Data fusion applications in the Air Quality Modeling Group

Joey Huang, Barron Henderson, Janica Gordon, Heather Simon, Tyler Fox
EPA/OAQPS/AQAD/AQMG
Huang.Jiaoyan@epa.gov
Henderson.Barron@epa.gov

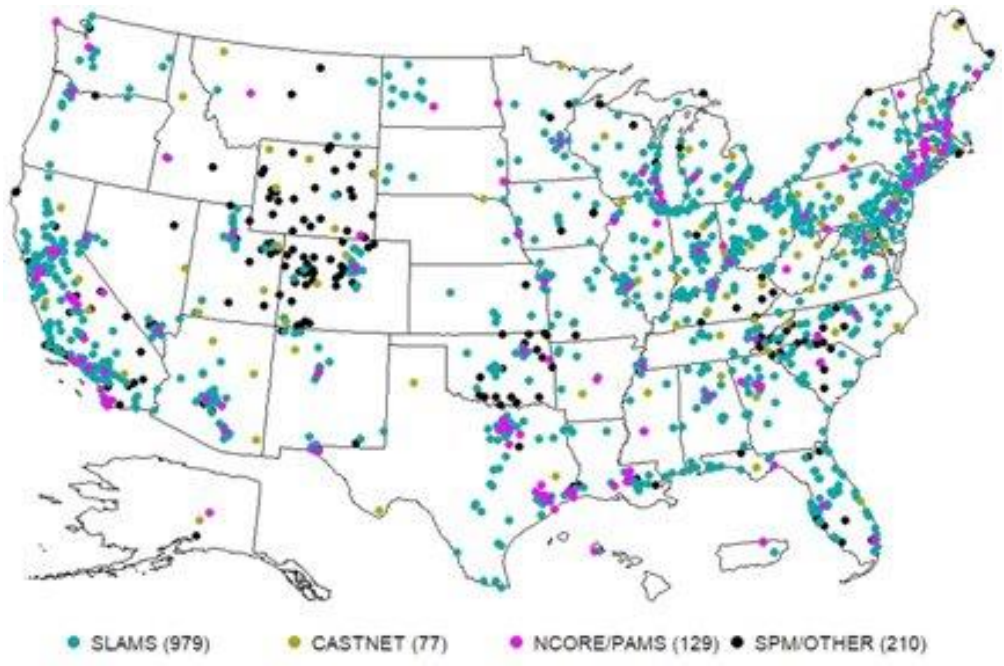
USEPA Air Sensor QA Workshop
July 2023, Durham, NC

Outline

- Why is data fusion important?
- How does the Air Quality Modeling Group use data fusion?
  - NAAQS Review Components
  - Retrospective Analysis
  - Future Year Projections
- Fusing Models and Observations for AirNow
  - Residual Kriging
  - EPA Traditional Approaches and Possibilities
- Summary

Why data fusion?

- Monitors tell us what is, but are limited in space, time, and composition.
- Models can provide complete coverage, but are limited by our ability to replicate processes in the atmosphere.



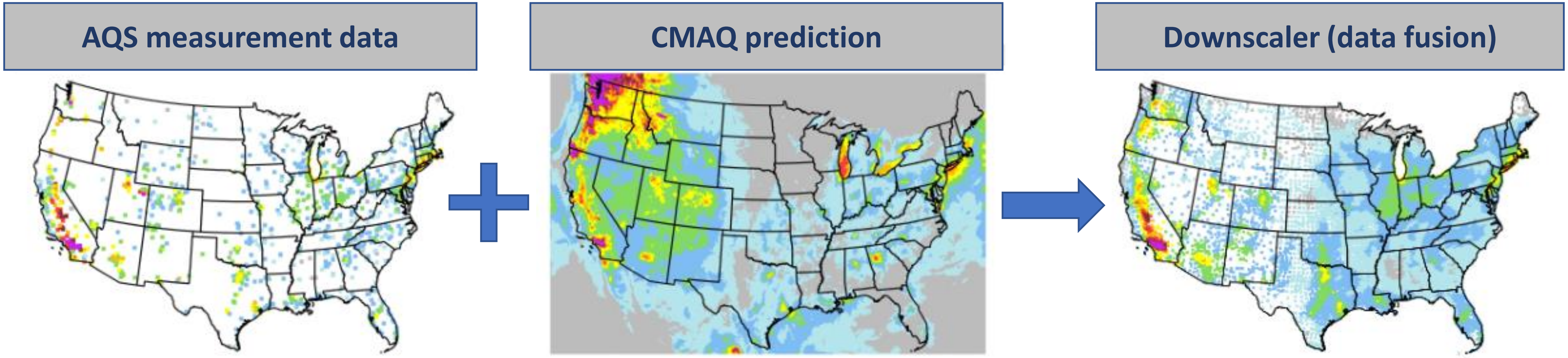
NAAQS Review Components and Data Fusion

Voronoi Neighbor Averaging: VNA
Enhanced Voronoi Averaging: eVNA

ISA: Integrated Science Assessment	REA: Health and Welfare Risk and Exposure Assessments	PA: Policy Assessment	RIA: Regulatory Impact Assessment
• Assesses the most policy relevant scientific evidence from health studies and draws weight-of-evidence conclusions for causality determinations • As part of the review of the overall body of scientific evidence, the ISA identifies at-risk populations and draws conclusions based on strength of evidence for health effects for the entire population, including at-risk groups.	• current estimates of air quality throughout the U.S. • Health REA assesses population exposures and health risks associated with recent ambient concentrations and with concentrations adjusted to simulate just meeting the current standard and potential alternative standards • Welfare REA assesses vegetation and ecosystem exposures and risks associated with recent ambient concentrations and with concentrations adjusted to simulate just meeting the current standard and potential alternative standards.	• PA presents and assesses the range of policy options that could be supported by the available scientific evidence and exposure/risk information. • The PA brings together the available scientific evidence, as assessed in the ISA, and exposure/risk information from the REA	• Future model projections that account for projected air quality changes throughout the US • Assesses the costs and benefits of attaining proposed alternative standard levels. Benefits derived from epi-based health improvements.
• ***Data fusion included in assessed literature***	• ***VNA*** *used for urban scale hourly fused surfaces for health assessments* • ***Downscaler*** *used for national seasonal fused surfaces for health assessment* • ***VNA*** *used for national seasonal fused surfaces for welfare assessment*		• **eVNA** used for national seasonal fused surfaces

Retrospective Analysis: CDC Phase Project

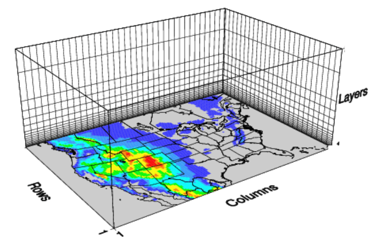
- For over a decade, EPA has developed an annual platform to characterize national surfaces of O_3 and $PM_{2.5}$ in collaboration with the CDC
  - CMAQ model output and measurement data are combined to create a fused surface that has better spatial coverage than monitors alone and less uncertainty than model data alone
  - Data are intended to help explore the association between environmental exposures and health impacts
  - Ozone and PM2.5 fused fields and associated documentation are currently available for 2002-2019 at https://www.epa.gov/hesc/rsig-related-downloadable-data-files#faqsd



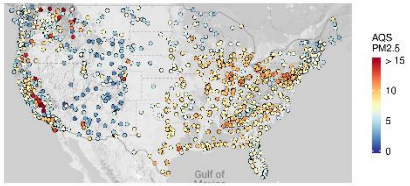
Community Multiscale Air Quality Modeling: CMAQ
Air Quality System: AQS
Centers for Disease Control: CDC

Future Year Projections:
Exposure Disparities for PM2.5 for 2011 and 2028

CMAQ Modeling



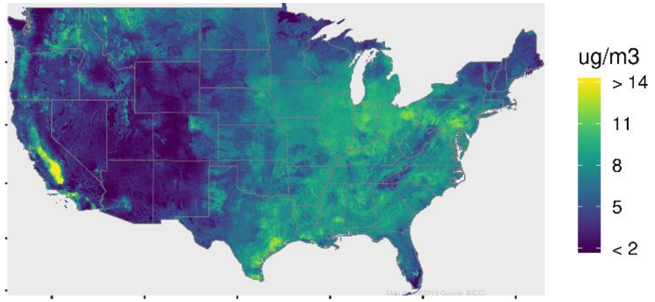
Monitoring



Remote Sensing

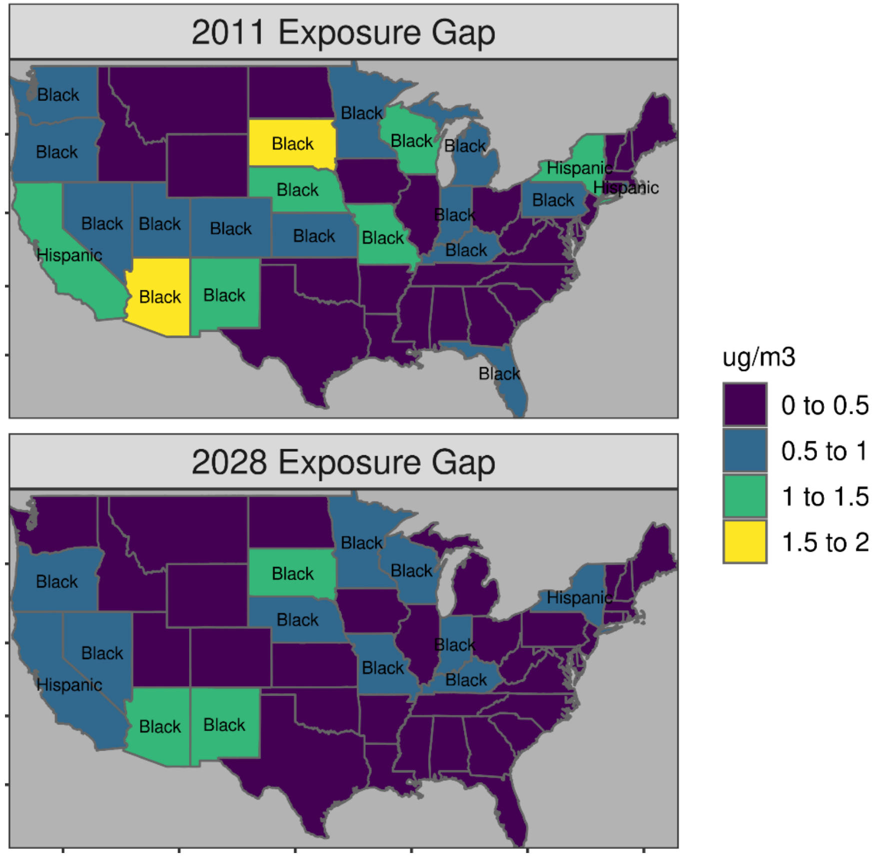
Other Data

Fused $PM_{2.5}$ Field



Project w/ Future "On-the-Books" AQ Modeling

Exposure Gap* by State Decreases from 2011 to 2028 but Most-Exposed Groups Persist

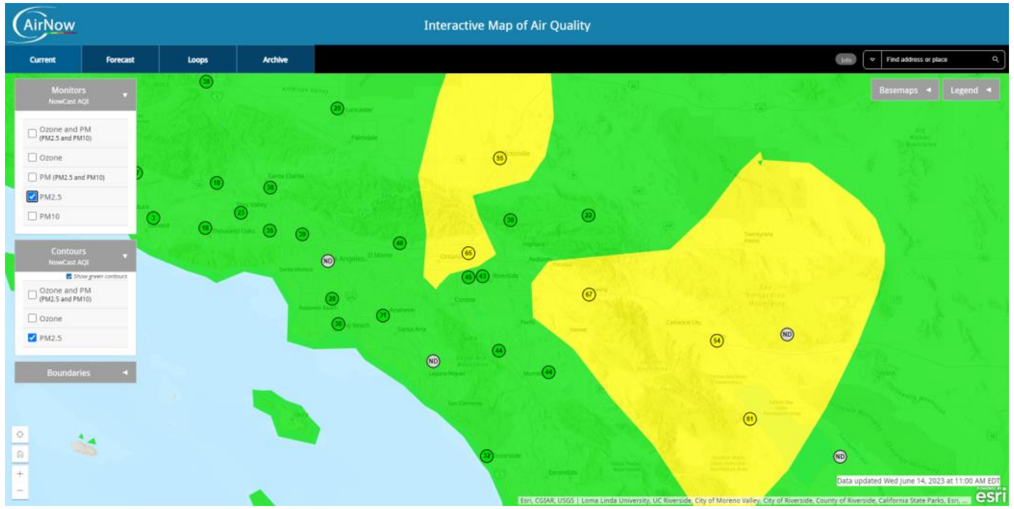


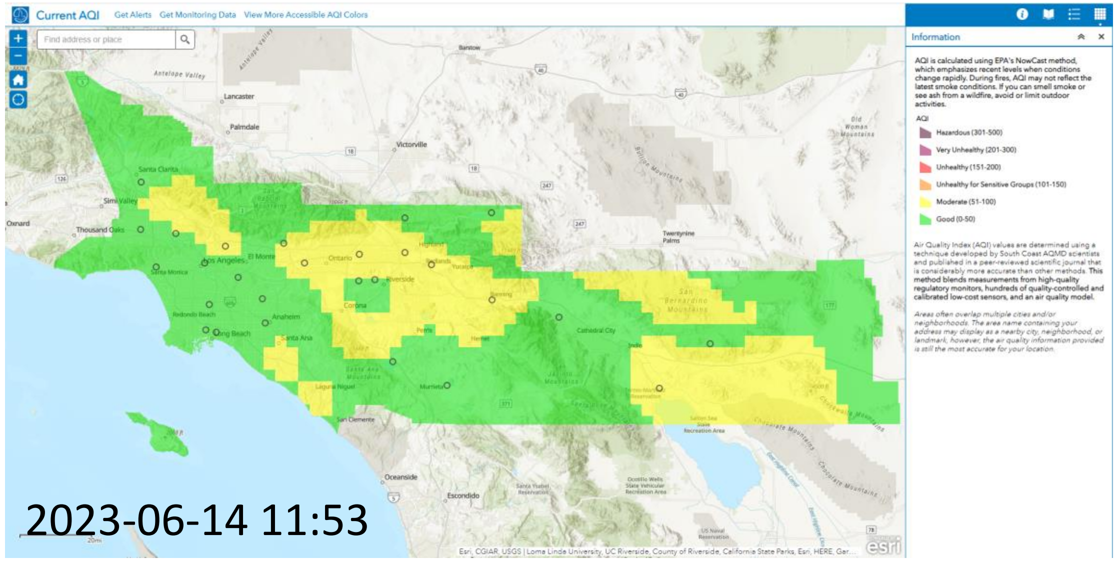
[Kelly et al. (2021)](https://doi.org/10.1016/j.envres.2020.110432) *Environmental Research* (https://doi.org/10.1016/j.envres.2020.110432)

*Exposure gap is defined here as the difference in population-weighted concentration between the most and least exposed group (but could be defined differently, e.g., gap between low-income non-white and high-income white)

Data Fusion for AirNow

- AirNow provides a map using inverse distance weighting* of mostly regulatory grade monitors.
- PurpleAir sensors have dramatically increased in prevalence
  - Widely increased the spatial coverage of monitored particulate matter.
  - Provides a measure where regulatory monitors are not.
- South Coast Air Quality Management District demonstrated that integrating PurpleAir improved their air quality estimates compared to inverse distance weighting.



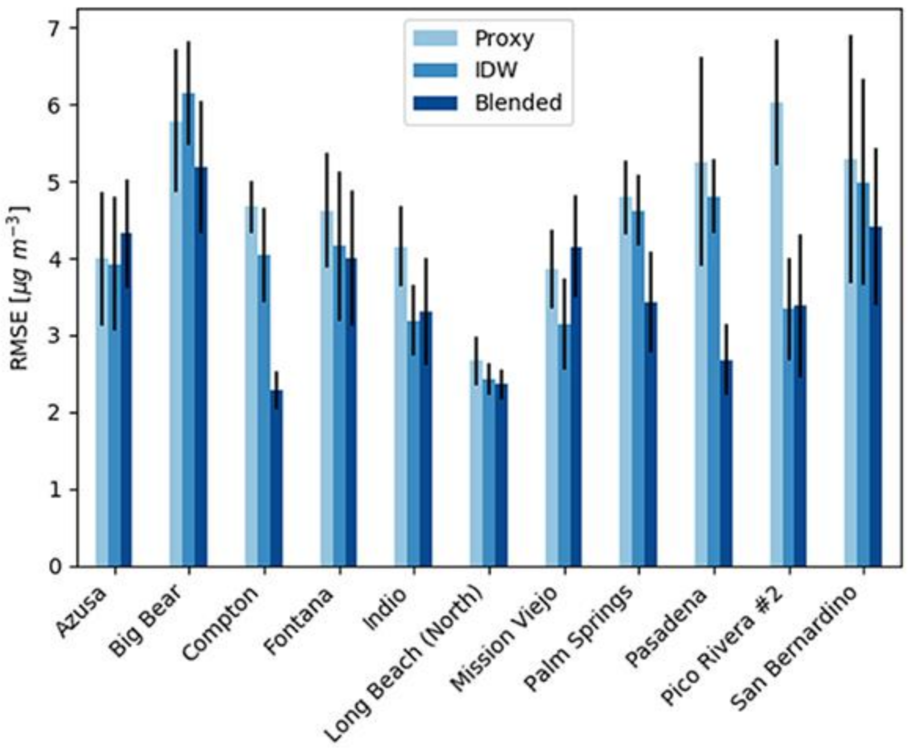


*10 nearest neighbors and weight ~ d^{-5}

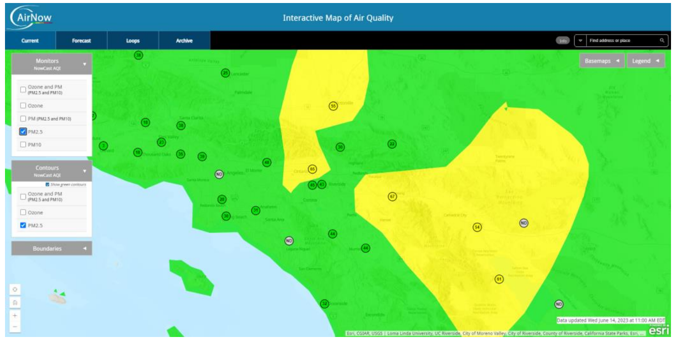
“South Coast” better than interpolation

- Schulte et al showed Residual Kriging had better performance than inverse distance interpolation or surrogate monitor.
- Residual Kriging is a way of interpolating model bias and then removing that bias from model.
  - Model: NOAA Air Quality Forecasts Capabilities
    - CMAQ initialized twice daily informed by EPA inventories
    - Twice a day hourly ozone and PM25 predictions
  - $Bias_n = Model_n - Observation_n$
    - Federal Equivalent Method hourly Ozone and PM25
    - PurpleAir averaged to hourly outputs
      - Corrected to FEM and Averaged to 5km grids
      - Aggregate is a “pseudo-station”
  - $Y = Model - Krig(Bias_n)$
    - Simple Kriging requires a semi-variogram
    - Variogram corrected for PurpleAir error correlation.

Schulte, N., Li, X., Ghosh, J. K., Fine, P. M., & Epstein, S. A. (2020). https://doi.org/10.1088/1748-9326/abb62b



EPA Traditional Approaches and Possibilities



- Universal Kriging is a good option, but are there tools we use at EPA that might be better?
- AirNow: $Y = sum(O_n * w_n)$
  - n in 10 nearest neighbors; $w_n = d^{-5}$
  - ***Super fast and super simple.***
- Downscaler
  - Hierarchical Bayesian Model (Berrocal et al. 2010, 2012) used for CDC PHASE project
  - ***Slower and complex – too slow for this application***
- eVNA: $Y = M * sum(O_n / M_n * w_n)$
  - Unmonitored Area Analysis and RIA
  - Interpolates Voronoi neighbors' multiplicative bias correction with weights = d^{-2}
  - ***Medium complexity, but very fast.***
- aVNA = $M + sum((O_n - M_n)*w_n)$
  - Simple reformulation of eVNA to apply additive bias (more like Residual Kriging)
  - ***Medium complexity, but very fast***

Should we apply eVNA or aVNA to pooled PurpleAir and AirNow obs?

Voronoi Diagram

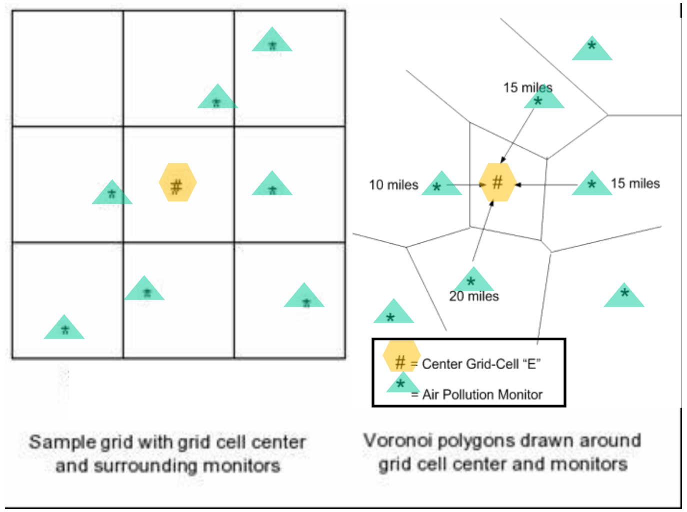
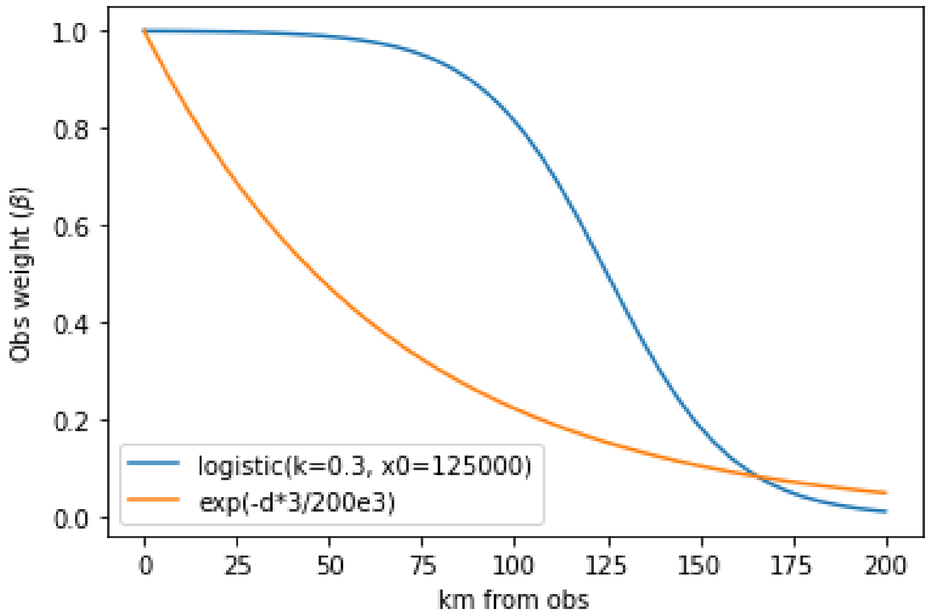
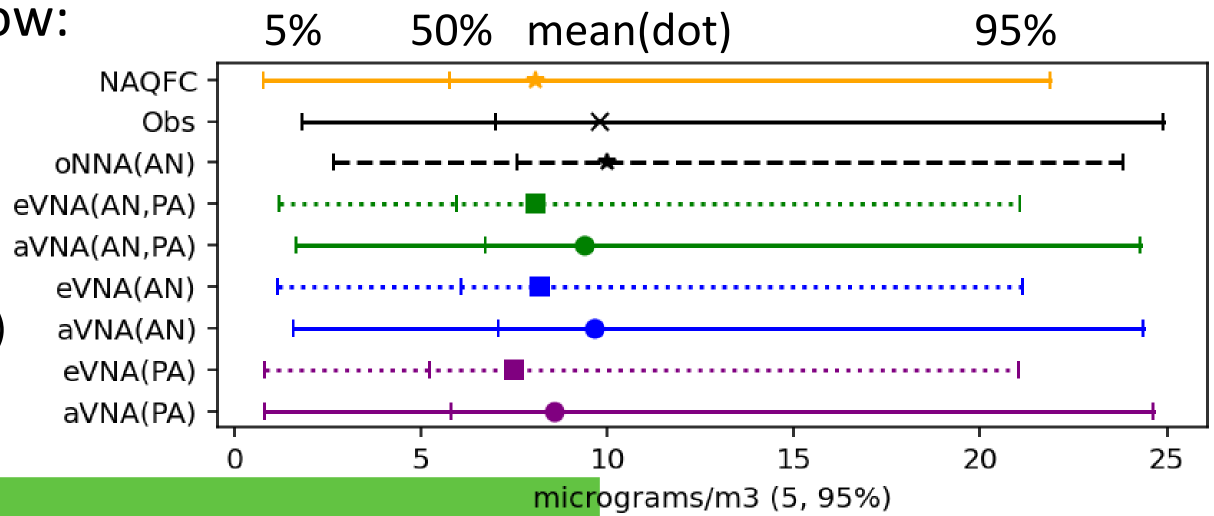


Figure courtesy of: Brian Timin

Separate fusion and estimate blending

- Not pooling data because of differential quality.
  - Bi et al. (2020): PurpleAir monitors down-weighted (0.23x) in a Random Forest model to preserve model performance.
  - Pooling PurpleAir with FEMs would ignore this.
- Ensemble Blending of NOAA and both fusions (Y_{PA}, Y_{AN})
  - $Y = \beta(\alpha_{PA}Y_{PA} + \alpha_{AN}Y_{AN}) + (1 - \beta) * Y_{NAQFC}$
  - $\alpha_{PA} = 0.25\ d_{PA}^{-2} / (0.25\ d_{PA}^{-2} + d_{AN}^{-2})$
  - $\alpha_{AN} = d_{AN}^{-2} / (0.25\ d_{PA}^{-2} + d_{AN}^{-2})$
  - β = see right figure
- National-scale annual cross-validation results show:
  - AirNow only or PurpleAir each marginally out-performs interpolation.
  - Combining AirNow and PurpleAir
    - Overall, quite good.
    - Improves root mean square error (*very good*)
    - Reduces variance compared to observations. (*less good*)





Summary and next steps

- Data fusion has capability to present air quality spatial variation between real atmosphere and modeling results which is important for regulatory review.
- Monitors, satellites, and models with data fusion tool can provide detailed air quality for environmental justice analysis.
- aVNA with AirNow ***and PurpleAir*** data has the best performance
  - Continue internal review
  - Anticipate a limited access roll-out for review by AirNow partners
  - Potentially roll-out to broader community

Appendix: Photochemical Modeling in the Risk and Exposure Assessment

- The national risk assessment requires a spatial field of pollutant concentrations covering the entire country
  - Ozone: seasonal average of 8-hr max, 8-hr block and 1-hr max; W126
  - PM2.5: annual average

- Fused fields created using enhanced Voronoi neighbor averaging (eVNA)
  - VNA: interpolation technique which uses inverse-distance-weighted averaging: monitor data

$$Species_{E,\,baseline} = \sum_{i=1}^{n} Weight_i \cdot Monitor_i$$

  - eVNA: supplements VNA with model data to adjust concentrations between monitors.
    - VNA concentrations are multiplied by the ratio of the modeled concentrations at the grid cell divided by the weighted average of the model concentration at the nearest neighbor monitor locations
    - Modeled spatial gradients are preserved
    - This ratio = 1 at the location of the monitor

$$Species_{E,\,baseline} = \sum_{i=1}^{n} Weight_i \cdot Monitor_i \cdot \frac{Model_{E,\,baseline}}{Model_{i,\,baseline}}$$

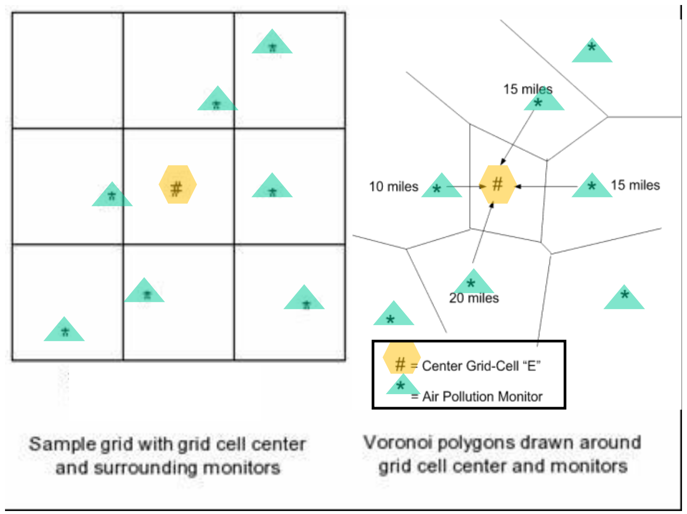


Figure courtesy of: Brian Timin

Appendix

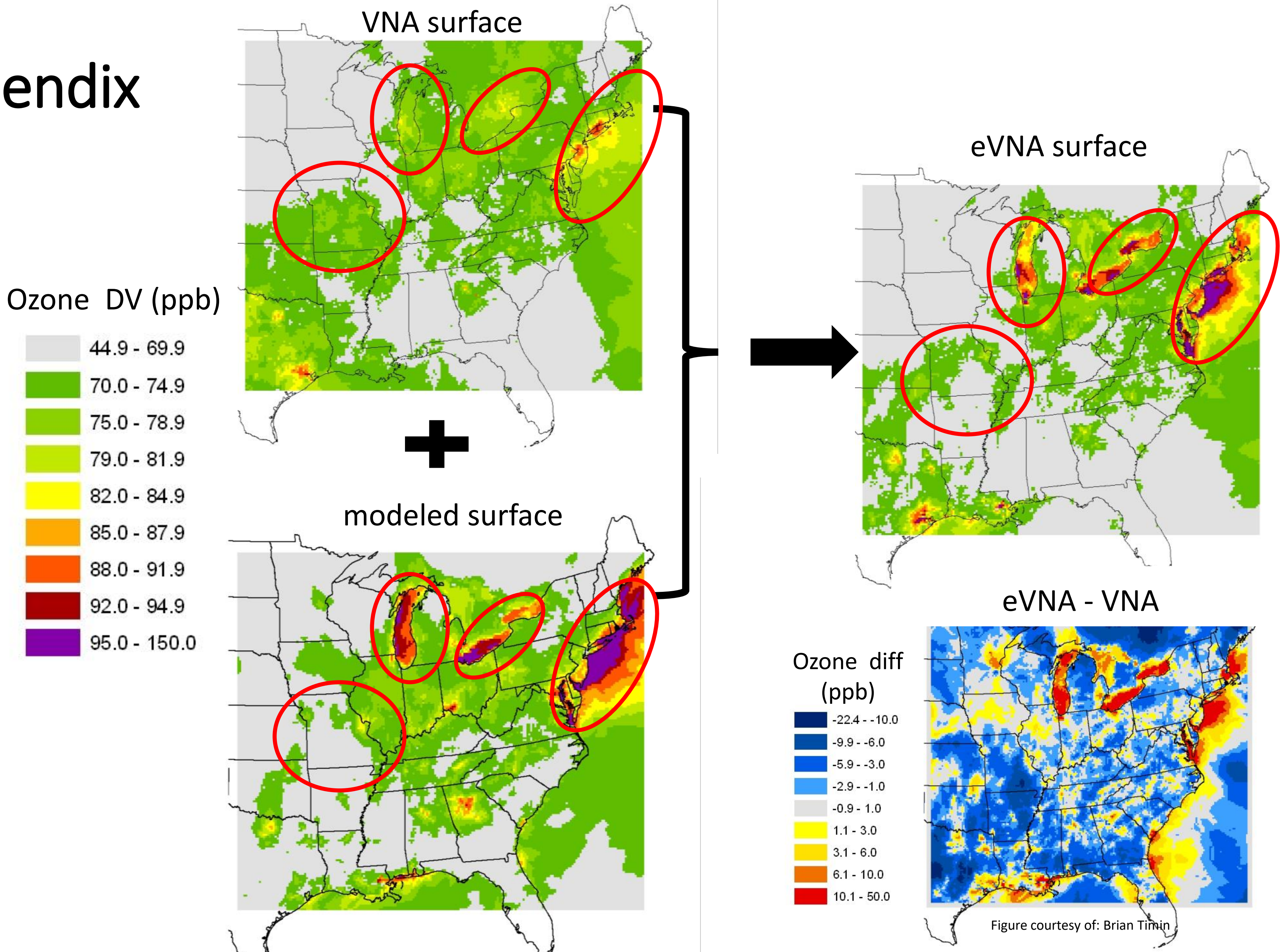
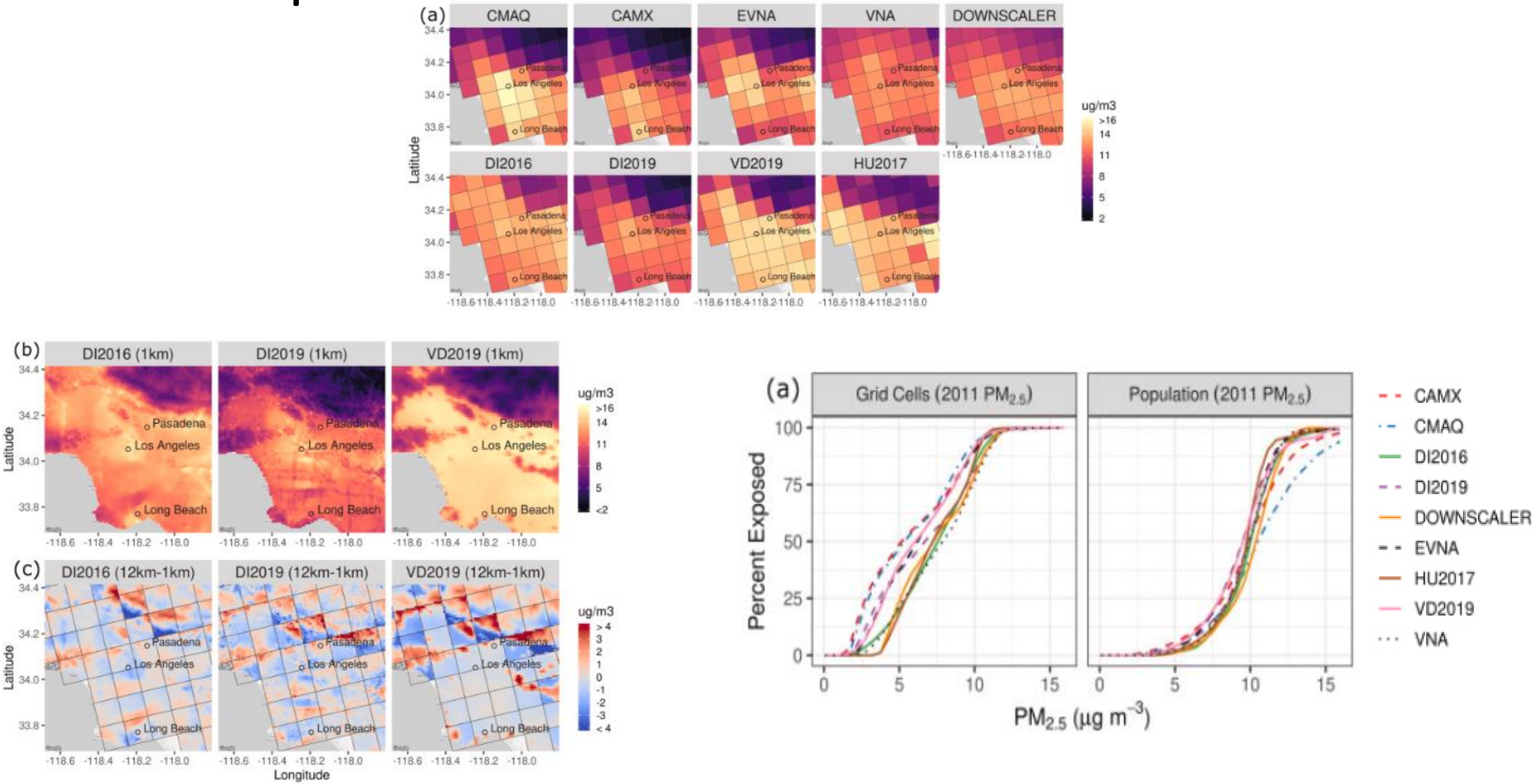
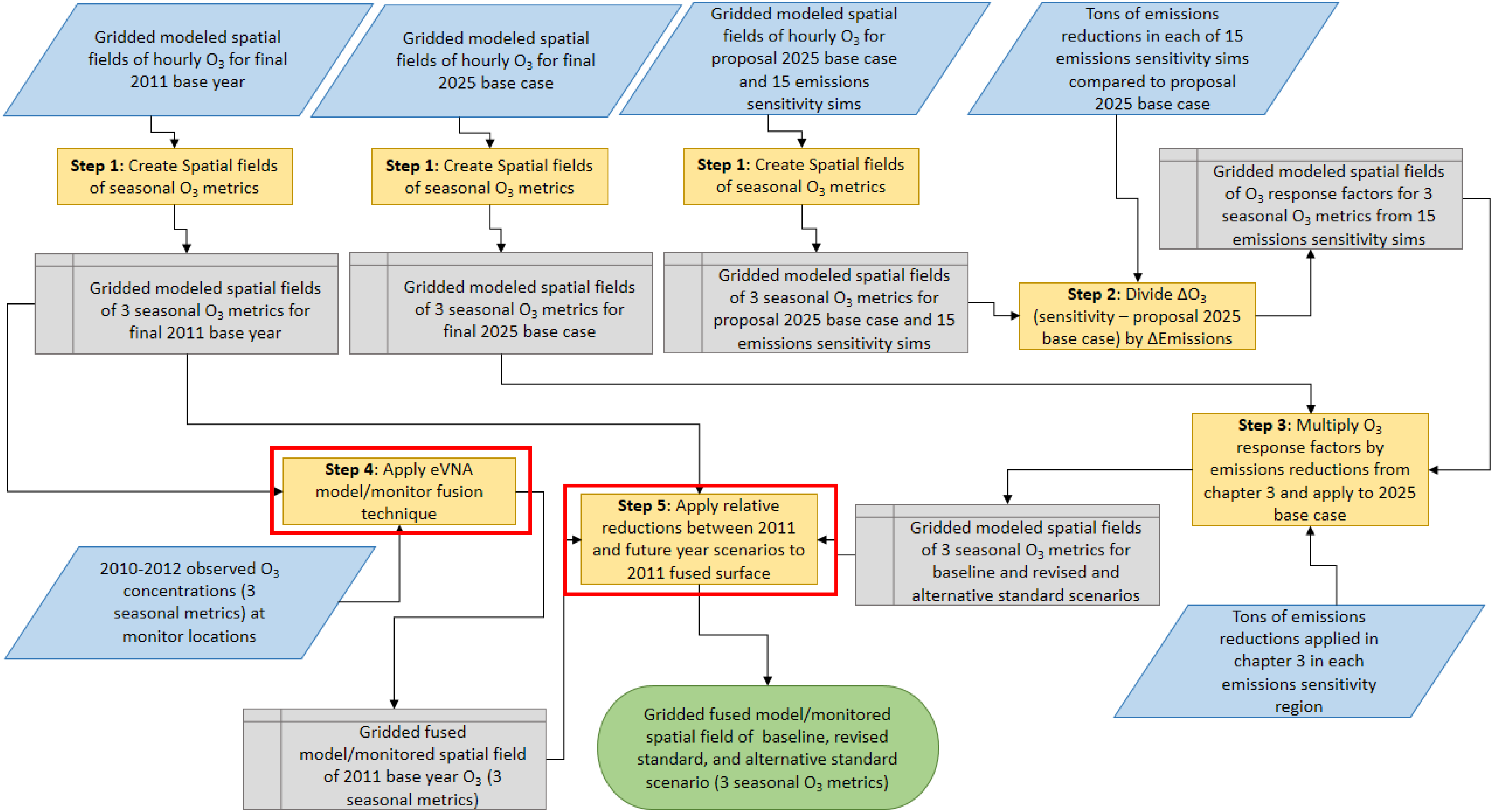


Figure courtesy of: Brian Timin

Appendix: Future Year Projections: Exposure Disparities for PM2.5 for 2011 and 2028



Appendix: Creating spatial surfaces in the RIA



Appendix: REA Analyses that Use CMAQ-HDDM Results

APEX Model

- Exposure Assessment and Clinical-based Risk Assessment
  - Ozone concentration inputs: 5 years of hourly spatial surfaces (census tract resolution) for 15 urban areas created by interpolating monitor values
  - Outputs:
    - **Exposure Assessment:** frequency of various populations experiencing exposures above benchmark levels of concern: 80, 70, 60 ppb
    - **Clinical-based Risk Assessment:** number of people who experience lung function decrements > 10%, 15%, 20%
  - Health outcomes most affected by exposure to "high" ozone concentrations

BenMAP

- Epidemiology-based Risk Assessment
  - Ozone concentration inputs:
    - **Urban area analysis:** daily time series of 8-hr max for area-wide average concentration ("composite monitor") in each city
    - **National analysis (current conditions only):** 3 national spatial surfaces of seasonal mean O_3
  - Outputs: ozone-related mortality, hospital admissions etc.
  - Uses linear, no-threshold C-R function, so all incremental changes in ozone impact estimates of total risk identically, regardless of the starting level of ozone
  - Health outcomes most affected by seasonal mean of area-wide average ozone