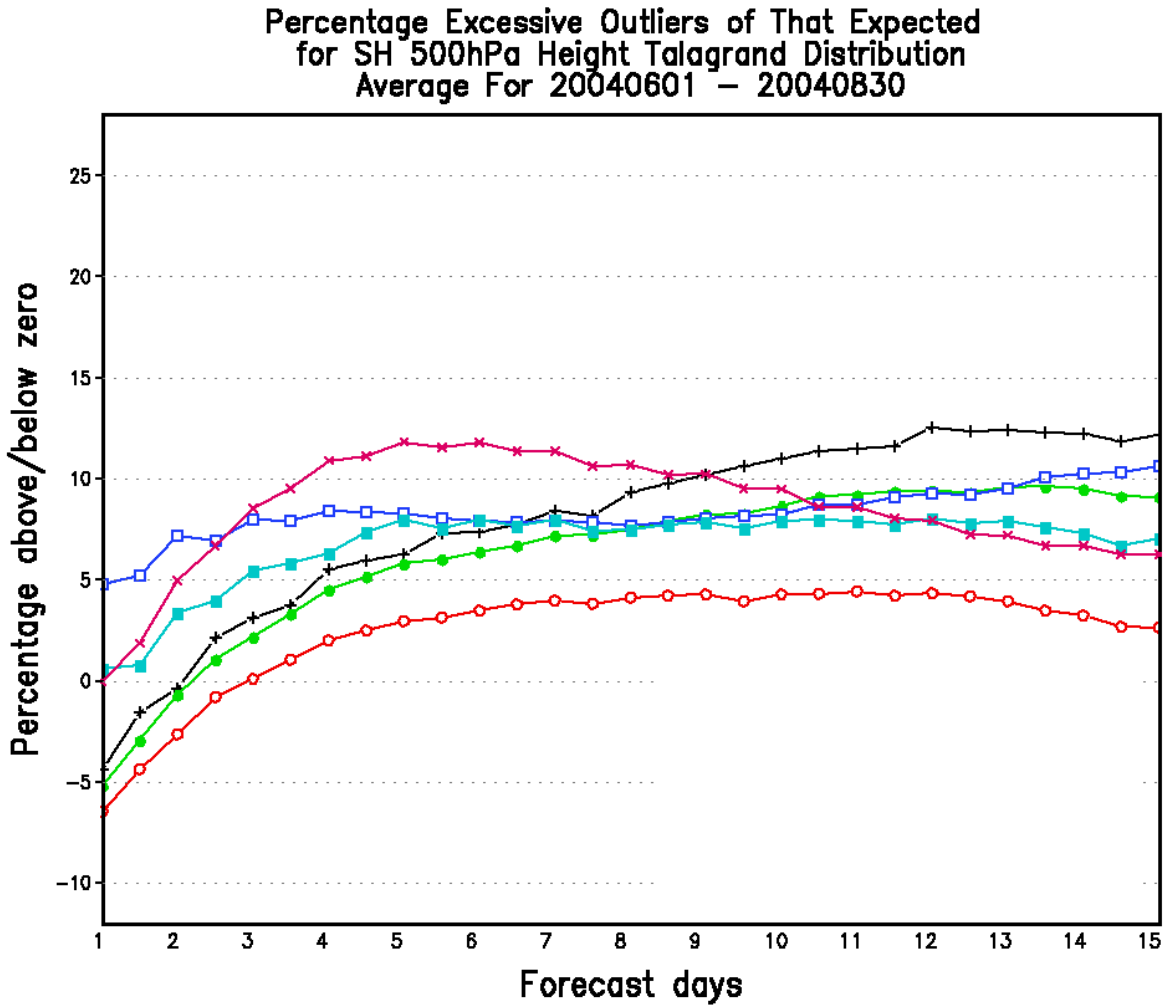


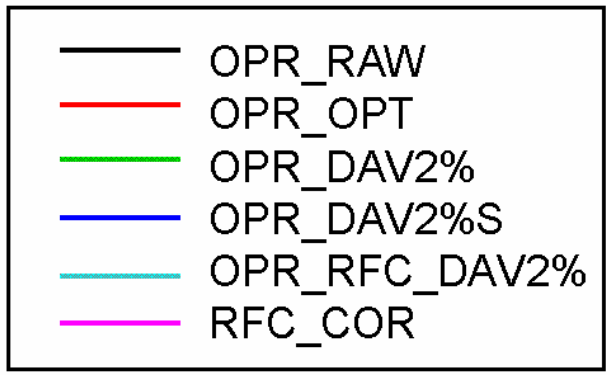


OPR_DAV2%S
no improvement vs.
OPR_DAC2%

OPR_RFC_DAV2%
no improvement for all lead time vs.
OPR_DAC2%

Excessive Outliers: 500 mb Height, 2004 Summer Southern Hemisphere

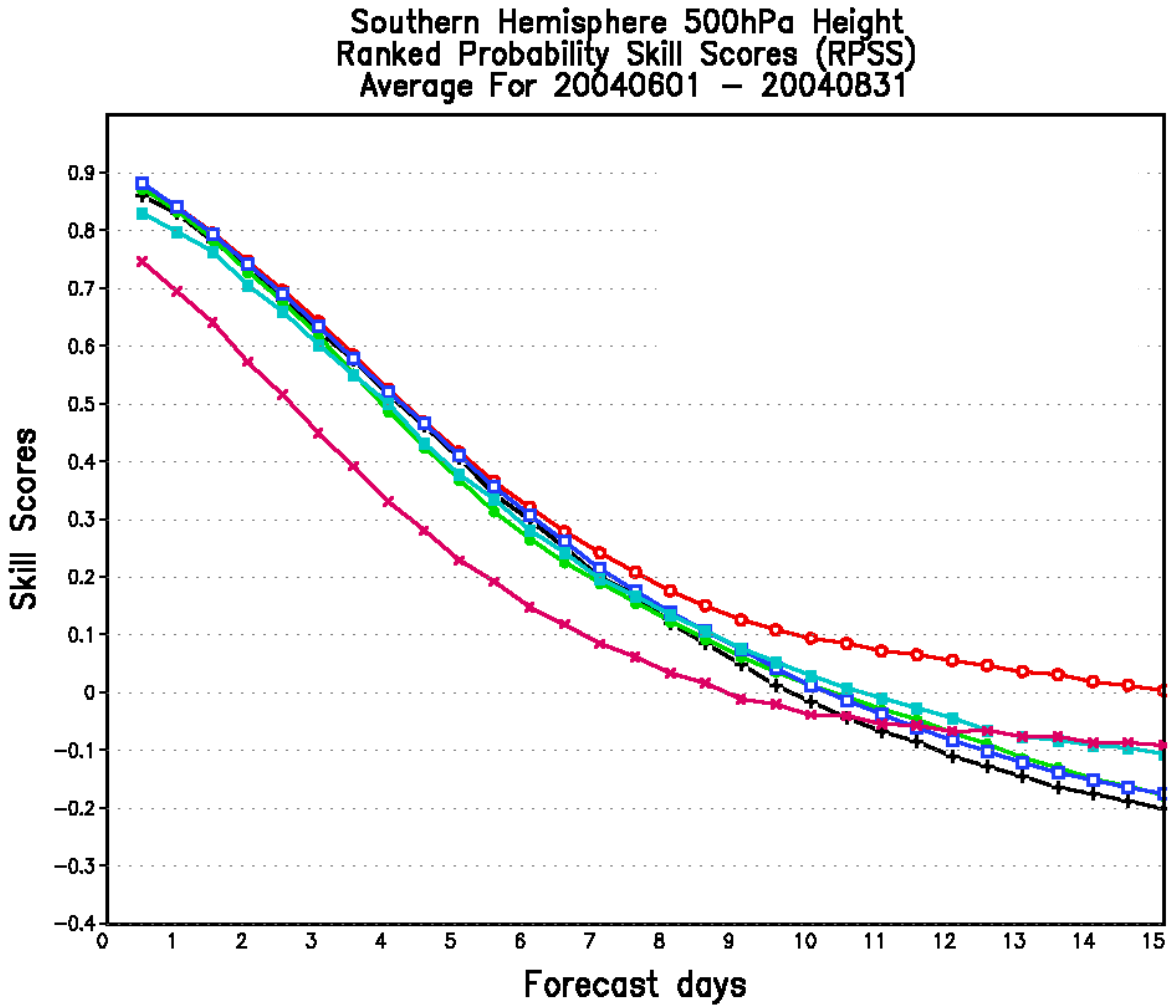


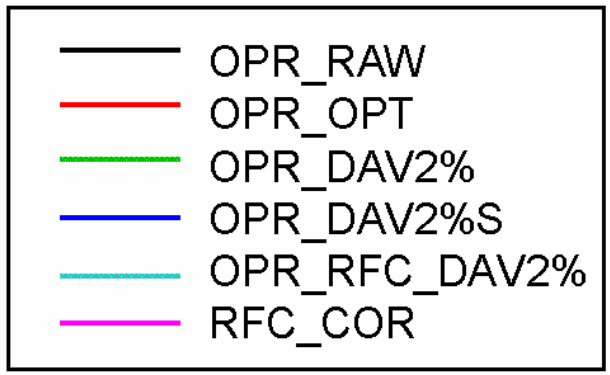


OPR_DAV2%S close to OPR_DAC2%

OPR_RFC_DAV2% improvement for week-2 vs. OPR_DAC2% and OPR_DAV2%S

RPSS: 500 mb Height, 2004 Summer Southern Hemisphere





OPR_DAV2%S
improves for week-1vs.
OPR_DAC2%

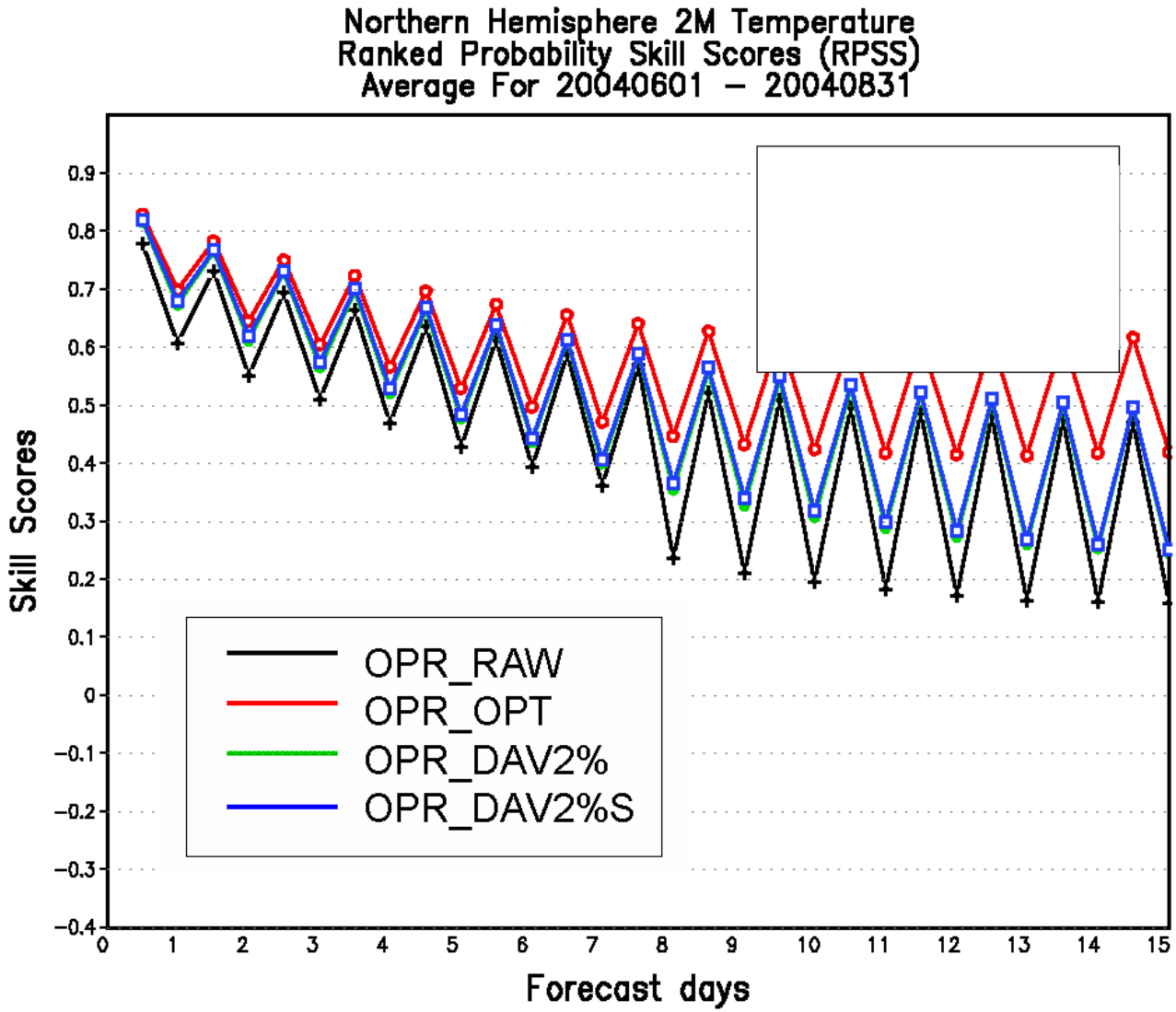
OPR_RFC_DAV2%
degrades for week-1,
improves for week-2 vs.
OPR_DAC2% and
OPR_DAV2%S

PRELIMINARY RESULTS (3)

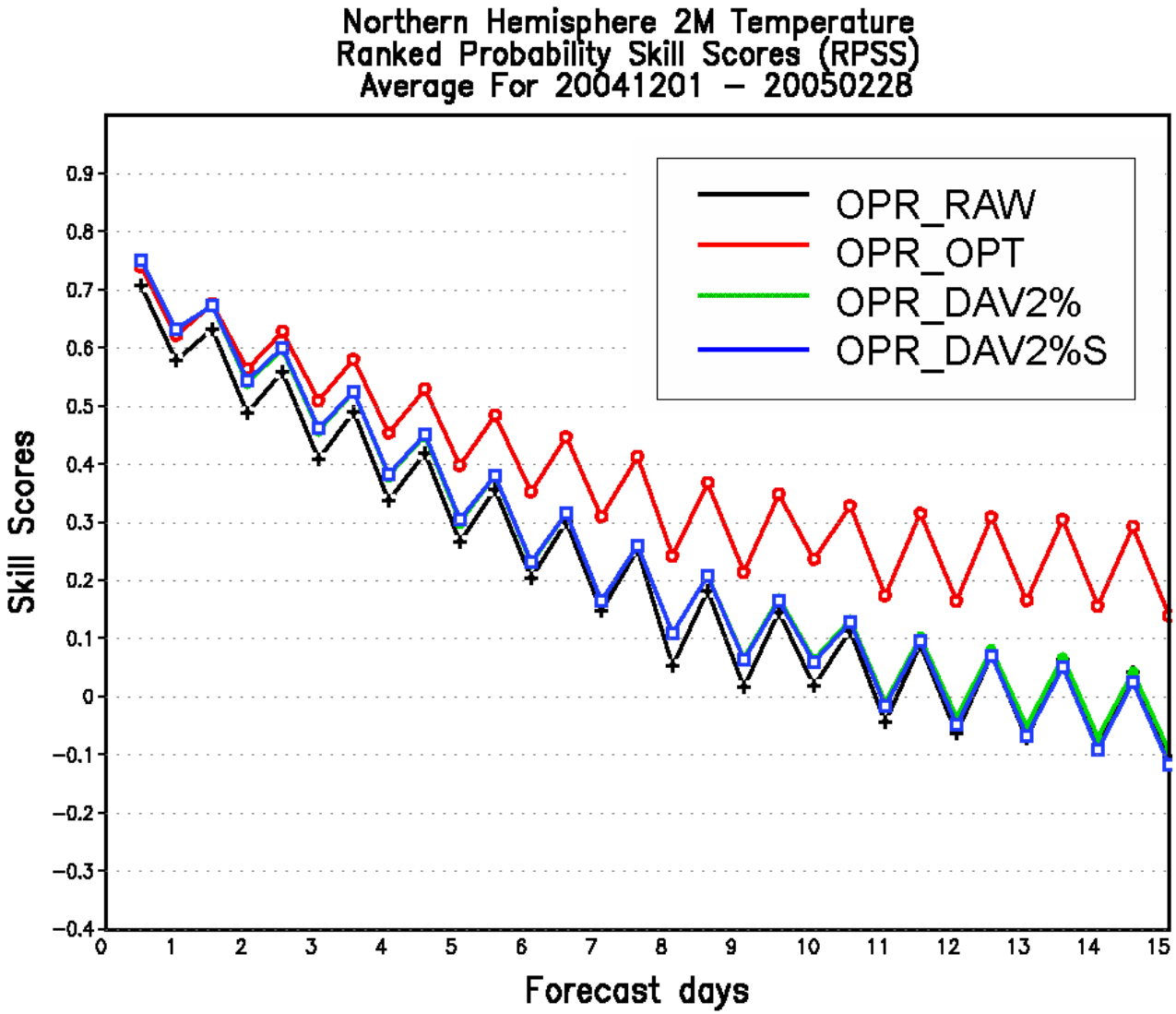
6. OPR_RFC_DAV2% (use both large hindcast data set and decaying average difference between operational and reforecast)
   - Show improvement for some measures and regions

7. Second Moment Bias-Correction Algorithm:
   - No significant improvement, the calculation of the 2nd moment ratio needs more consideration

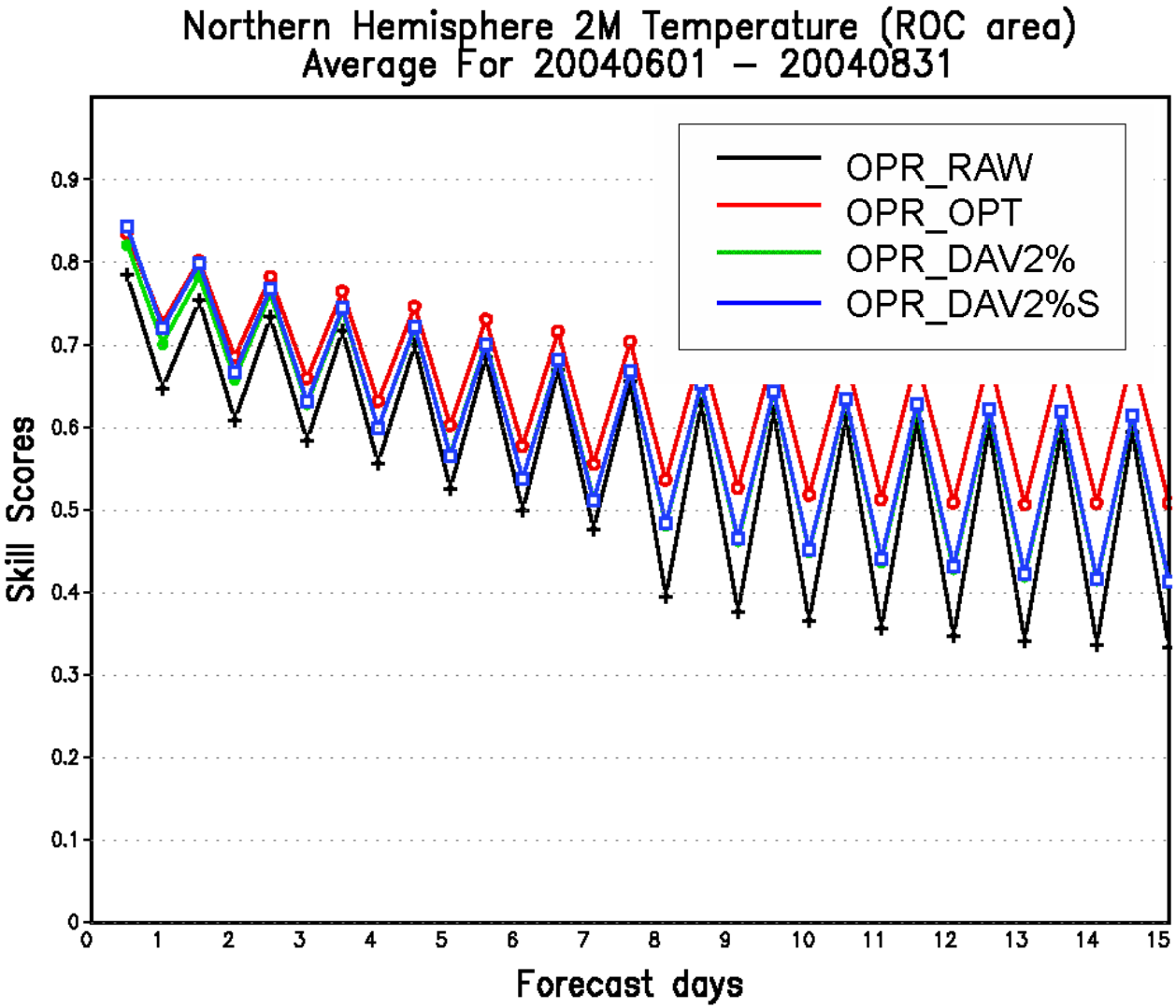
RPSS: 2m Temperature, 2004 Summer Northern Hemisphere



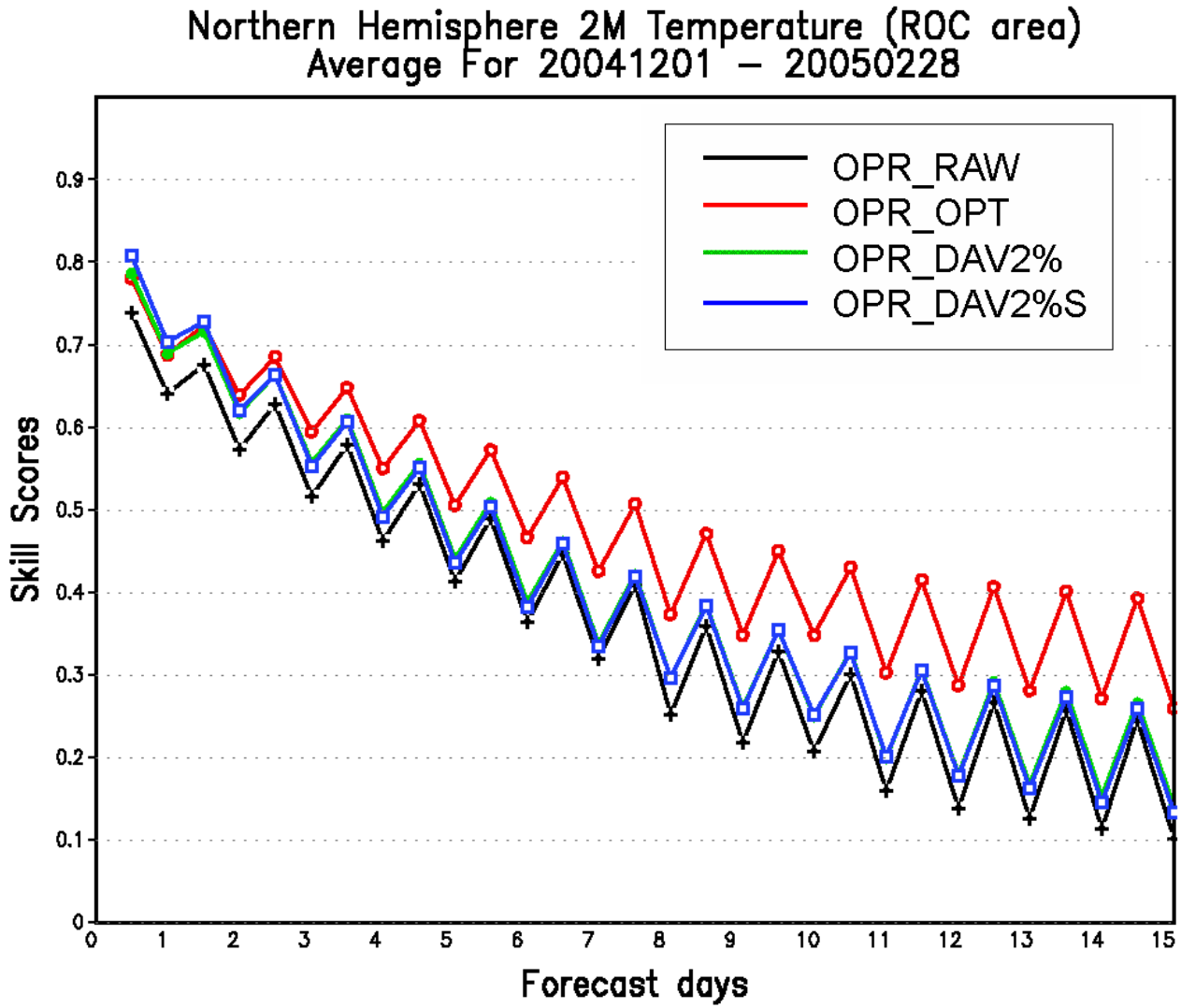
RPSS: 2m Temperature, 2004 Winter Northern Hemisphere



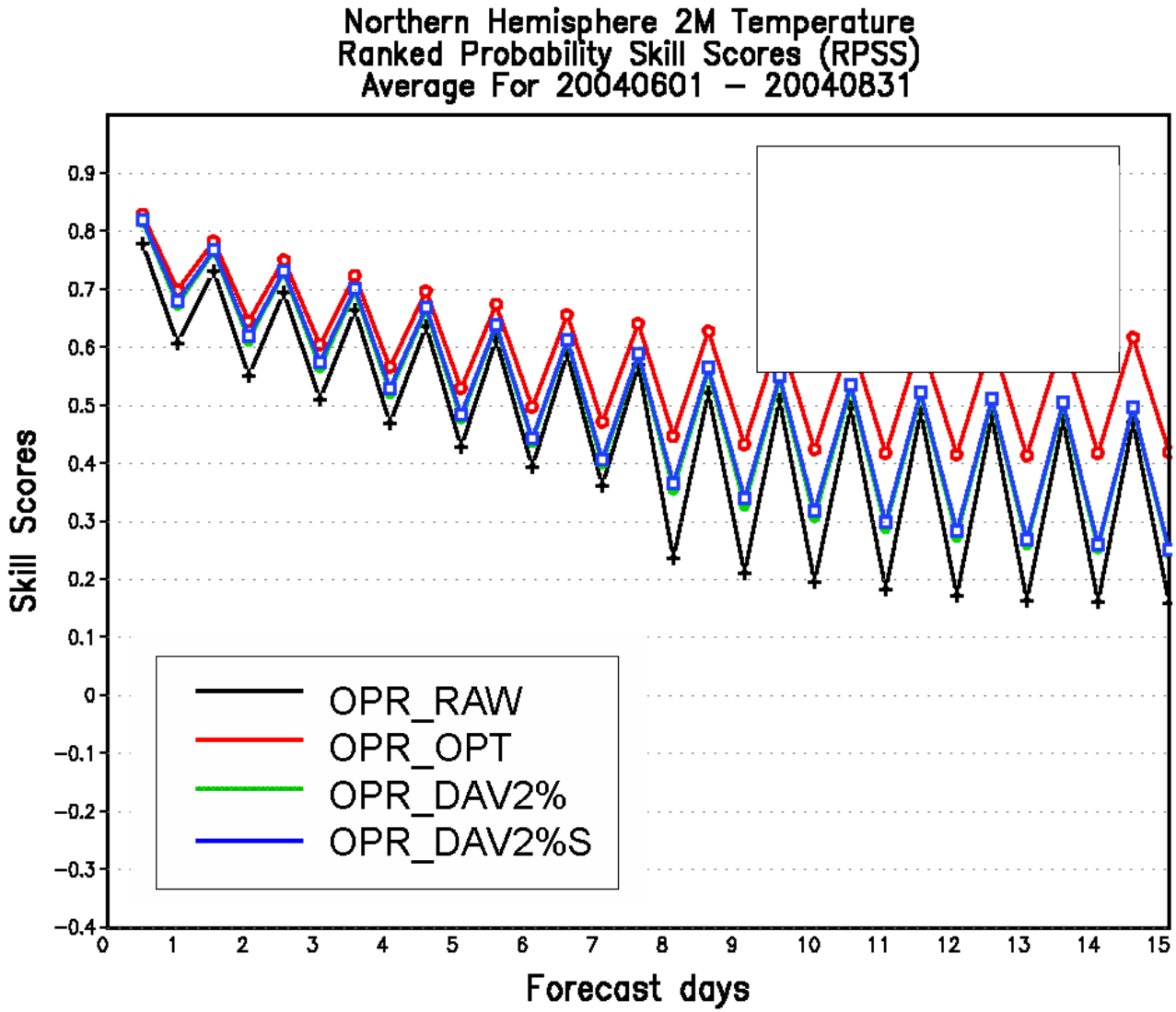
ROC: 2m Temperature, 2004 Summer Northern Hemisphere



ROC: 2m Temperature, 2004 Winter Northern Hemisphere



RPSS: 2m Temperature, 2004 Summer Northern Hemisphere



TENTATIVE CONCLUSIONS

- Adaptive, regime dependent bias correction works well for first few days (almost as good as “optimal”)
  - Frequent updates of DA/NWP modeling system possible
- Climate mean bias correction can add value, especially for wk2 prob. fcsts
  - Generation of large hind-cast ensemble is expensive but can be helpful
- The best performing methods can be selected for use of other ensemble fcst. variables, U, V, cumulative frequency distribution for QPF
- Use of up-to-date data assimilation/NWP techniques imperative at all ranges

OPEN QUESTIONS

- How to gain benefits of both
  - Frequent updates to DA/NWP system ***AND***
  - Large hind-cast data set?
- Are week-2 biases dependent on specific version of DA/wk-1 model used?
  - Will test; if not,
- Will a “hybrid” system work?
  - Use latest DA/NWP system for week-1, with adaptive bias correction
  - Branch off at D5 with less frequently upgraded model with large hindcast data set
    - Combine benefits of improved short-range performance & large wk2 hind-cast data set
- Alternatively, can a large hind-cast dataset be generated before each (major) DA/NWP model upgrade?
- Do we need to consider additional new criteria for operational DA/NWP implementations?
  - Compare objective scores for operational & experimental systems
    - Current practice: Compute scores ***without bias correction***
      - Good for model development purposes
    - Possible additional new way: Compute also scores ***after bias correction***
      - Needed as additional test before operational implementation?
- Implement new DA/NWP systems only if bias corrected fcsts improve?
  - Minor changes may not require new hind-casts
  - Major changes will need generation of associated hind-cast dataset?

Questions and Comments?

More plots on http://www.emc.ncep.noaa.gov/gmb/ens/