

The use of a multi-model ensemble in local operational ozone forecasting

William F. Ryan

Jeremy Geiger

Department of Meteorology
The Pennsylvania State University

3rd International Workshop on Air Quality Research,
Washington, DC, November, 2011

Brief Background on Operational O₃ Forecasting in the United States

- Forecasts are typically issued on the metropolitan scale by local (non-NOAA) forecasters, and are verified against the domain-wide peak 8-hour O₃ at monitors operated by state or local AQ agencies.
 - In PHL, roughly 10-12 monitors.
- Key forecast threshold is an 8-hour average concentration ≥ 76 ppbv (Code Orange).
- In order to activate “Air Quality Action Day” plans, forecasts are issued ~ 1800 UTC on the previous day – an 18-30 hour forecast.

Standard Forecast Techniques

- Outside the margins of the season, climatology is not useful but persistence (one day lag) is a skillful forecast.
 - Persistence “explains” ~ 36% of variance in peak 8-h O₃ in PHL.
- Persistence can be coupled with forecast back trajectories to assess regional transport (expected residual layer concentrations).
- Simple statistical models using meteorological and persistence predictors were useful but have become less so in the changing emissions environment post-2002.
- Conceptual models are useful at the regional scale, and for longer range forecasts, as multi-day, severe high O₃ events in the mid-Atlantic typically follow certain “classic” synoptic scale patterns.
- However, daily metropolitan scale peak O₃ concentrations vary on the meso-scale. These effects that may be successfully simulated by the latest generation of coupled chemistry-meteorological models.
- Post-processed model guidance is underway in Canada but has not been implemented in the US.
 - Though bias correction methods are available, see AQMOS later.

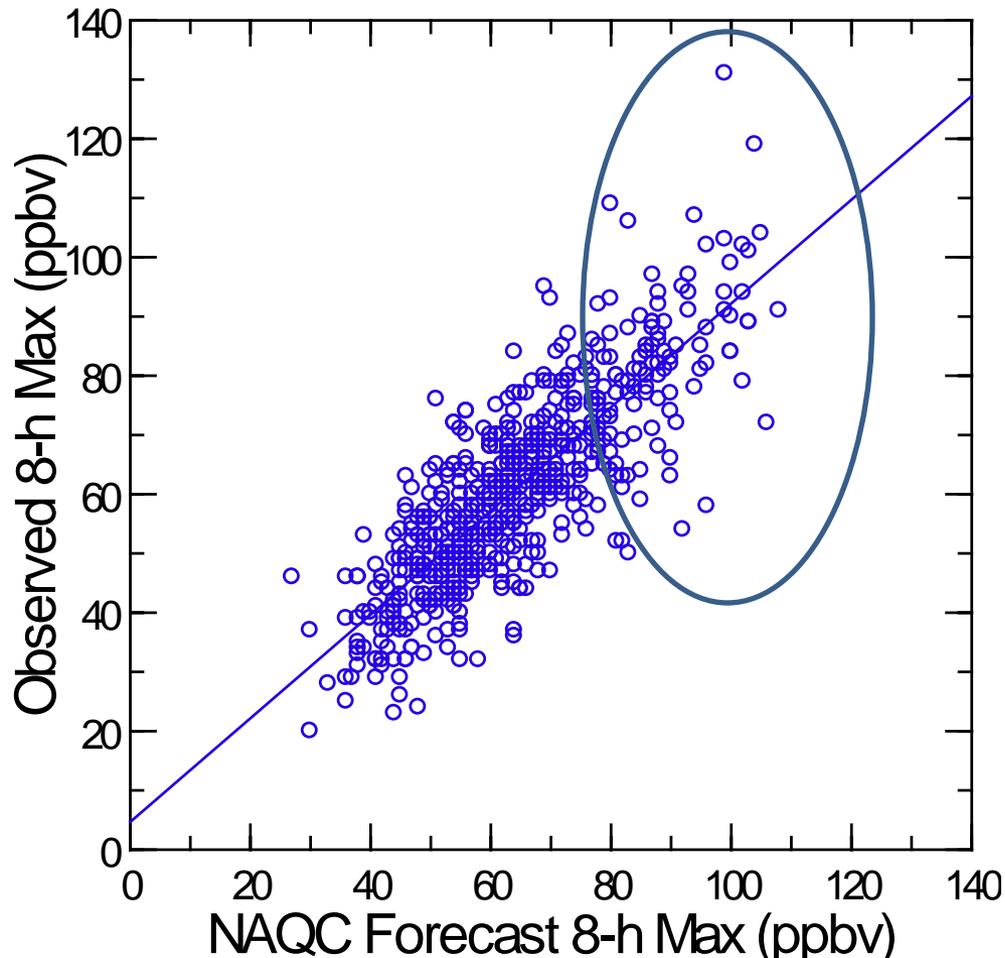
Hypothesis

- Several skillful numerical O₃ forecast models are now reliably available in a timely manner for use in preparing operational O₃ forecasts.
 - These models vary in terms of their meteorological model drivers, emissions data bases and chemical mechanisms.
- An “ensemble” of these models will provide better forecasts at the critical forecast threshold (Code Orange) than a single model.

Useful Reading

- Djalalova, I., et al, 2010: Ensemble and bias-correction techniques for air quality model forecasts of surface O₃ and PM_{2.5} during the TEXAQS-II experiment of 2006, *Atmos. Environ.*, **44**, 455-467.
- McKeen, S., et al, 2005: Assessment of an ensemble of seven real-time ozone forecasts over eastern North America during the summer of 2004, *J. Geophys. Res.*, **110**, doi:10.1029/2005JD005858.

NOAA Operational Forecast Model Results - Philadelphia Peak 8-Hour Domain-Wide O₃ (May-September, 2007-2010)



Bias: +3.3 ppbv

Median Absolute Error:
7.0 ppbv

Mean Absolute Error:
8.0 ppbv

Best Linear Fit:

$$[\text{O}_3]_{\text{OBS}} = 4.7 + 0.875 * [\text{O}_3]_{\text{FC}}$$

Increased spread in the
higher end of the distribution.

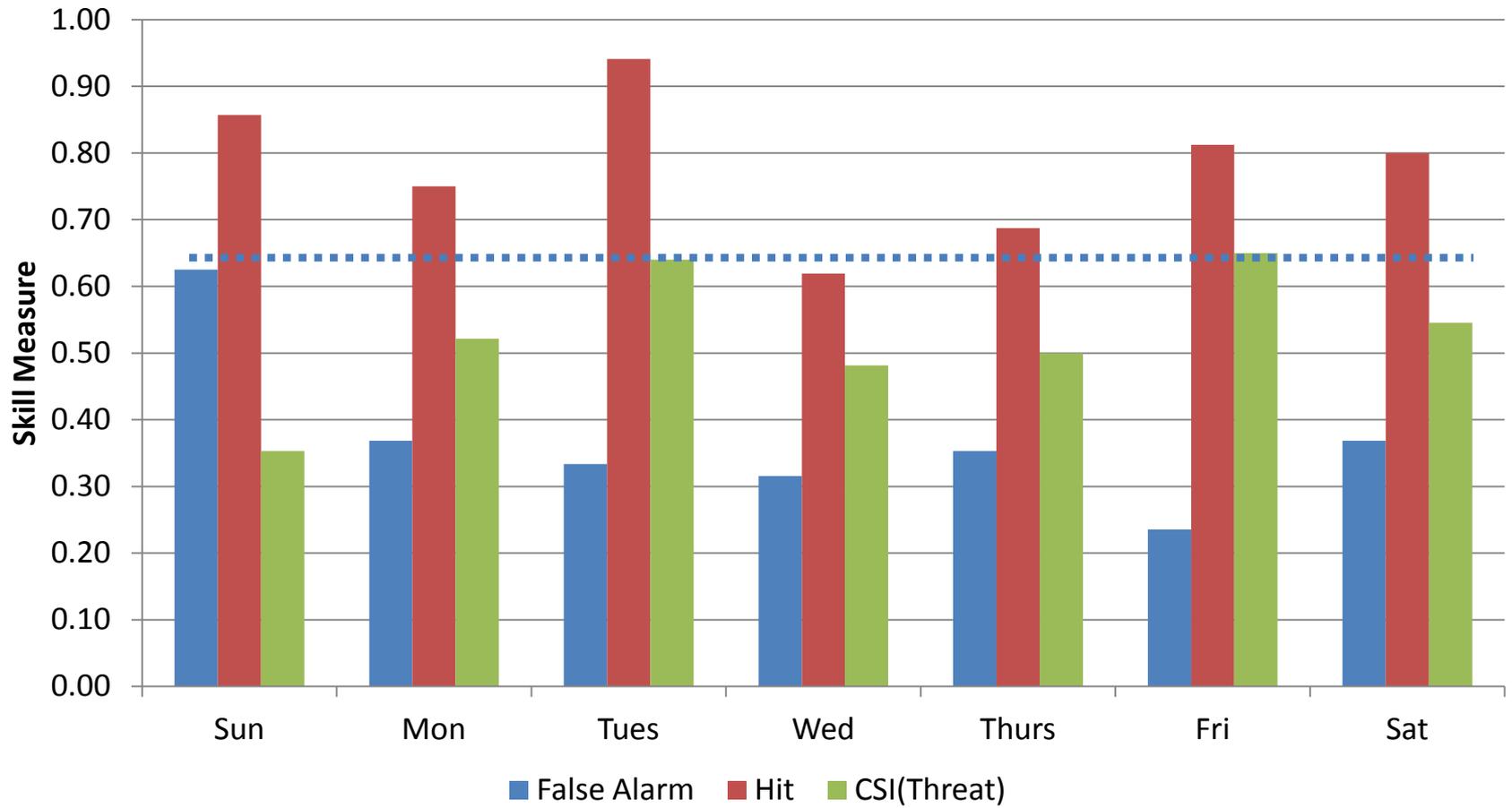
For 76 ppbv Threshold:

Hit Rate: 0.77

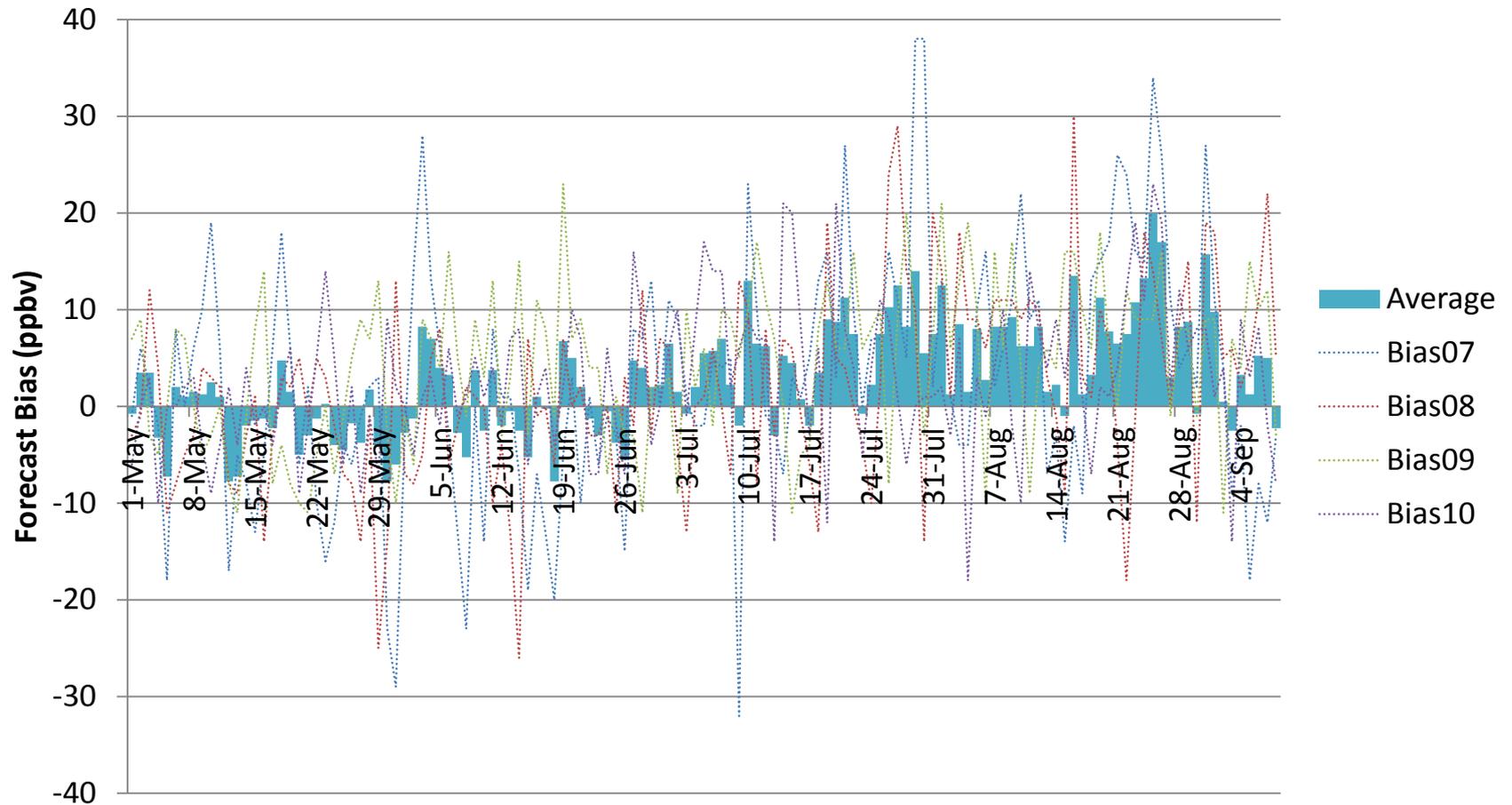
False Alarm: 0.37

Threat Score: 0.58

NOAA Operational Model Sunday False Alarms



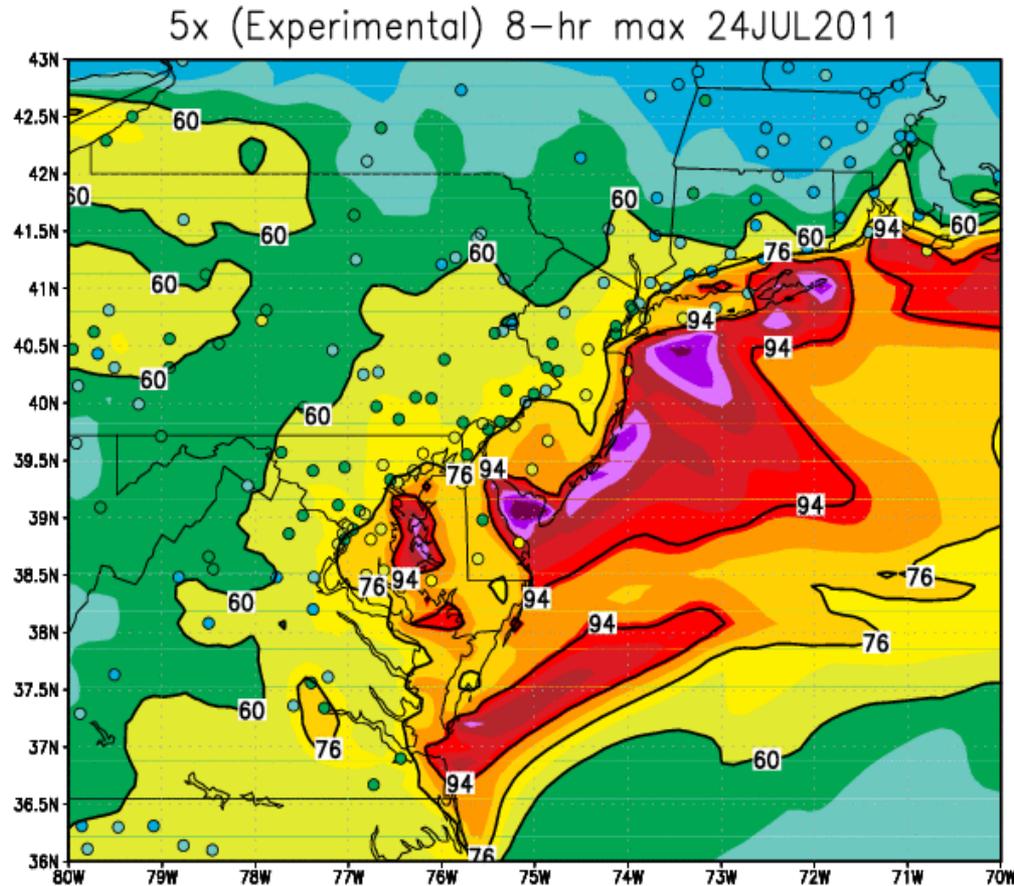
NOAA Model Seasonal Drift



Forecast Models Used in Ensemble

- NAQC (NOAA) – Queried at Monitor Locations
 - 1200 UTC Run valid following day (24-36 h forecast)
 - <http://www.emc.ncep.noaa.gov/mmb/aq/>
- ZIP/NAQC – NOAA Model Queried at all Domain Land Areas
 - Data extracted from AQMOS (Sonoma Tech)
 - <http://aqmos.sonomatech.com/login.cfm>
- AQMOS – NOAA Model with Seasonal Bias Correction
 - <http://aqmos.sonomatech.com/faq/>
- Barons Meteorological Services – MAQSIP RT
 - 0600 UTC Run
 - <http://www.baronams.com/products/>
- SUNY-Albany
 - 1200 UTC Run, “NYSDEC_3x12z”, CMAQ 4.7.1
 - http://asrc.albany.edu/research/aqf/aqvis/tomorrowforecast_maps.htm

Comment on Determination of Peak Model 8-Hour O₃



Numerical forecast models tend to develop strong sea-land O₃ gradients.

At left: See high O₃ in embayments and along NJ coast.

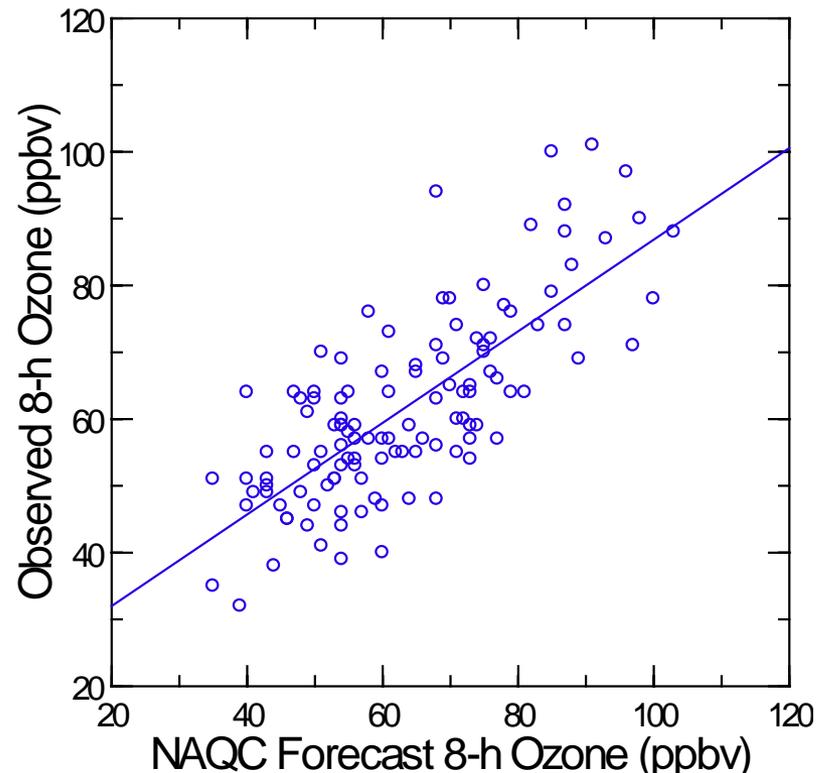
If use entire area to verify model, rather than forecasts at monitor locations, the number of false alarms of high O₃ more than double and bias increases by 7 ppbv (2011 results, NOAA Operational Model).

Ensembles Tested in 2011

- ENS1: NAQC, SUNY, BARONS, ZIP
- **ENS2: NAQC, SUNY, BARONS, ZIP, AQMOS**
- ENS3: NAQC, SUNY, BARONS, AQMOS
- ENS4: NAQC, ZIP
- ENS5: NAQC, ZIP, AQMOS
- ENS6: ZIP, AQMOS
- ENS7: NAQC, NAQC Experimental
- ENS8: NAQC, SUNY, AQMOS
- **ENS9: NAQC, SUNY, ZIP, AQMOS**
- ENS10: NAQC, SUNY, Barons
- **ENS11: SUNY, Barons**

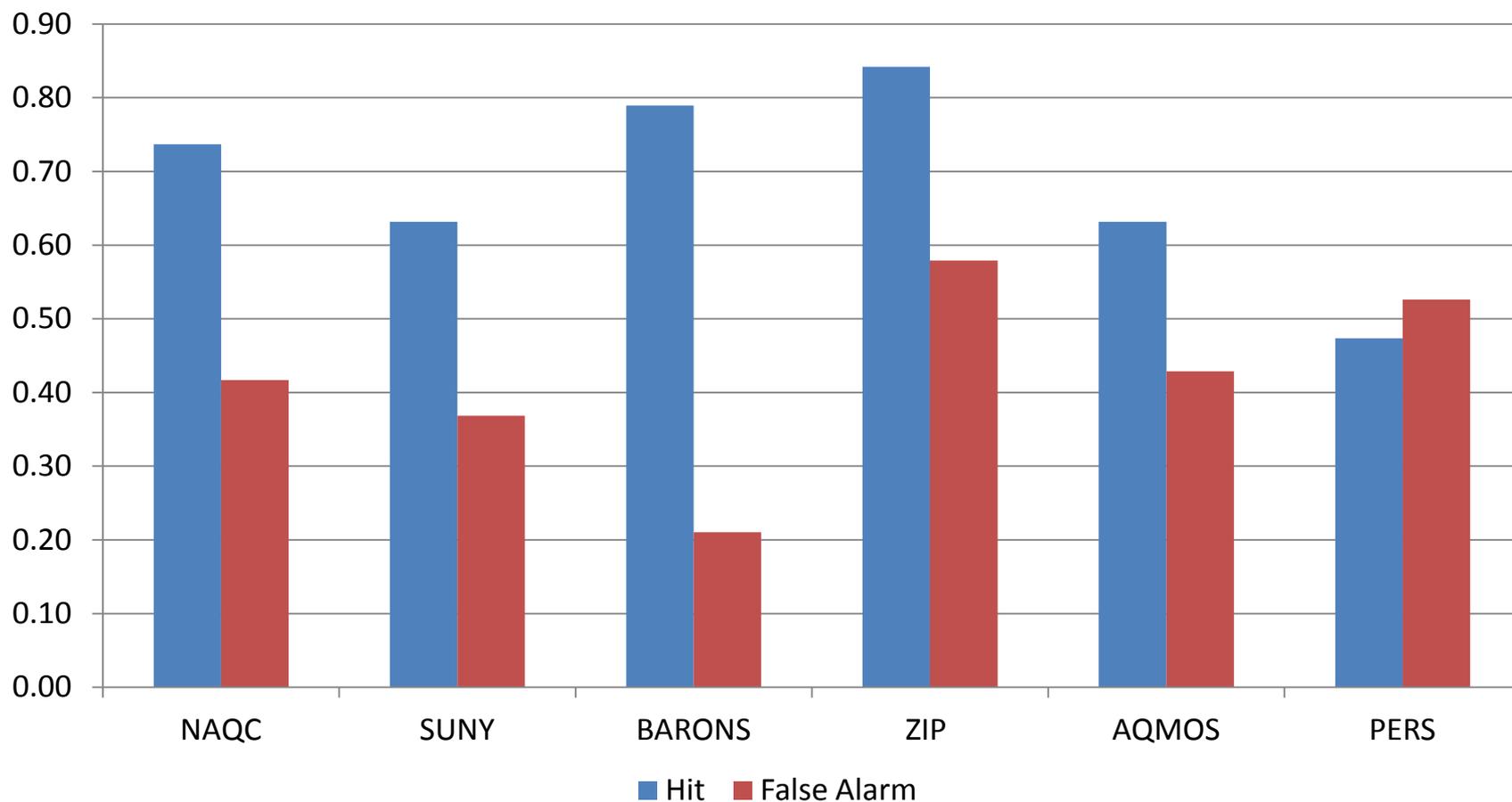
NAQC Forecast Results for 2011 (PHL)

- Bias: + 1.7 ppbv
- $r = 0.75$
- Best Fit:
 - $[O_3]_{OBS} = 18.3 + 0.68 * [O_3]_{NAQC}$
- For Code Orange Threshold:
 - Hit Rate: 0.74
 - False Alarm: 0.42
 - Threat: 0.48

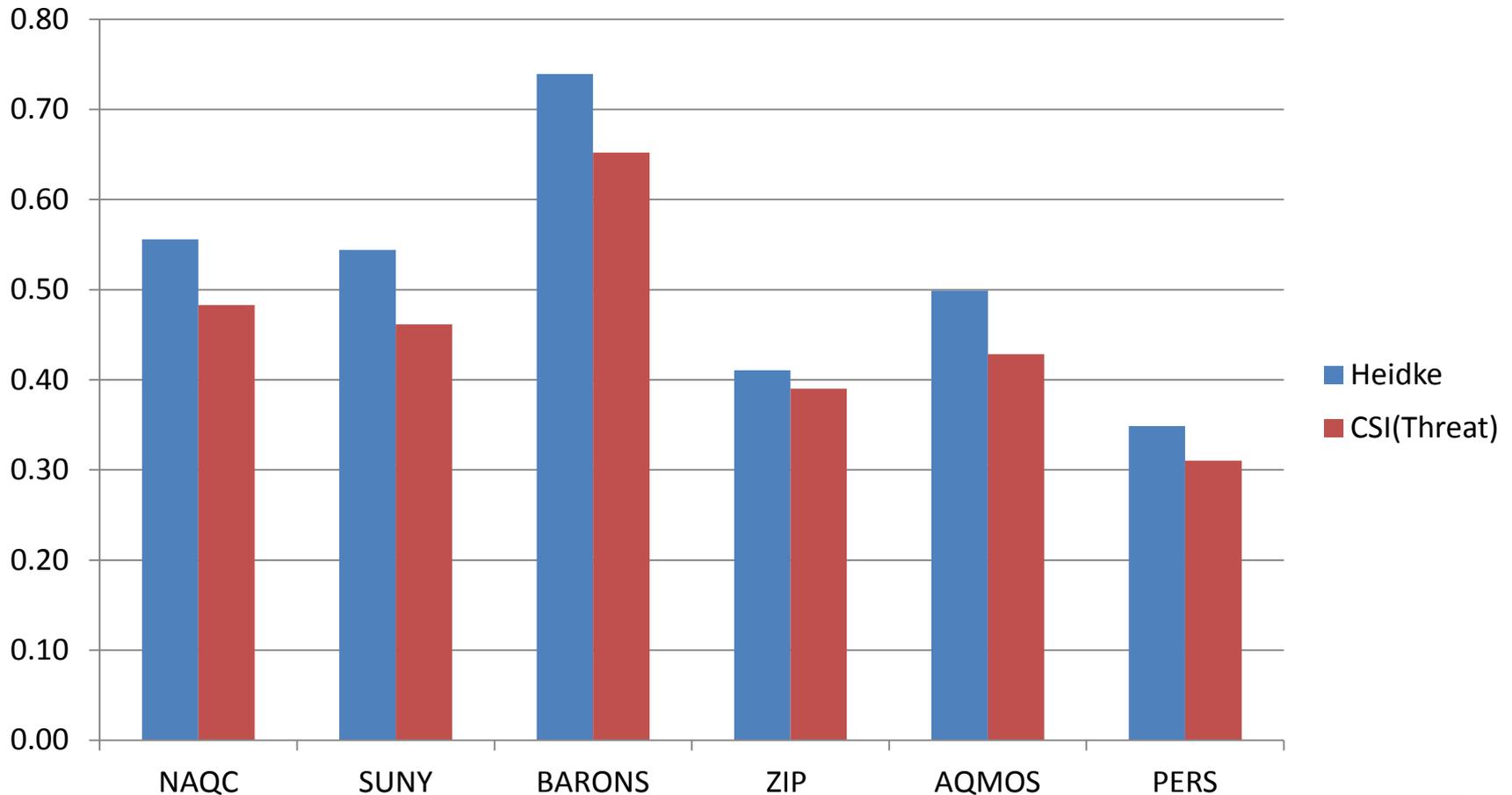


Hit and False Alarm Rates for Individual Models

Threshold: 76 ppbv 8-h Average (Code Orange)

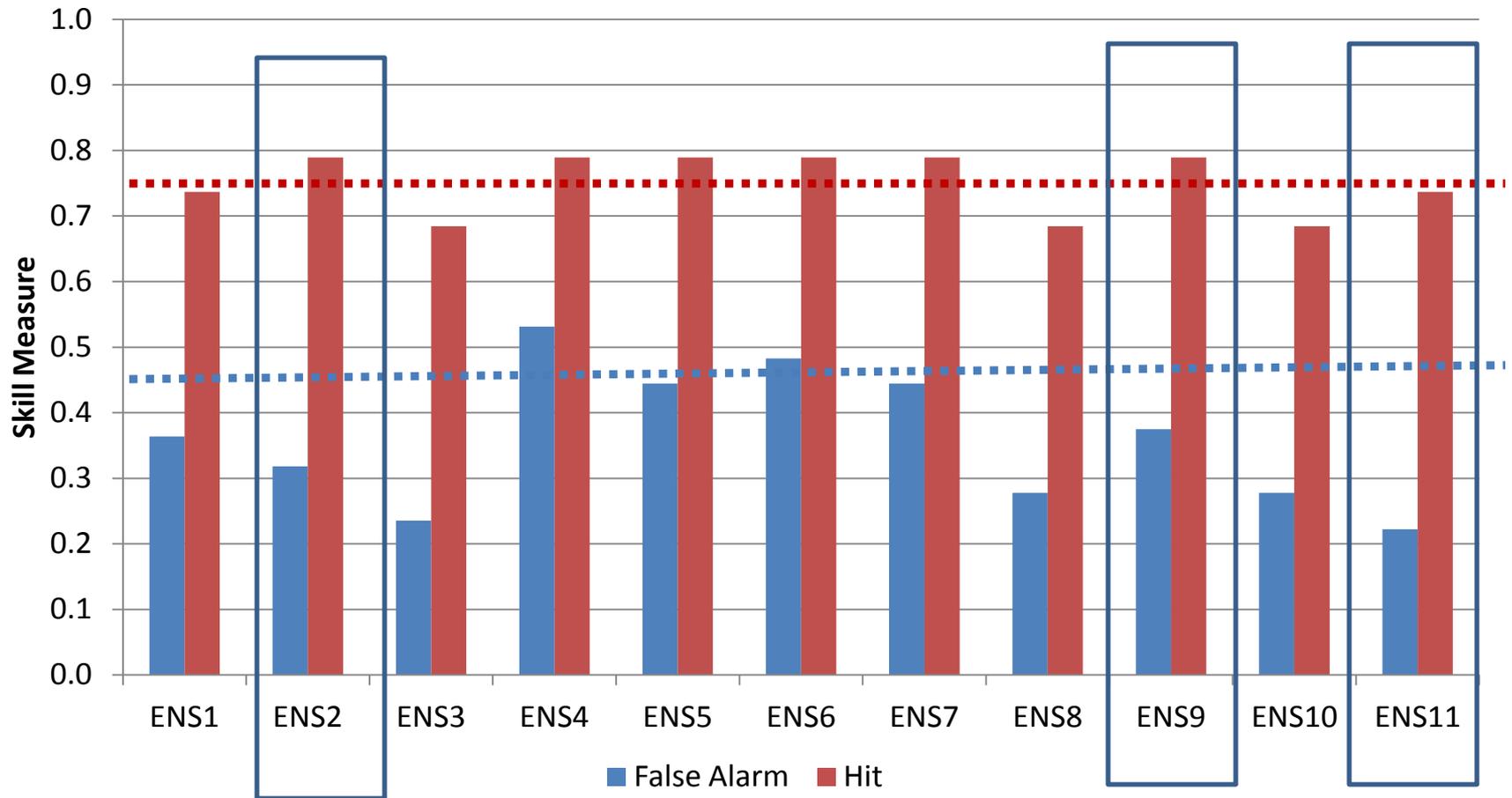


Selected Skill Scores for Individual Models



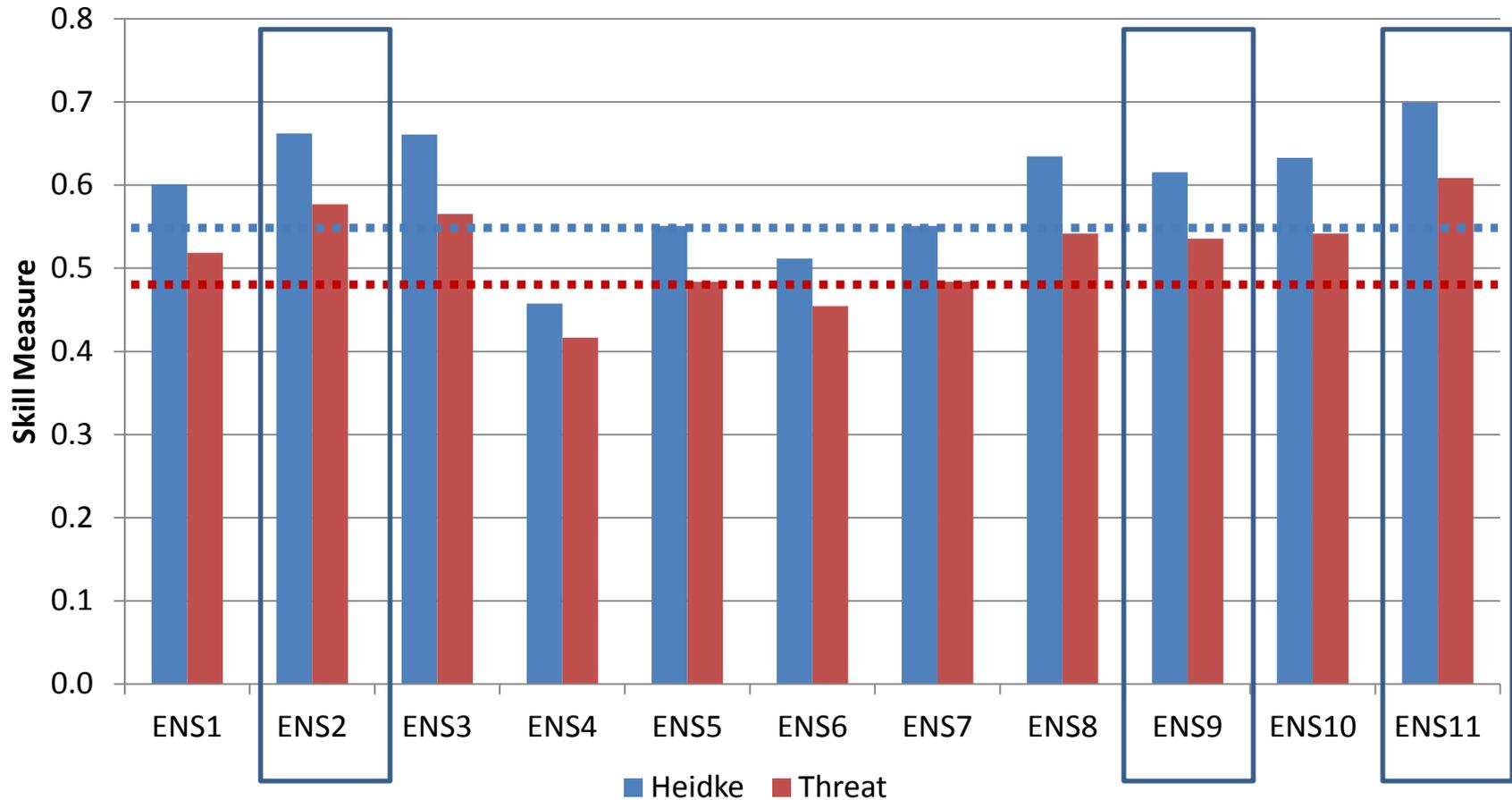
Heidke score measures skill relative to random forecast with range $[-1,+1]$.
Threat score removes influence of “correct null” forecasts, useful for rare events

Hit and False Alarm Rates for Ensembles



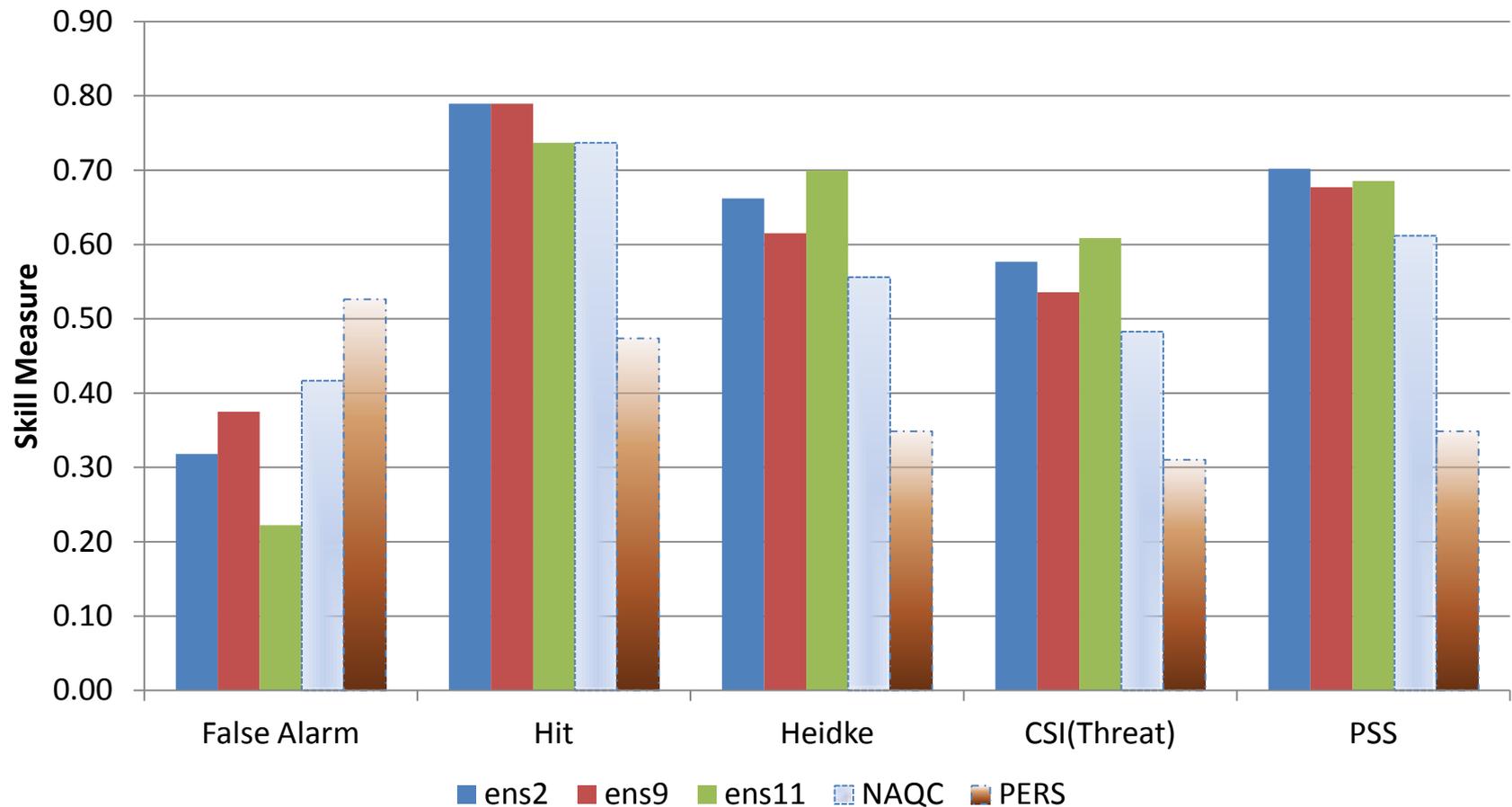
ENS2: All models; ENS9: All models except Barons; ENS11: SUNY, Barons only.
Dashed lines are stand alone NAQC model results.

Selected Skill Scores for Ensembles



ENS2: All models; ENS9: All models except Barons; ENS11: SUNY, Barons only.
Dashed lines are stand alone NAQC model results.

Skill Measures for Selected Ensembles, NAQC Model and Persistence

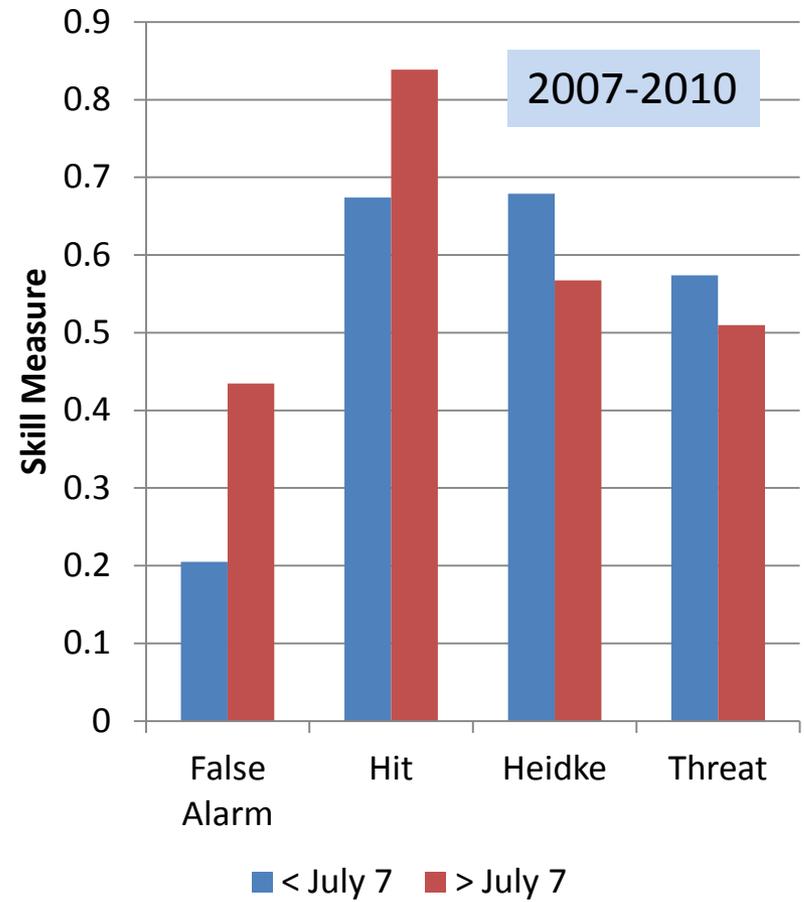
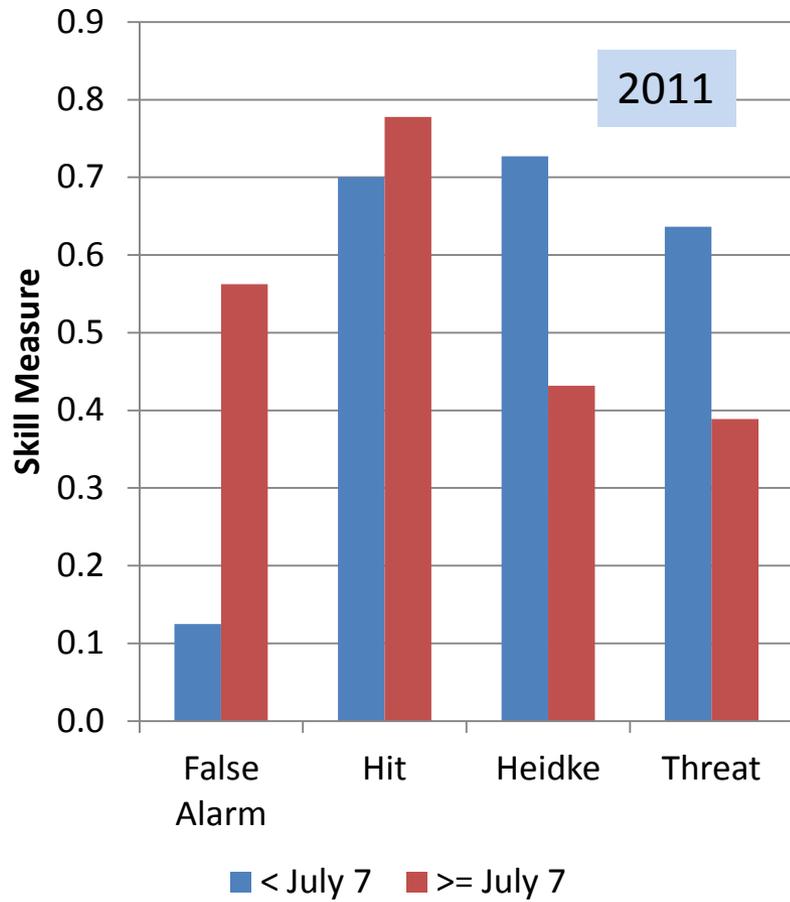


ENS2: All models; ENS9: All models except Barons; ENS11: SUNY, Barons only.
Dashed lines are stand alone NAQC model results.

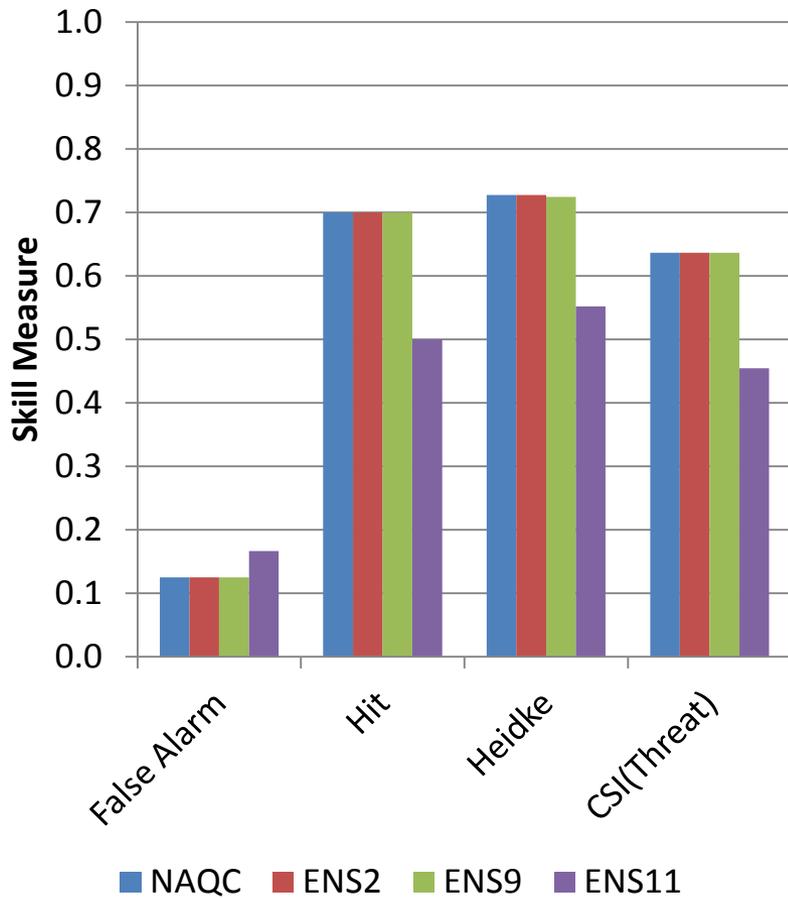
Summary: Overall Ensemble Results

- Two groups performed best
 - Ensembles 1-3: NAQC, SUNY and Barons models with various combinations of NAQC-based forecasts (AQMOS, ZIP)
 - Ensembles 8-10: NAQC and SUNY, with various combinations of the other models.
 - Ensemble 11: SUNY and Barons only.
- Do the ensembles address other shortcomings of the NAQC model?
 - Seasonal drift in bias
 - Weekday/Weekend effects

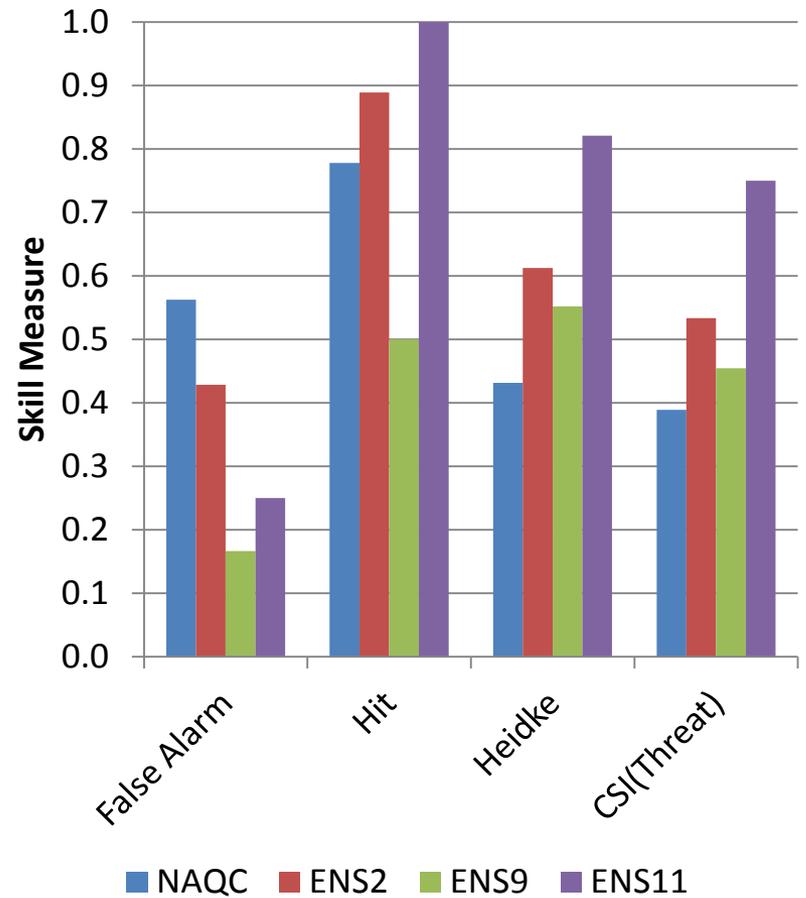
Seasonal Drift: NAQC Skill Scores for Early and Late Summer for 2011 (left) and 2007-2010 (right)



Seasonal Drift: Ensemble Skill

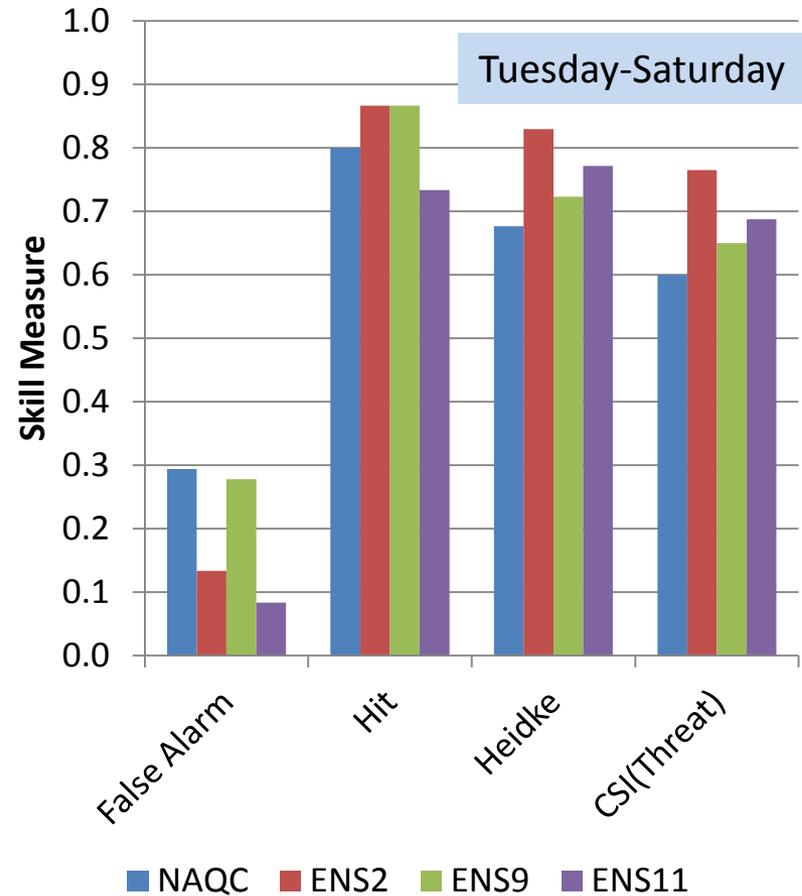
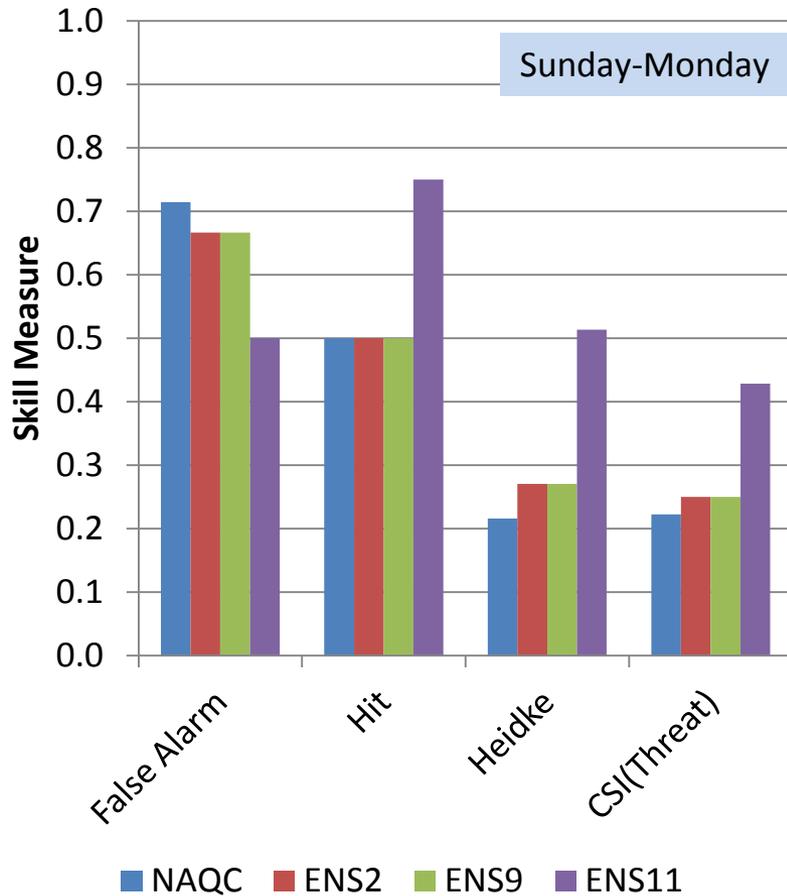


Early Summer (< July 7)



Late Summer (≥ July 7)

Forecast Skill by Day of Week



Conclusions

- A variety of ensemble combinations can improve on the performance of the NAQC model.
- The seasonal drift problem with the NAQC can be partially corrected using non-NAQC models.
- The Sunday-Monday false alarm problem can also be improved with non-NAQC members.

Acknowledgements

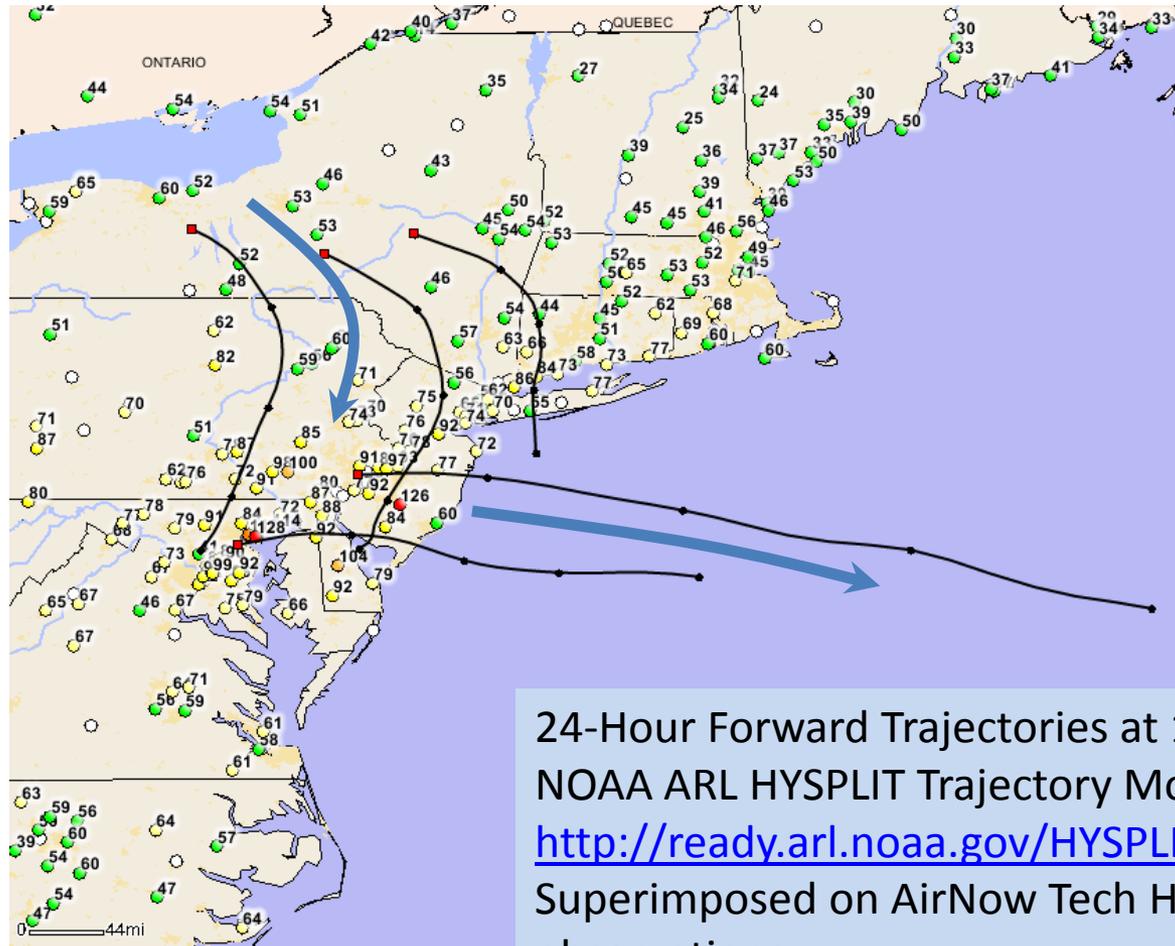
- Funding for air quality forecasting in the Philadelphia metropolitan area is supported by the Delaware Valley Regional Planning Commission (Sean Greene), the Commonwealth of Pennsylvania Department of the Environmental Protection and the State of Delaware.
- Thanks to Dianne Miller (Sonoma Tech), John McHenry (Barons), as well as SUNY-Albany and the New York Department of Environmental Conservation for access to their model forecasts.

Supporting Slides

Skill Score Measures

- Heidke Skill Score
 - Range: [-1,+1]
 - Compares the proportion of correct forecasts to a no skill random forecast that is constrained in that marginal total (hit, miss, false alarm) are equal to observed.
- Threat Score
 - Also known as Critical Success Index (CSI) or Gilbert Skill Score
 - Range: [0,1]
 - Excludes the “null” forecast, so that is a measure of hits divided by sum of hits, misses and false alarms.

Example of Use of Back Trajectories Coupled with Observations (August 10, 2011)



24-Hour Forward Trajectories at 1000 m AGL

NOAA ARL HYSPLIT Trajectory Model

<http://ready.arl.noaa.gov/HYSPLIT.php>

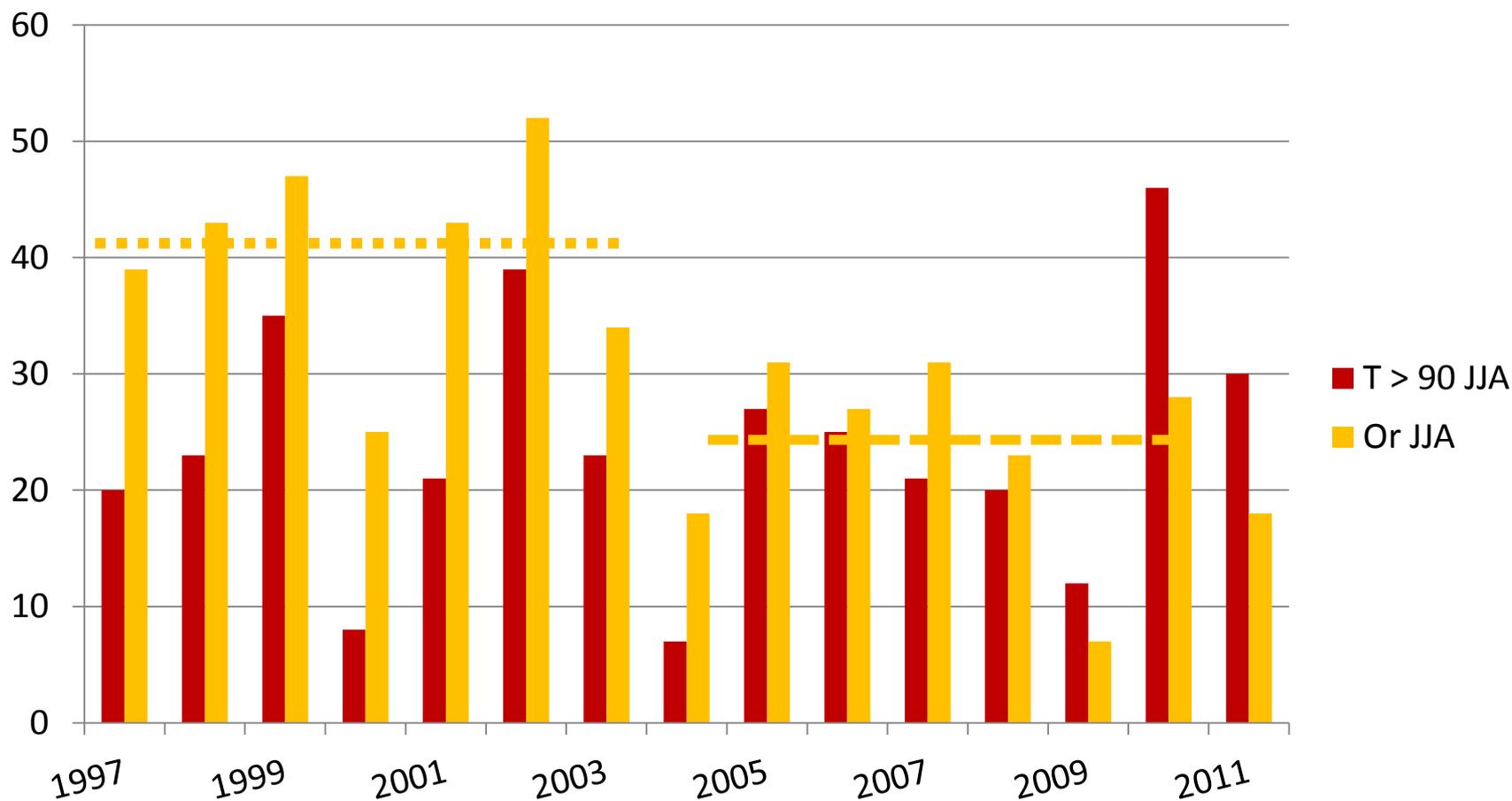
Superimposed on AirNow Tech Hourly O₃ observations:

<http://www.airnowtech.org/>

Weather Summary for Summer 2011

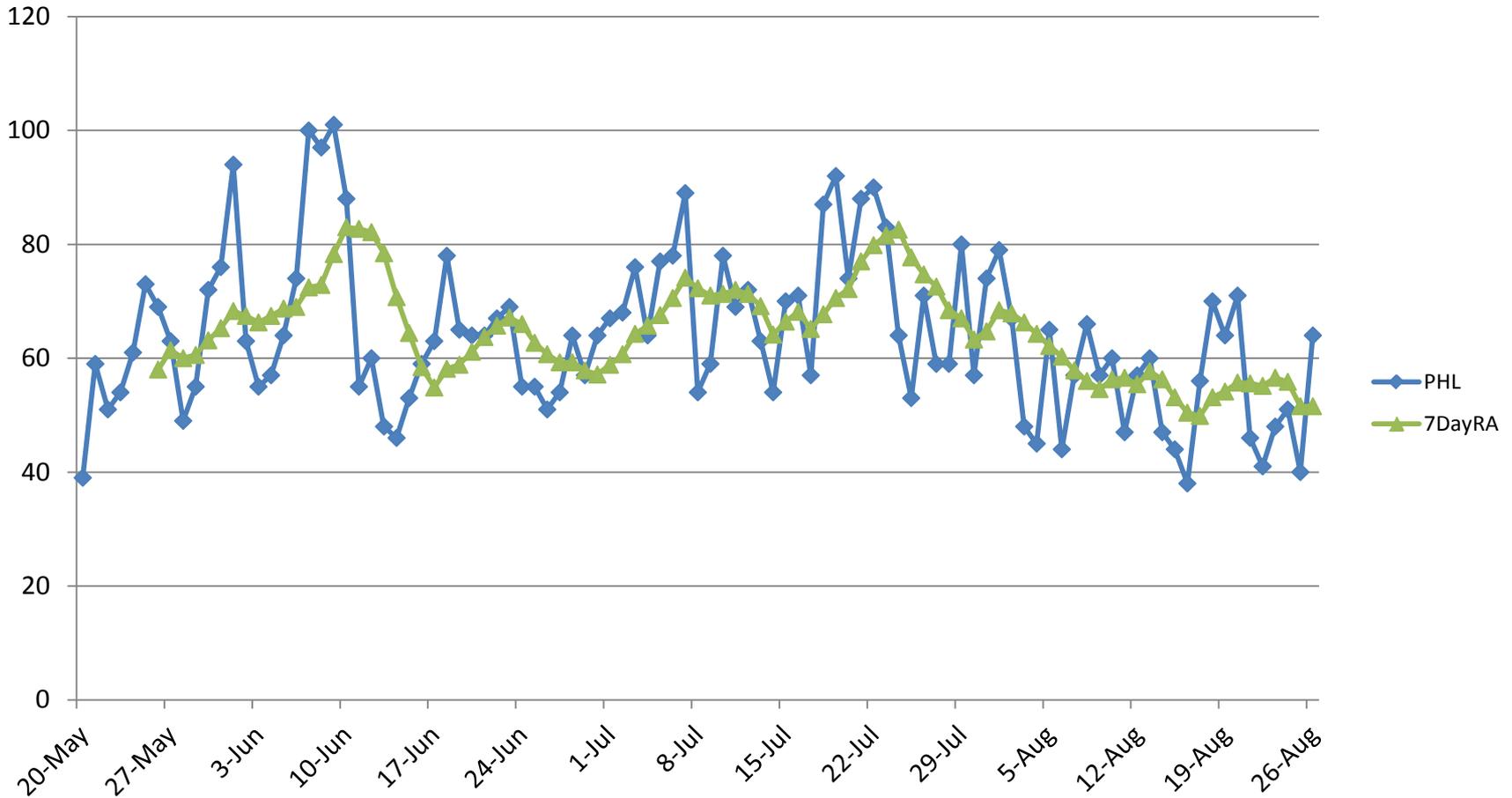
- In Philadelphia, and the mid-Atlantic as a region, 2011 was a warmer than average summer.
 - 30 days $\geq 90^{\circ}$ F compared to average of 23 (1997-2010)
- Warmest temperatures occurred from May 30-August 3 – the heart of the O₃ season – with very wet conditions in August and September.
- In PHL, 19 days were Code Orange or Red compared to an average of 24 (2004-2010).
 - Prior to regional NO_x controls, average of 40 Code Orange or Red days (1997-2003).

Weather Summary for Philadelphia (Summer, 2011): Higher than Average Frequency of Hot Weather

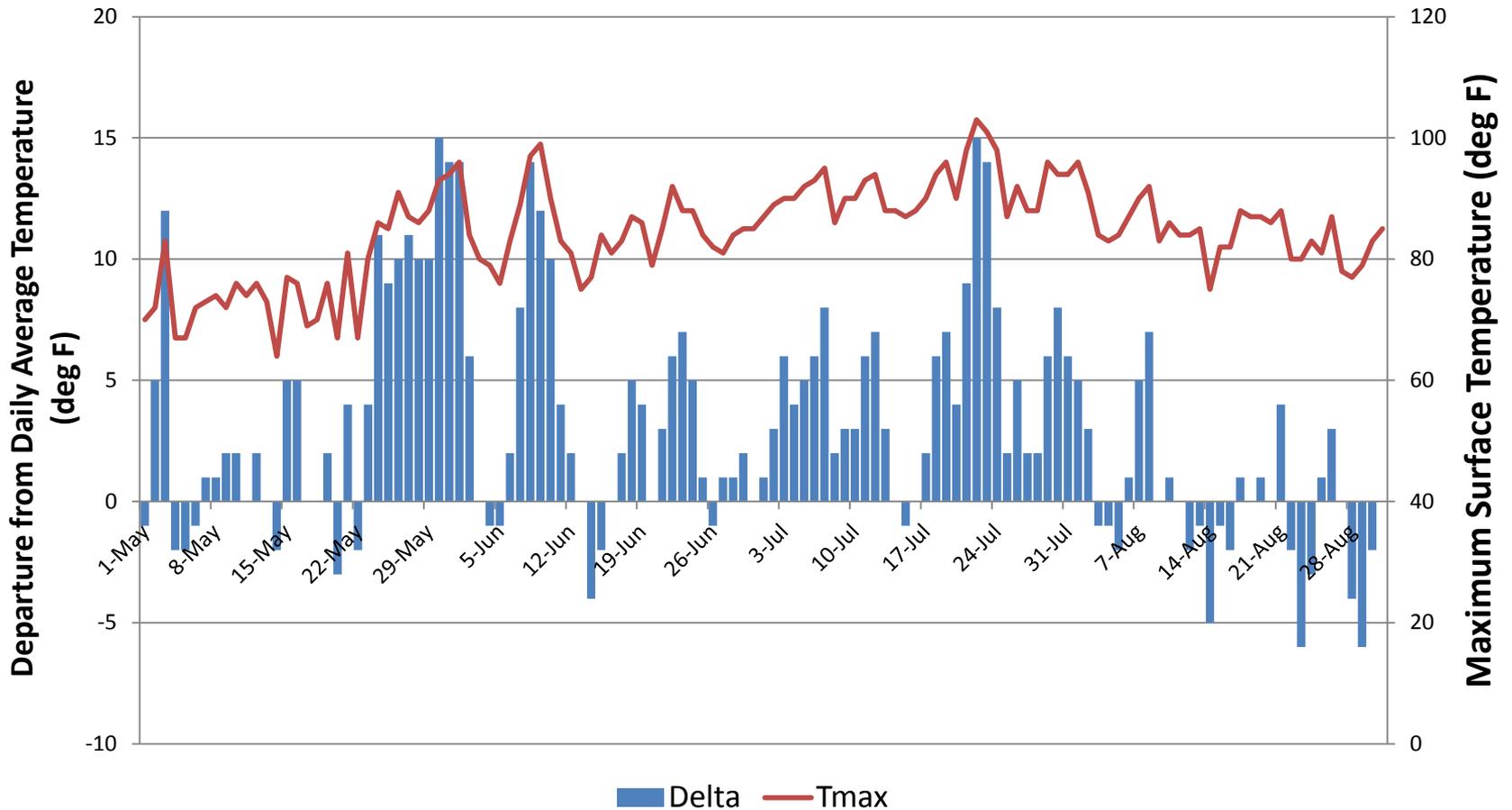


Dotted orange line is mean number of Code Orange days in two periods: 1997-2003 and 2004-2010, reflecting "NO_x SIP Rule" emissions changes in 2002-2003

Daily Peak and 7-Day Running Average for PHL

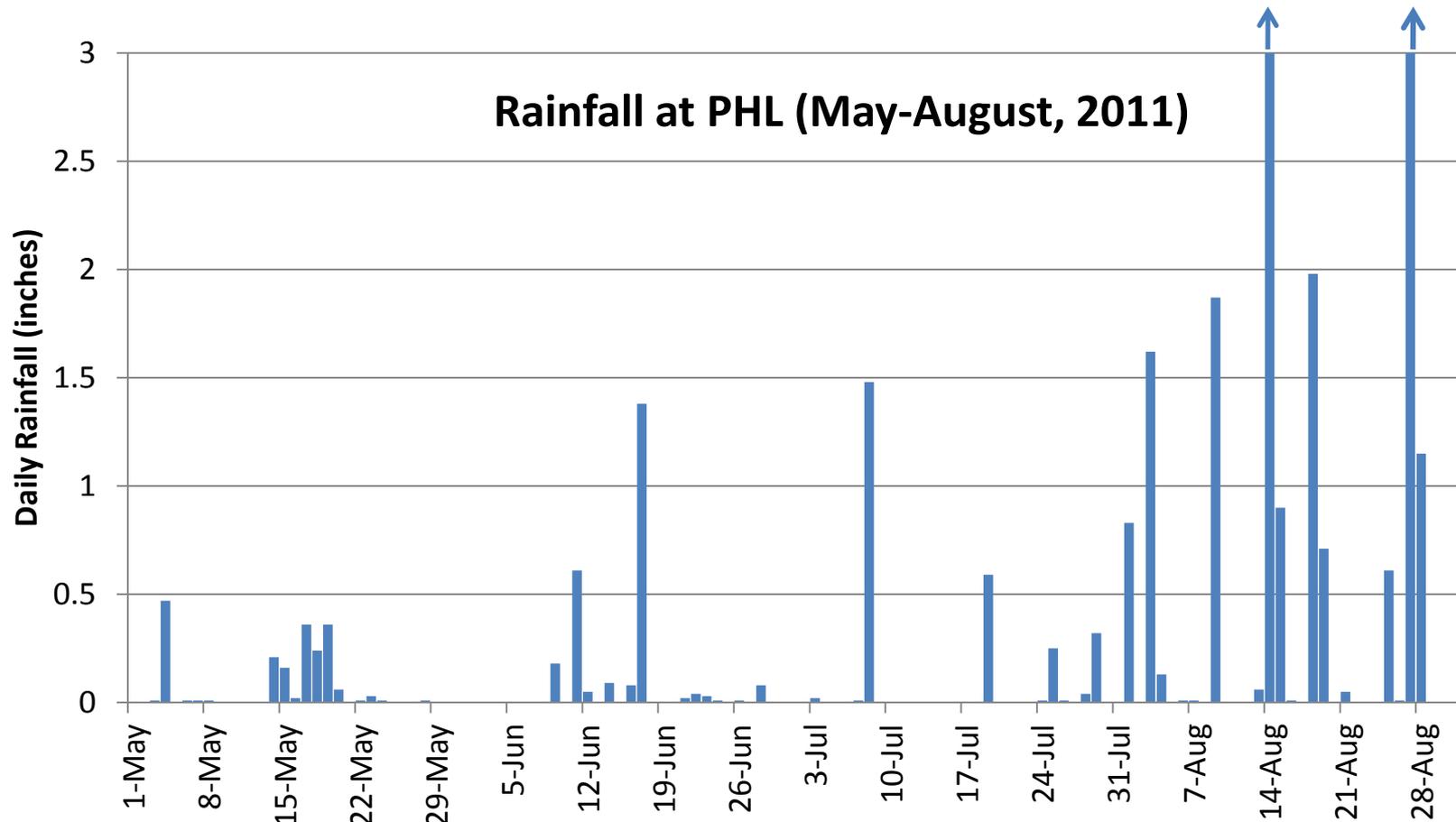


Warmest weather was concentrated during the climatological peak of the O₃ season



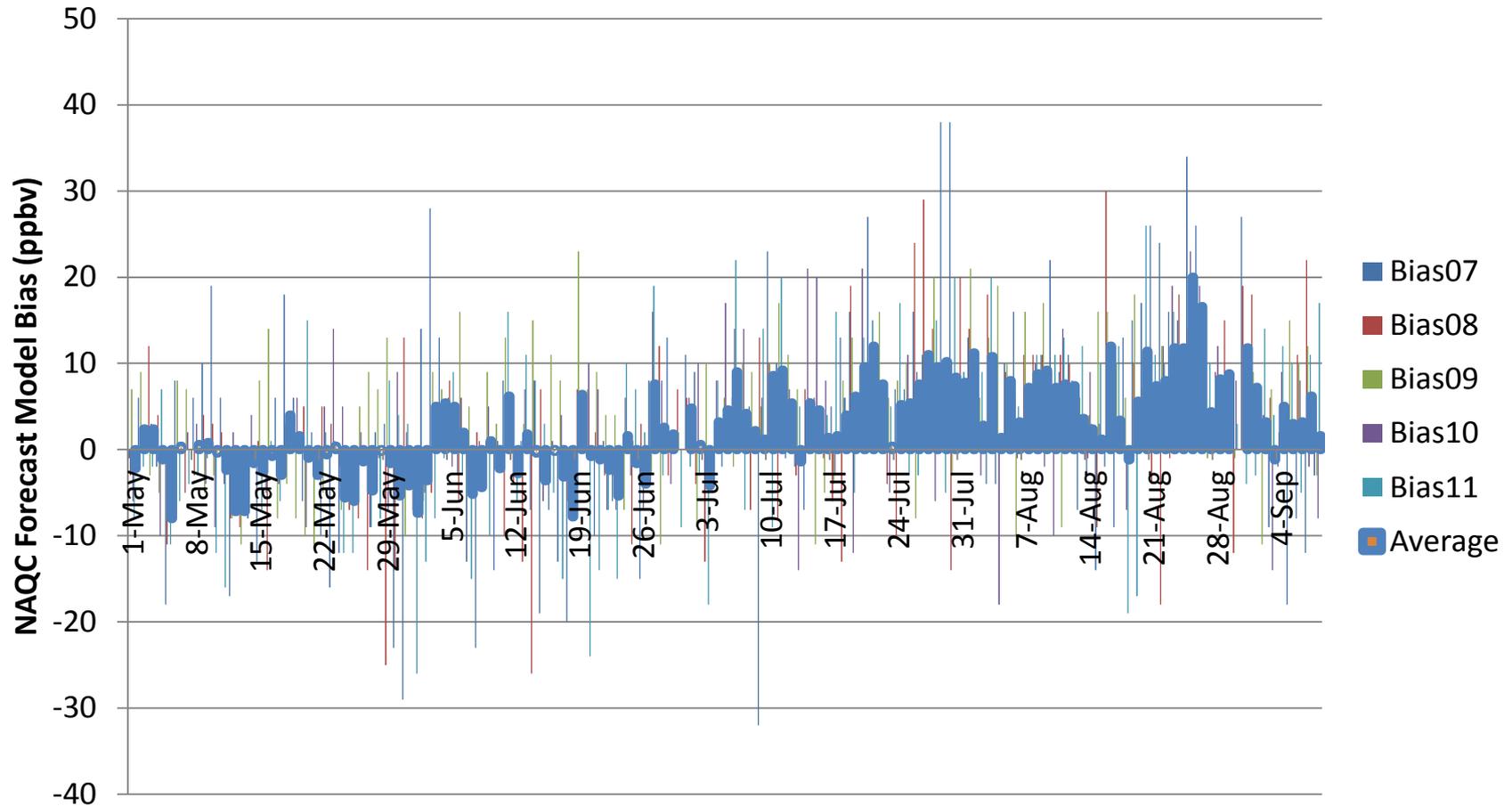
Blue bars are departure from average temperature at PHL, red line is daily maximum surface temperature (deg F).

Ozone conducive weather comes to an end with very wet conditions in August



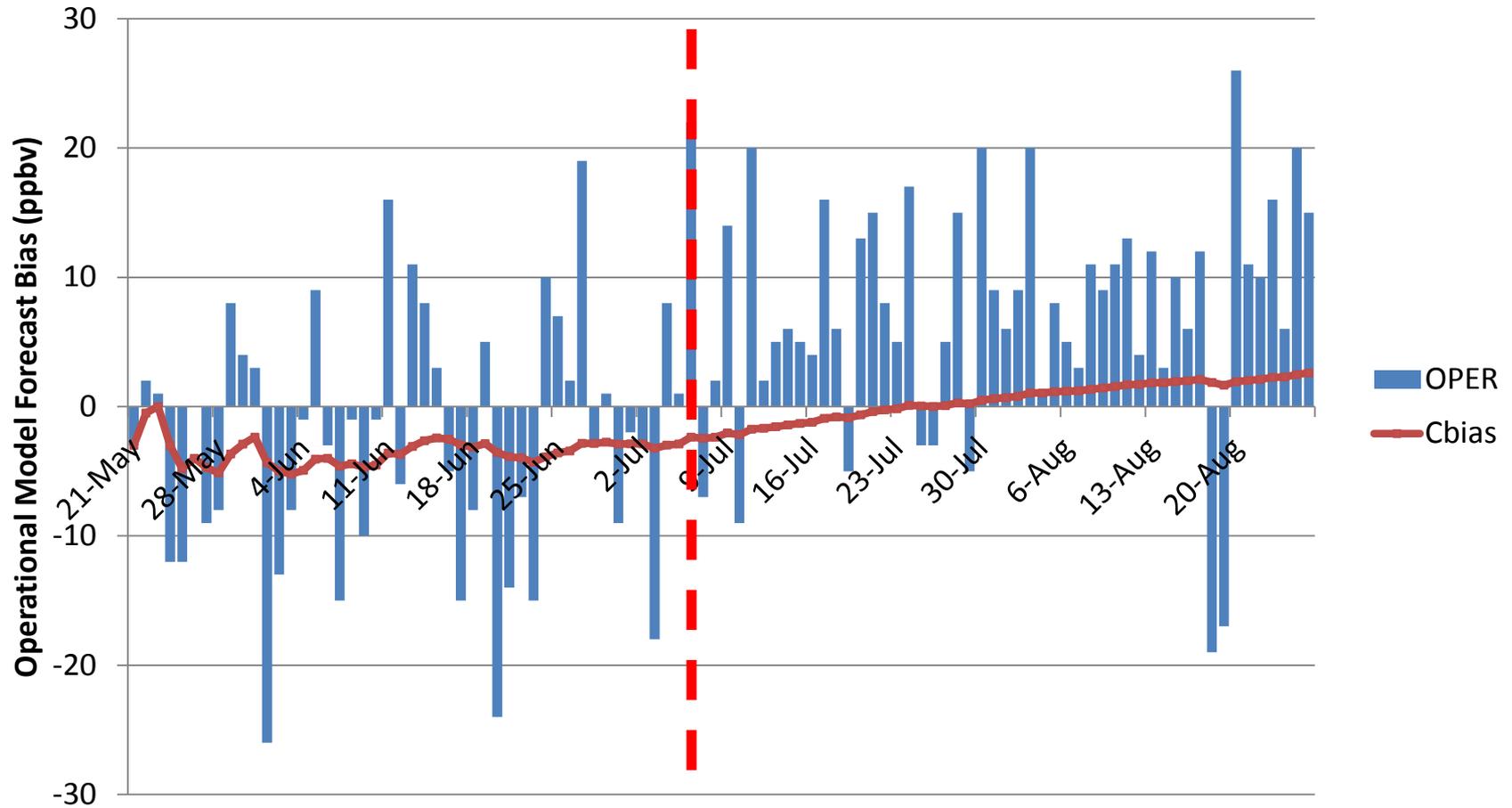
August 15 (4.84") and August 29 (4.55") are off scale, wettest August on record at PHL with 19.21"

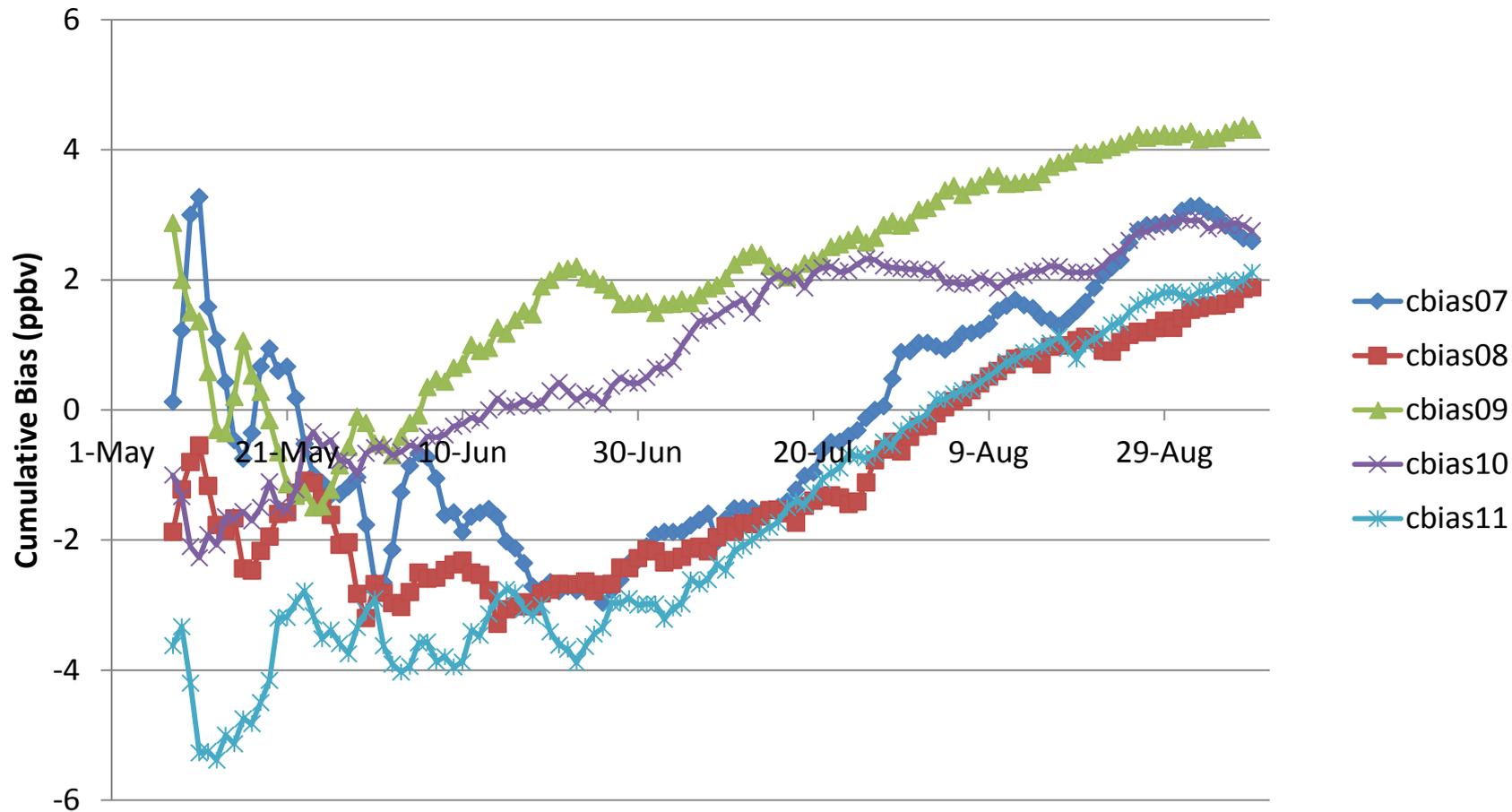
Seasonal Drift in Model Bias NAQC for PHL (2007-2011)

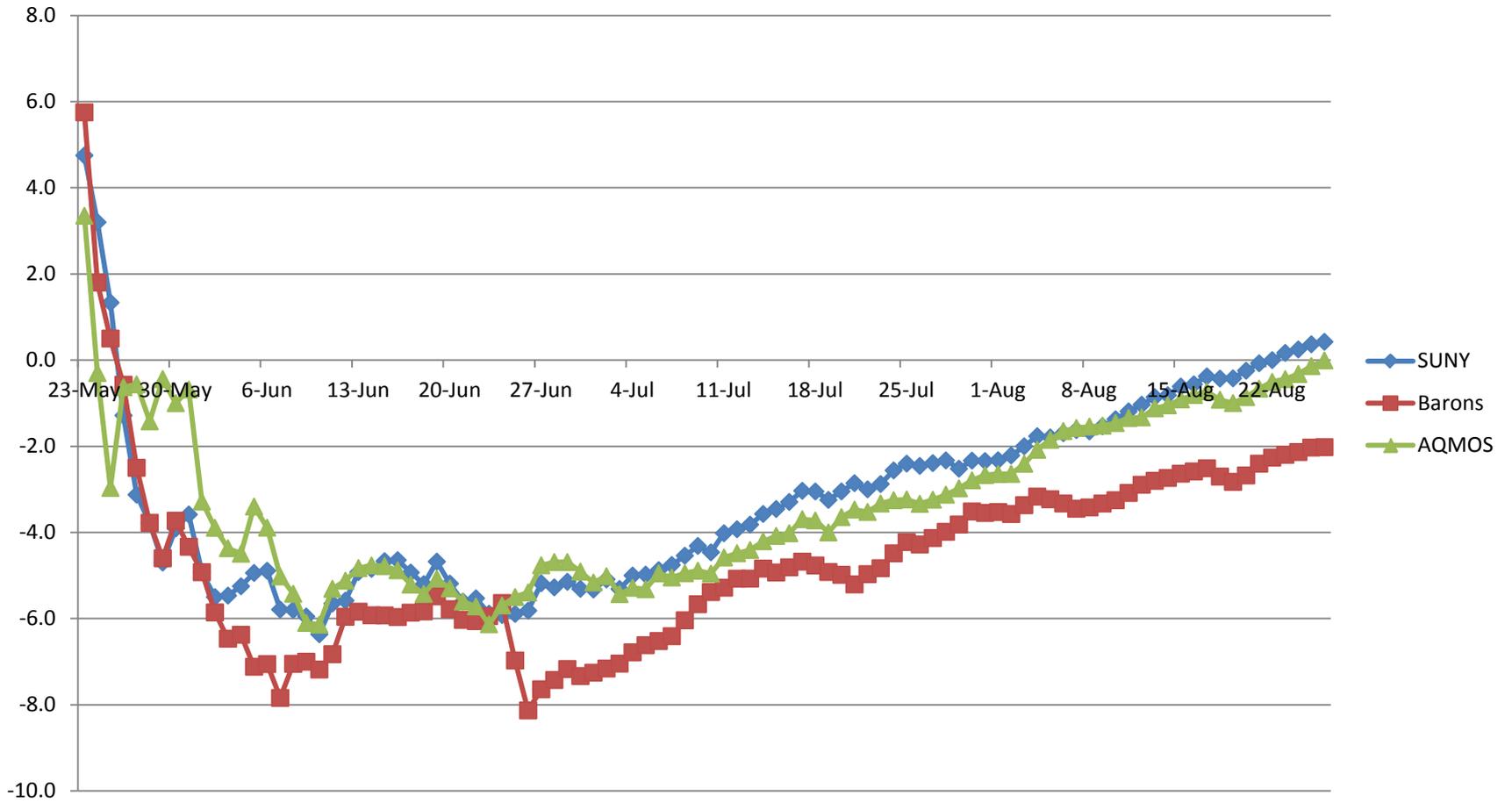


Seasonal Drift in Model Bias

NAQC for PHL, 2011







AQMOS

