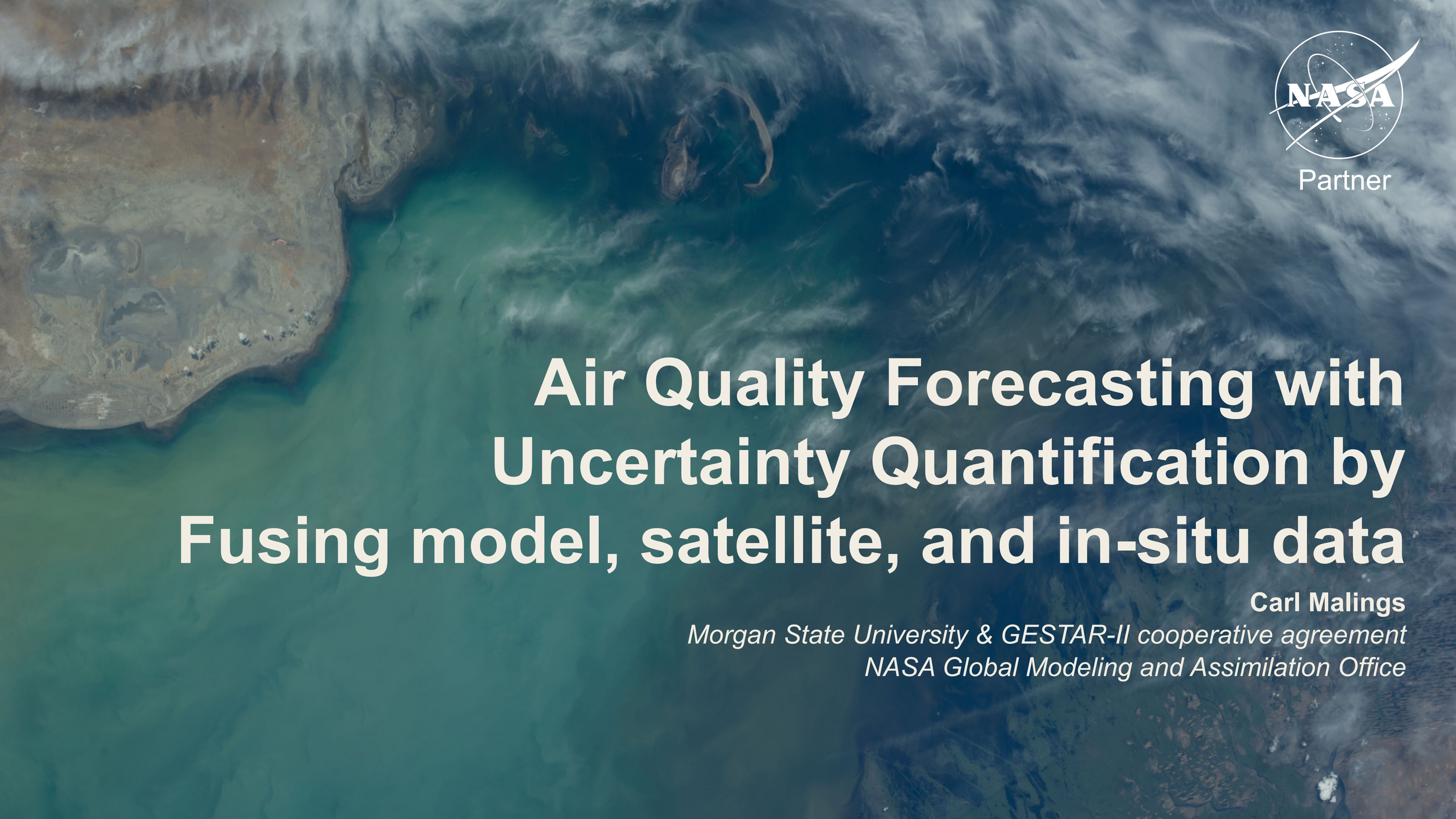


Partner

Air Quality Forecasting with Uncertainty Quantification by Fusing model, satellite, and in-situ data

Carl Malings

Morgan State University & GESTAR-II cooperative agreement

NASA Global Modeling and Assimilation Office

How can air quality managers use many data sources together?

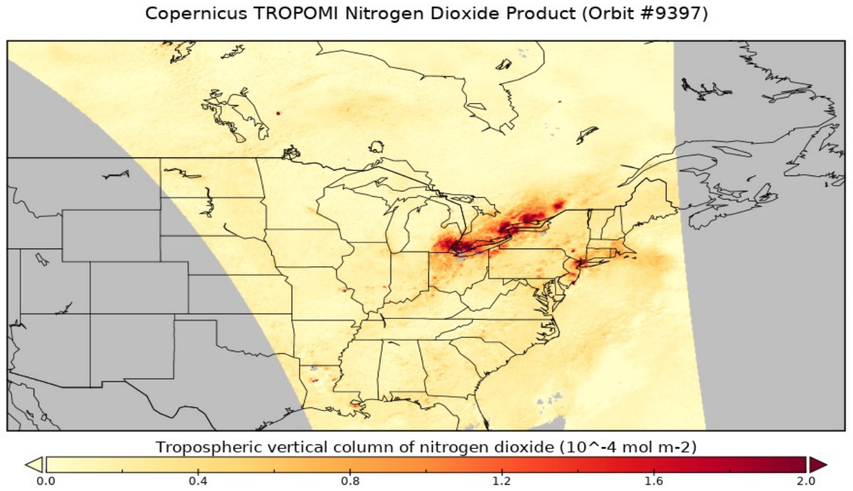
Partner



regulatory monitoring

+ accurate
- expensive
? representativity

form the "backbone" of the monitoring system, but insufficient alone



satellite retrievals

+ global coverage
- low time resolution
- column-integrated

good coverage and frequency, but need ground validation

low-cost monitoring

+ relatively inexpensive
+ dense/remote deployment
- greater noise and bias

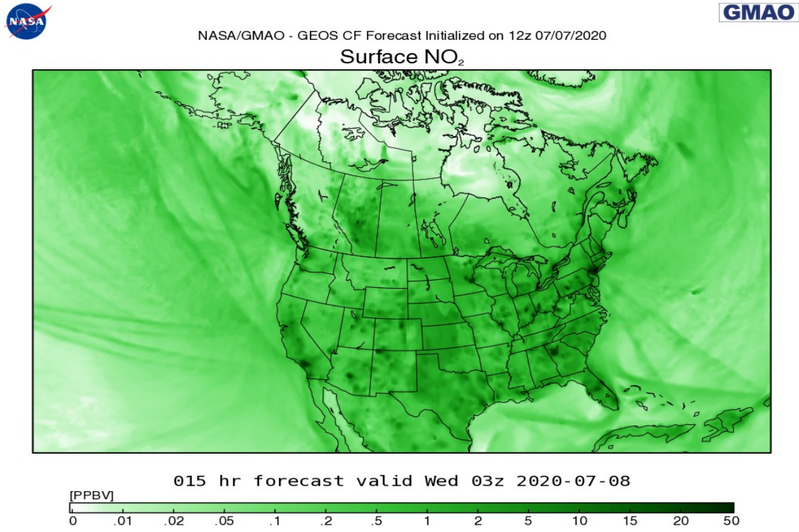
calibration is an open issue, but network density might offset these shortcomings



simulation models

+ global coverage
+ forecasting
- limited resolution

provide complete maps and forecasts, but need validation



The goal of our project is to…

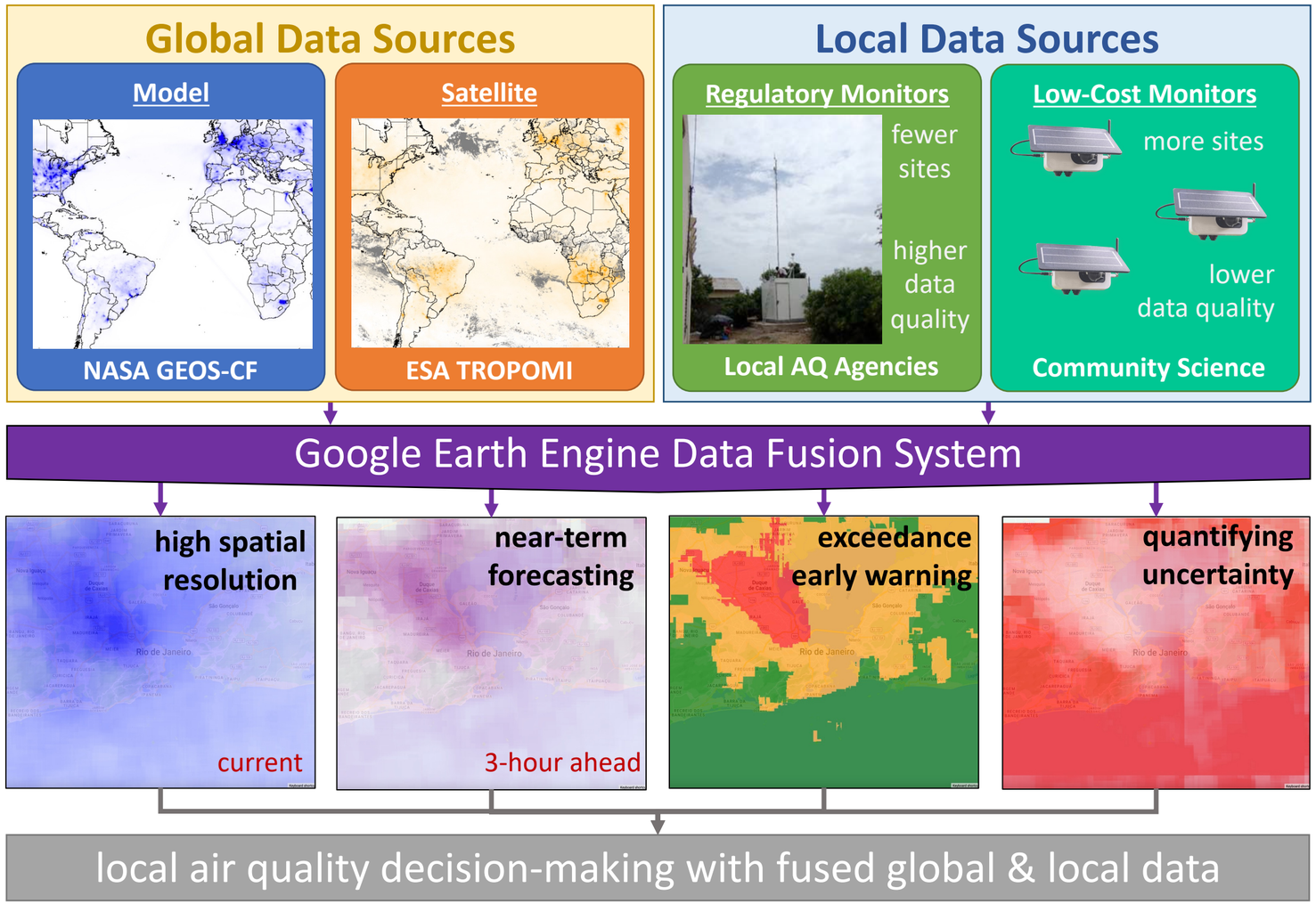
Partner

…integrate diverse **global** and **local** air quality data sources…

…using the cloud computing platform of **Google Earth Engine**…

…to provide synthesized **estimates** and **forecasts** of air quality at a **local scale** but with a **global scope**…

…freely accessible by air quality managers worldwide, facilitating their **decision-making** processes.

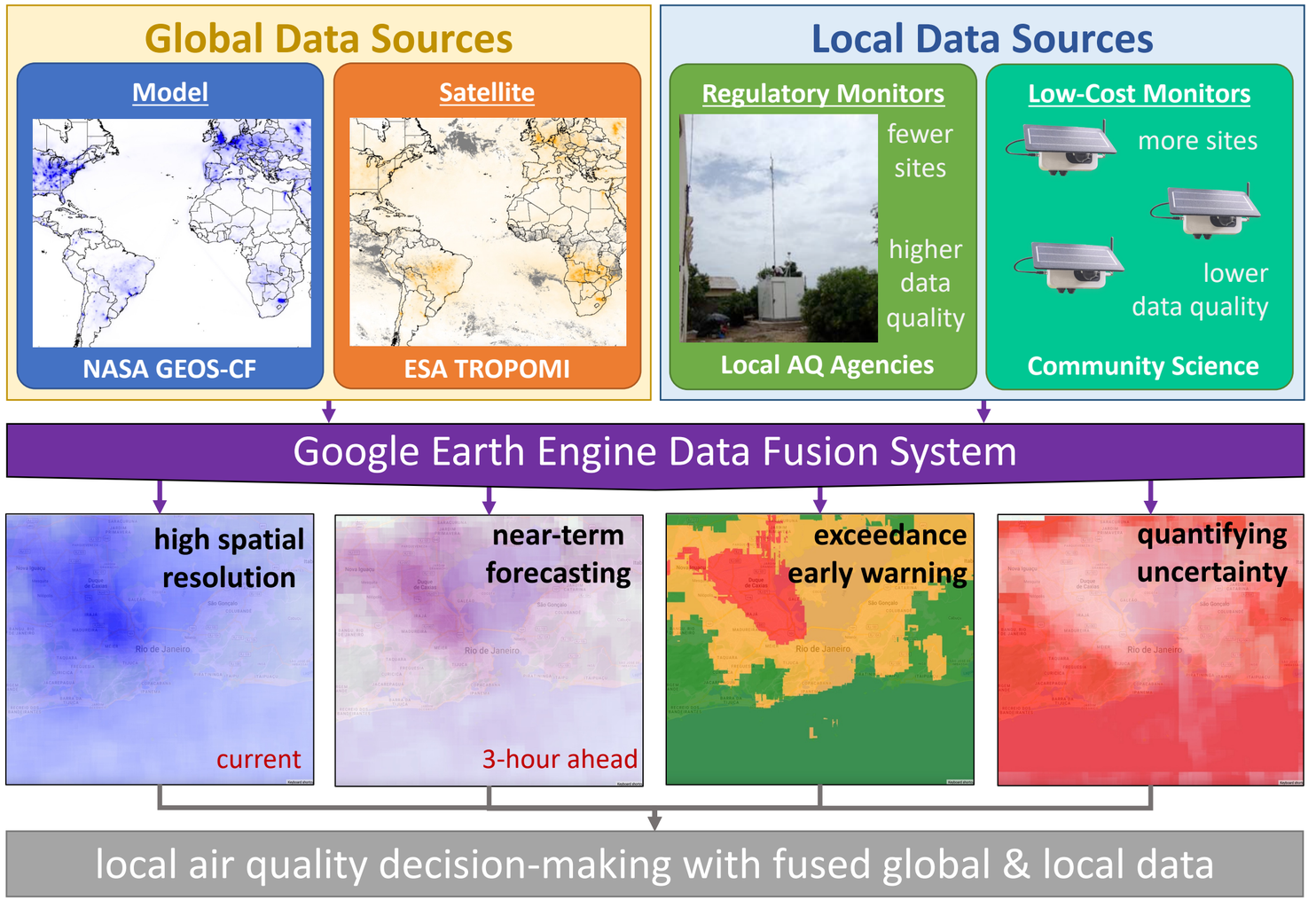


Source: NASA GMAO Science Snapshot "Google Earth Engine Data Fusion Tool to support Air Quality Managers"

Project Team & Partners

Partner

- **NASA GMAO:** basic algorithm development & refinement
- **Clarity:** low-cost sensor integration
- **Sonoma Technologies:** data fusion system implementation & user interface
- **WUSTL:** air quality data integration expertise (monthly/annual timescales)
- **Columbia LDEO:** experience training end-users in AQ data interpretation
- **UNEP:** integration with global users
  Dakar, Senegal
  Rio de Janeiro, Brazil
- **US EPA:** integration with US end-users
  Oregon, Colorado, Idaho, Louisiana



Source: NASA GMAO Science Snapshot "Google Earth Engine Data Fusion Tool to support Air Quality Managers"

Why quantify uncertainty in air quality forecasting?

Partner

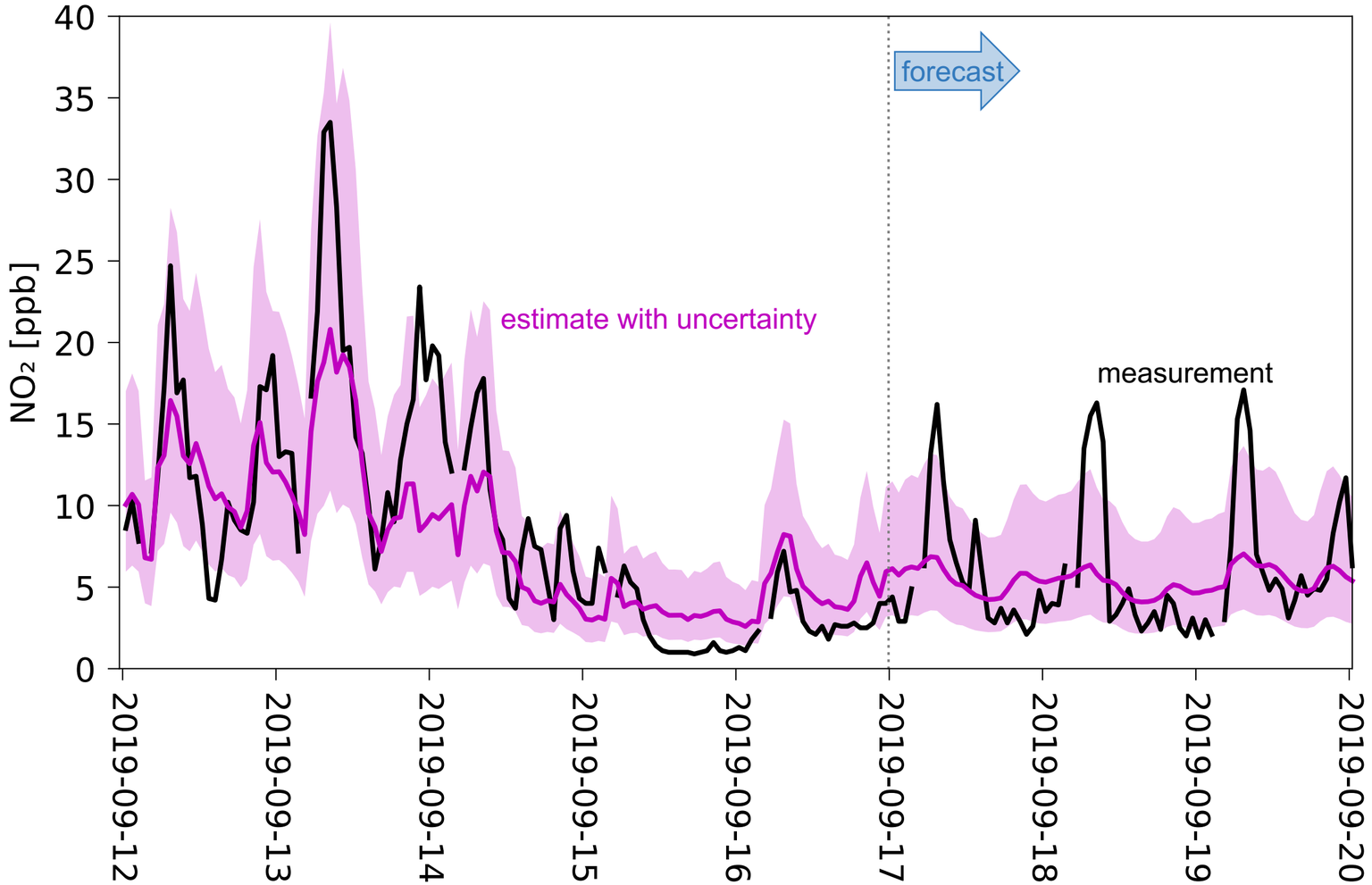
Provide a prior estimate of the relative confidence in a forecast

Convey probabilities of specific events, e.g., exceedance of standards

Identify a range of likely outcomes

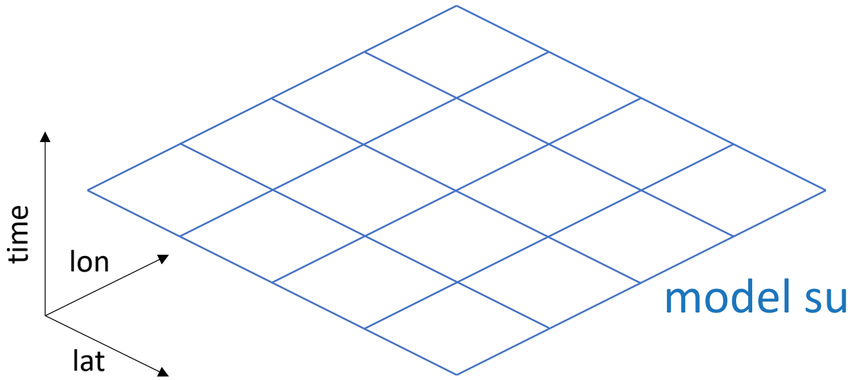
Quantify the impacts of different data sources in reducing uncertainties

Identify the potential to reduce uncertainties through additional data collection



Phase 1: Model Only

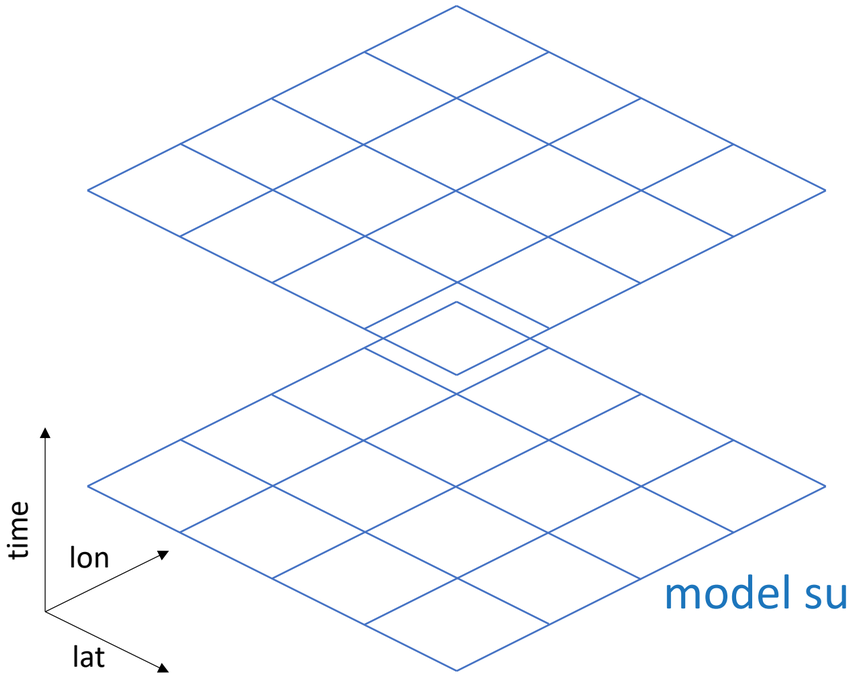
Phase	Estimate
1	forecast model (GEOS-CF)



model surface concentration $M(x,t)$

Phase 1: Model Only

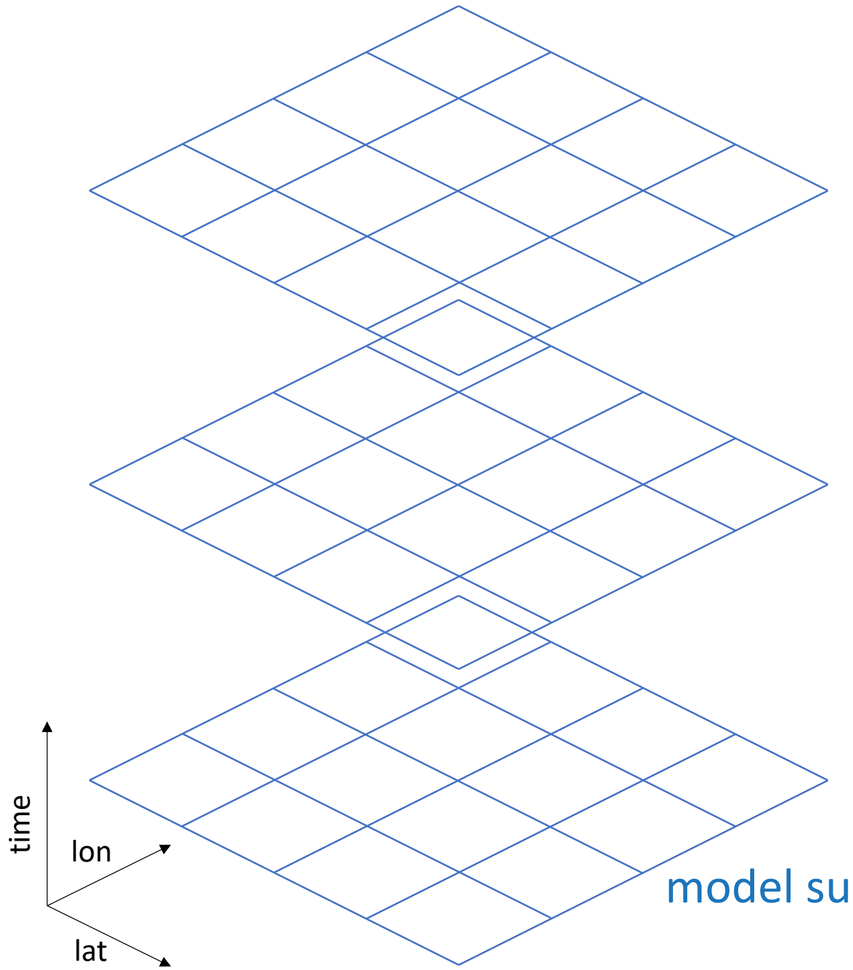
Phase	Estimate
1	forecast model (GEOS-CF)



model surface concentration $M(x,t)$

Phase 1: Model Only

Phase	Estimate
1	forecast model (GEOS-CF)

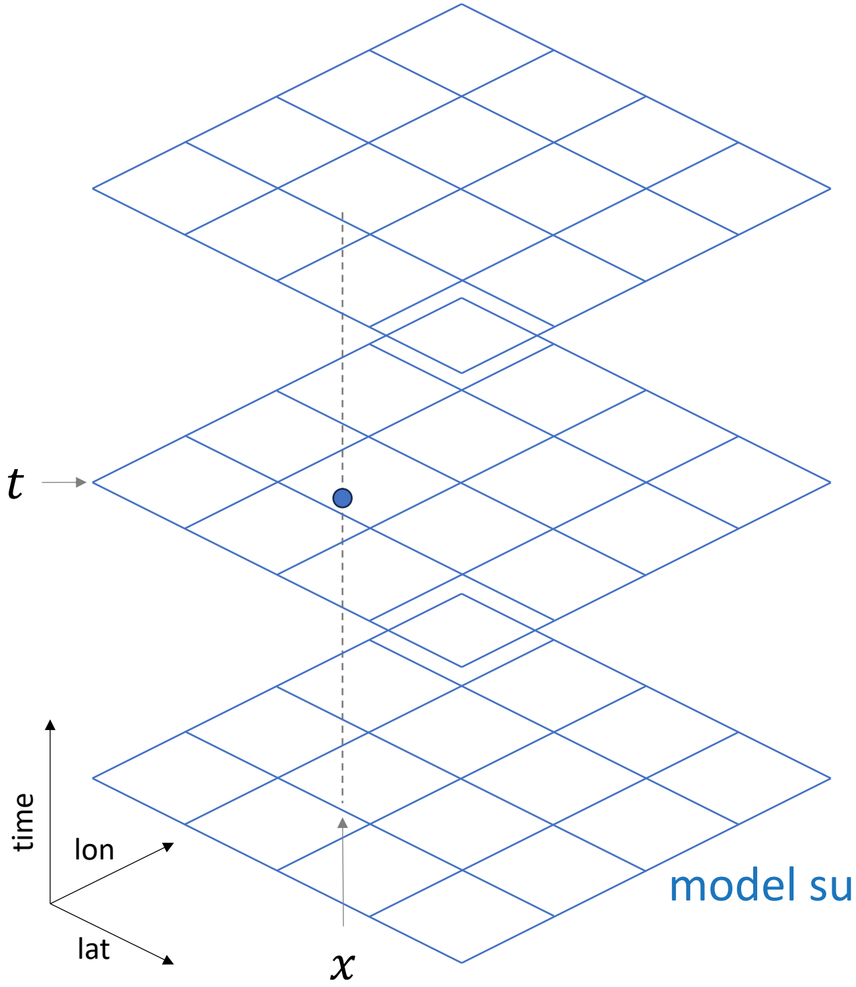


model surface concentration $M(x,t)$

Phase 1: Model Only

Phase	Estimate
1	forecast model (GEOS-CF)

$$F_1(x,t) = M(x,t)$$

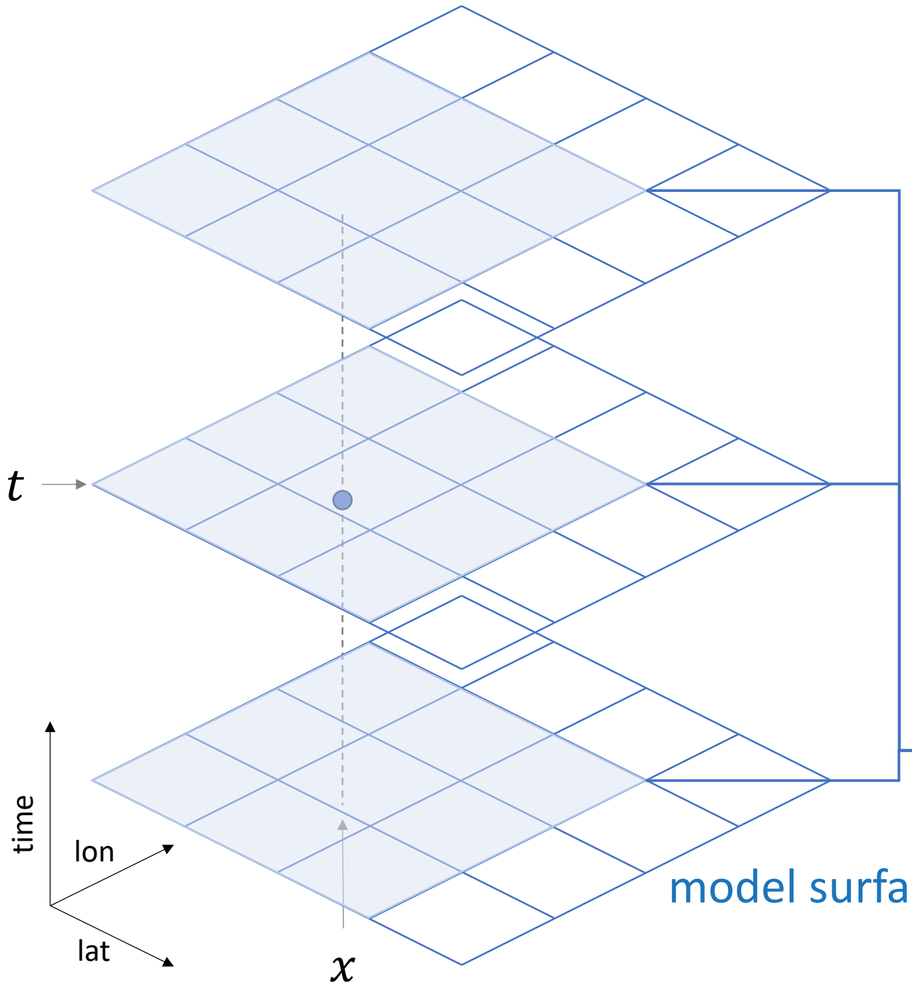


model surface concentration $M(x,t)$

Phase 1: Model Only

Partner

Phase	Estimate	Uncertainty
1	forecast model (GEOS-CF)	cell-to-cell variability of model



$$F_1(x,t) = M(x,t)$$

$$V_1(x,t) = V_{F1}(x,t,\tau)$$ ← uncertainty due to forecasting by τ ahead (ignore this for now…)

data fusion uncertainty (variance) at phase 1

$$+V_M(x,t)$$ ← uncertainty due to model internal variability

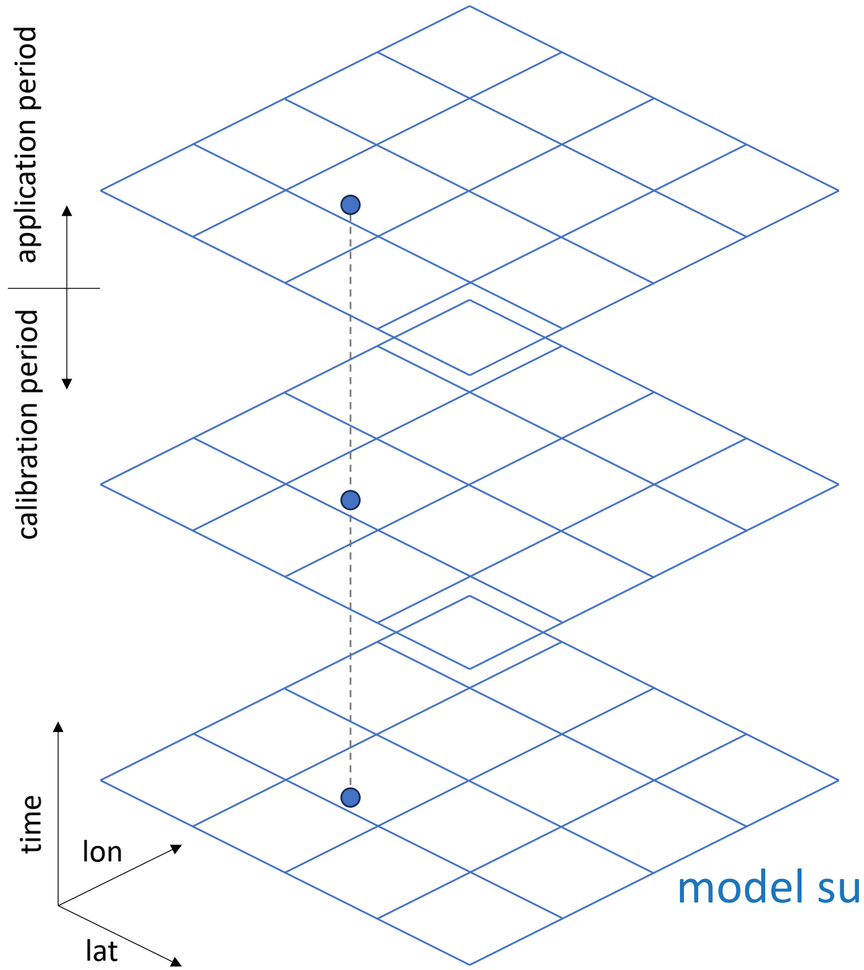
$$+V_{B1}(x,t)$$ ← uncertainty due to model bias

$$+V_{R1}(x,t)$$ ← uncertainty due to spatial representativity (model scale)

$$V_M(x,t) \approx \mathbb{E}_{x' \in X_n(x), t' \in T_n(t)} \left[\left(M(x',t') - M(x,t) \right)^2 \right]$$

model surface concentration $M(x,t)$

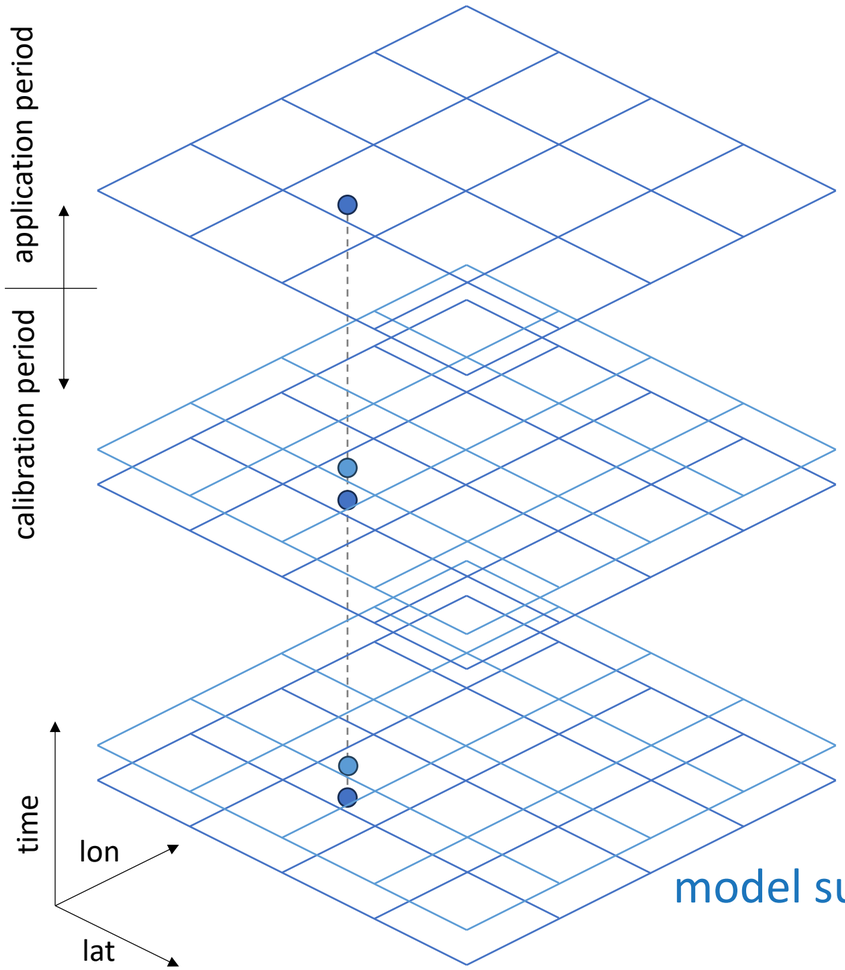
Phase 2: Model & Satellite



Phase	Estimate	Uncertainty
1	forecast model (GEOS-CF)	cell-to-cell variability of model
2		

model surface concentration $M(x,t)$

Phase 2: Model & Satellite



Phase	Estimate	Uncertainty
1	forecast model (GEOS-CF)	cell-to-cell variability of model
2		

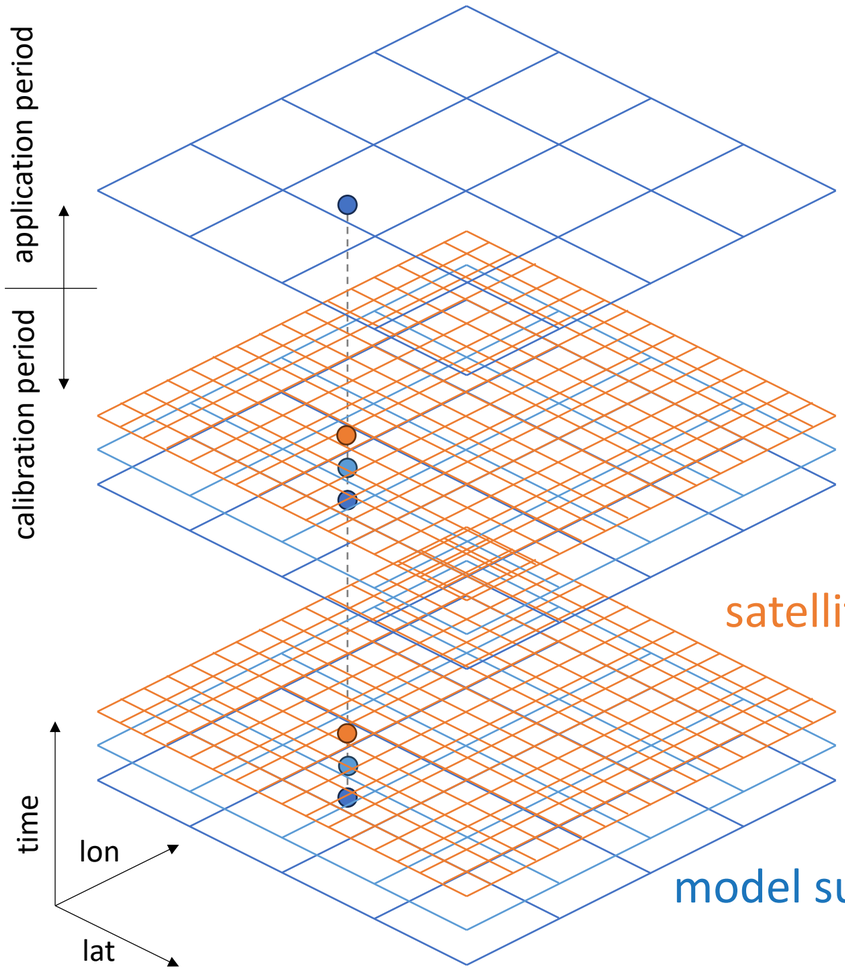
model column concentration $M_{col}(x,t)$

model surface concentration $M(x,t)$

Phase 2: Model & Satellite

Partner

Phase	Estimate	Uncertainty
1	forecast model (GEOS-CF)	cell-to-cell variability of model
2		



satellite column concentration $S_{col}(x,t)$

model column concentration $M_{col}(x,t)$

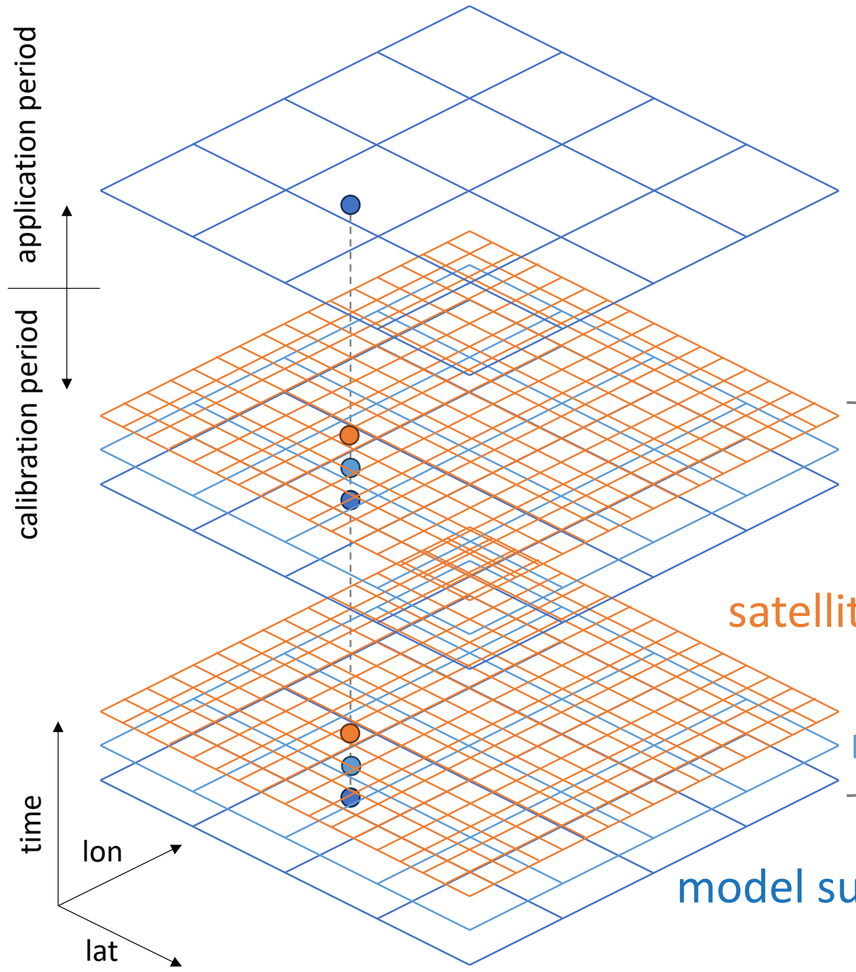
model surface concentration $M(x,t)$

Phase 2: Model & Satellite

Phase	Estimate	Uncertainty
1	forecast model (GEOS-CF)	cell-to-cell variability of model
2	satellite (TROPOMI) informs sub-model-grid variability	

$$F_2(x,t) = F_1(x,t) + D(x,t)$$

$$D(x,t) = \mathbb{E}_{t' \in T_{c,overpass}(t)}\left[\left(S_{col}(x,t') - M_{col}(x,t')\right)\phi(x,t')\,\psi(x,t,t')\right]$$



satellite column concentration $S_{col}(x,t)$

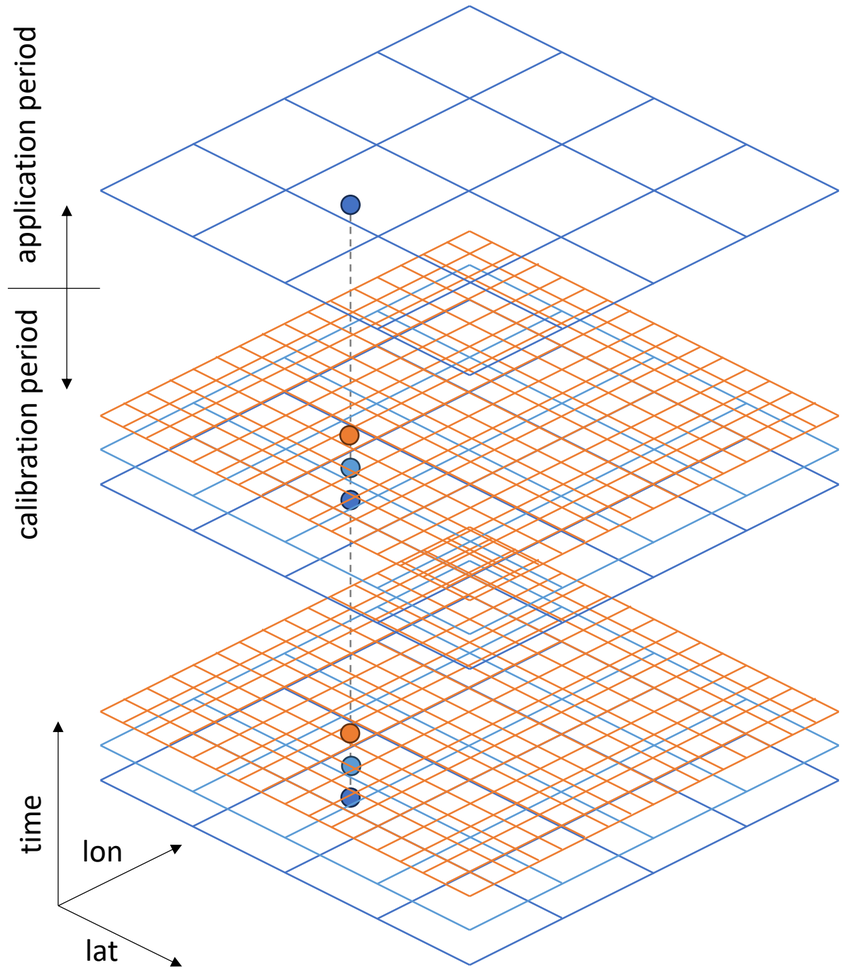
model column concentration $M_{col}(x,t)$

model surface concentration $M(x,t)$

$$\phi(x,t') \approx \frac{M(x,t')}{M_{col}(x,t')}$$ surface-to-column relationship

$$\psi(x,t,t') \approx \frac{M(x,t)}{M(x,t')}$$ target-time-to-overpass-time relationship

Phase 2: Model & Satellite



Phase	Estimate	Uncertainty
1	forecast model (GEOS-CF)	cell-to-cell variability of model
2	satellite (TROPOMI) informs sub-model-grid variability	satellite-to-model and surface-to-column ratios vary over time

$V_2(x,t) = V_{F2}(x,t,\tau)$ ← uncertainty due to forecasting by τ ahead

$+V_M(x,t)$ ← uncertainty due to model internal variability

$+V_D(x,t)$ ← uncertainty in satellite-to-model differences

$+2V_{MD}(x,t)$ ← co-variance of satellite-to-model differences with model outputs

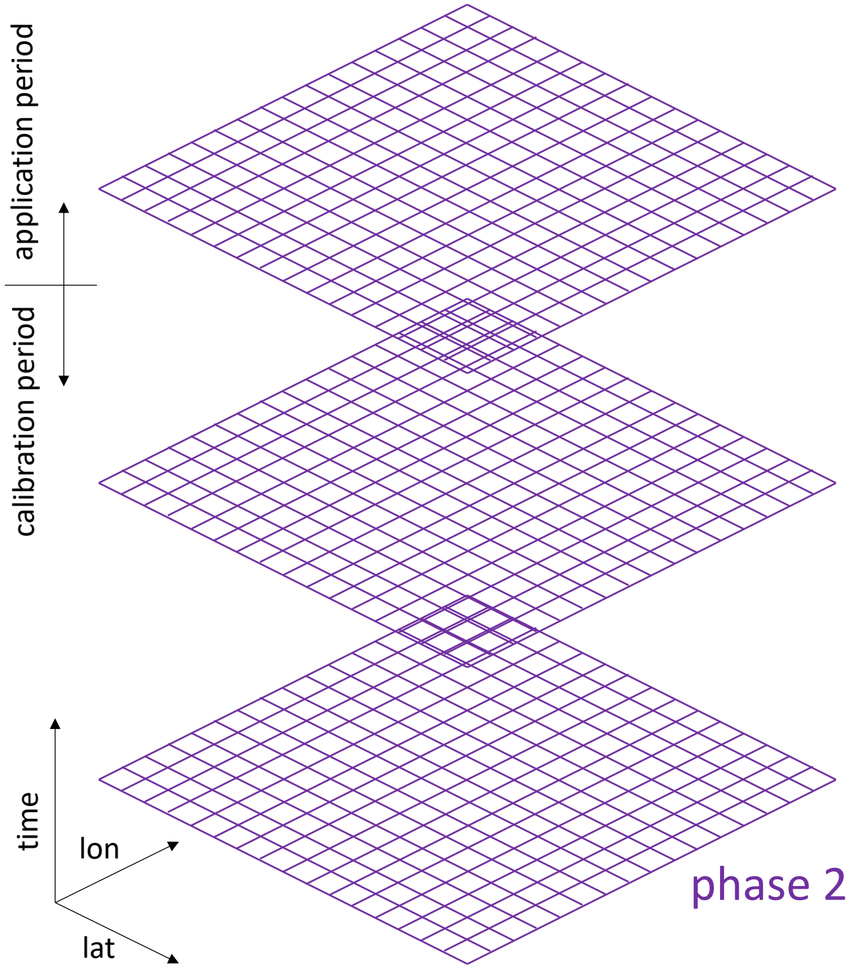
$+\mathrm{V}_{B2}(x,t)$ ← uncertainty due to model & satellite bias

$+\,V_{R2}(x,t)$ ← uncertainty due to spatial representativity (satellite scale)

$$V_D(x,t) \approx \mathbb{V}_{t' \in T_{c,overpass}(t)}\left[\left(S_{col}(x,t') - M_{col}(x,t')\right)\phi(x,t')\,\psi(x,t,t')\right]$$

$$V_{MD}(x,t) \approx \mathbb{E}_{x' \in X_n(x), t' \in T_n(t)}\left[\left(M(x',t') - M(x,t)\right)\left(D(x',t') - D(x,t)\right)\right]$$

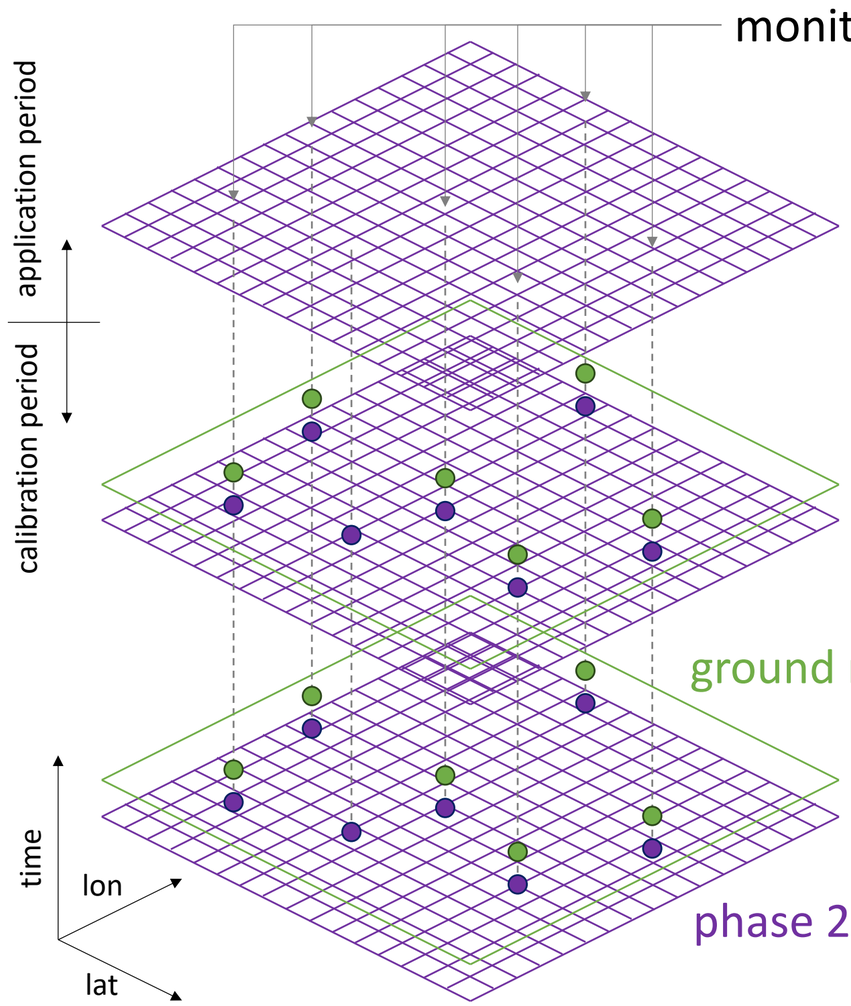
Phase 3: Model & Satellite & Ground



phase 2 estimates $F_2(x,t)$

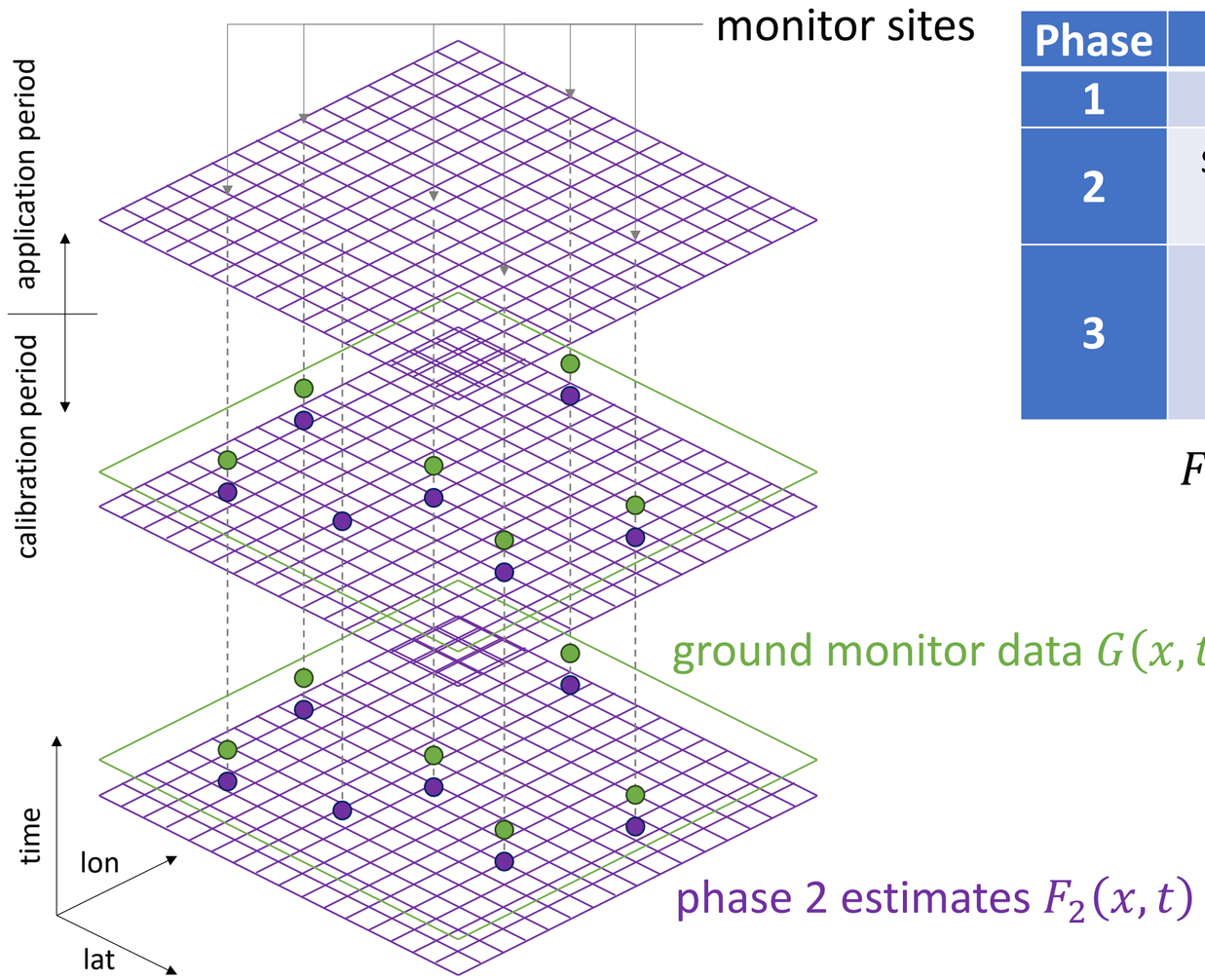
Phase	Estimate	Uncertainty
1	forecast model (GEOS-CF)	cell-to-cell variability of model
2	satellite (TROPOMI) informs sub-model-grid variability	satellite-to-model and surface-to-column ratios vary over time
3		

Phase 3: Model & Satellite & Ground



Phase	Estimate	Uncertainty
1	forecast model (GEOS-CF)	cell-to-cell variability of model
2	satellite (TROPOMI) informs sub-model-grid variability	satellite-to-model and surface-to-column ratios vary over time
3		

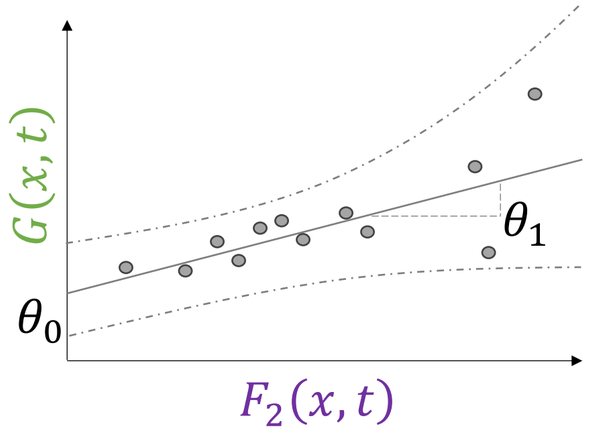
Phase 3: Model & Satellite & Ground



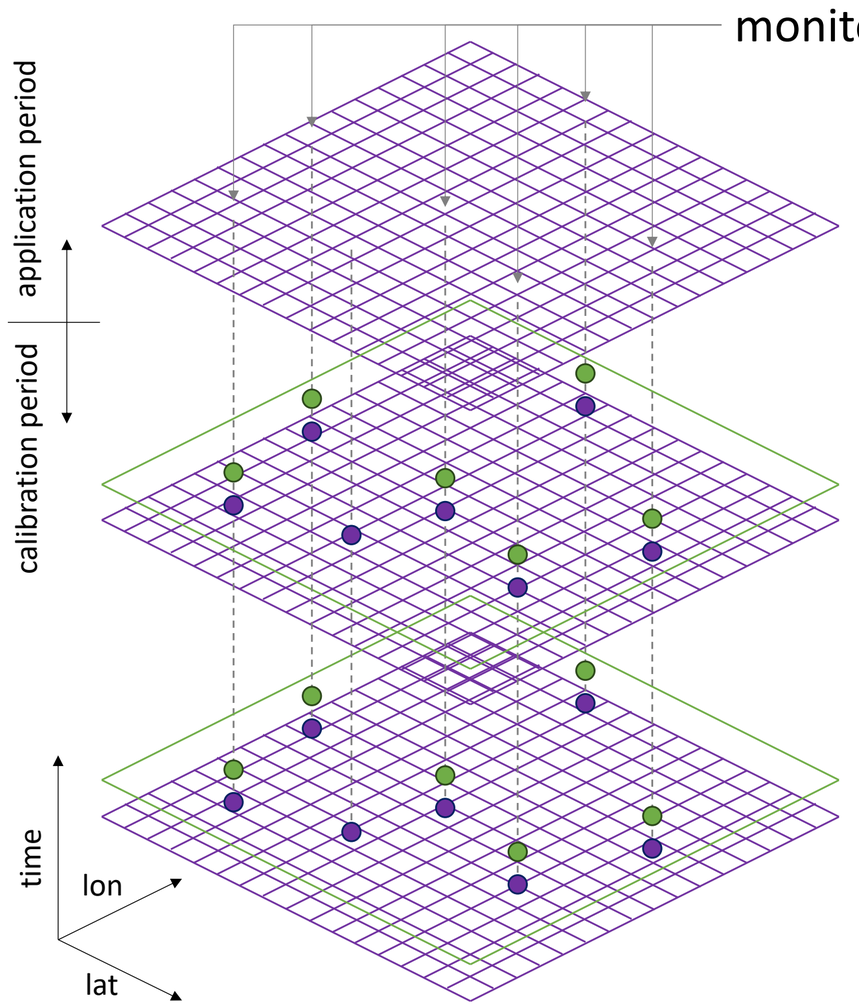
Phase	Estimate	Uncertainty
1	forecast model (GEOS-CF)	cell-to-cell variability of model
2	satellite (TROPOMI) informs sub-model-grid variability	satellite-to-model and surface-to-column ratios vary over time
3	phase 2 corrected to match surface monitor data	

$$F_3(x,t) = \theta_1 F_2(x,t) + \theta_0$$

$$\theta_0, \theta_1 = \mathbb{LR}_{t' \in T_c(t), x' \in X_c(x)}[G(x',t') \sim F_2(x',t')]$$



Phase 3: Model & Satellite & Ground



Phase	Estimate	Uncertainty
1	forecast model (GEOS-CF)	cell-to-cell variability of model
2	satellite (TROPOMI) informs sub-model-grid variability	satellite-to-model and surface-to-column ratios vary over time
3	phase 2 corrected to match surface monitor data	uncertain regression parameters between phase 2 output and surface monitor data

$$V_3(x,t) = V_{F3}(x,t,\tau)$$ ← uncertainty due to forecasting by τ ahead

$$+\theta_1^2[V_M(x,t) + V_D(x,t) + 2V_{MD}(x,t)]$$ ← rescaled from phase 2

$$+\mathrm{var}[\theta_1]F_2(x,t)^2$$

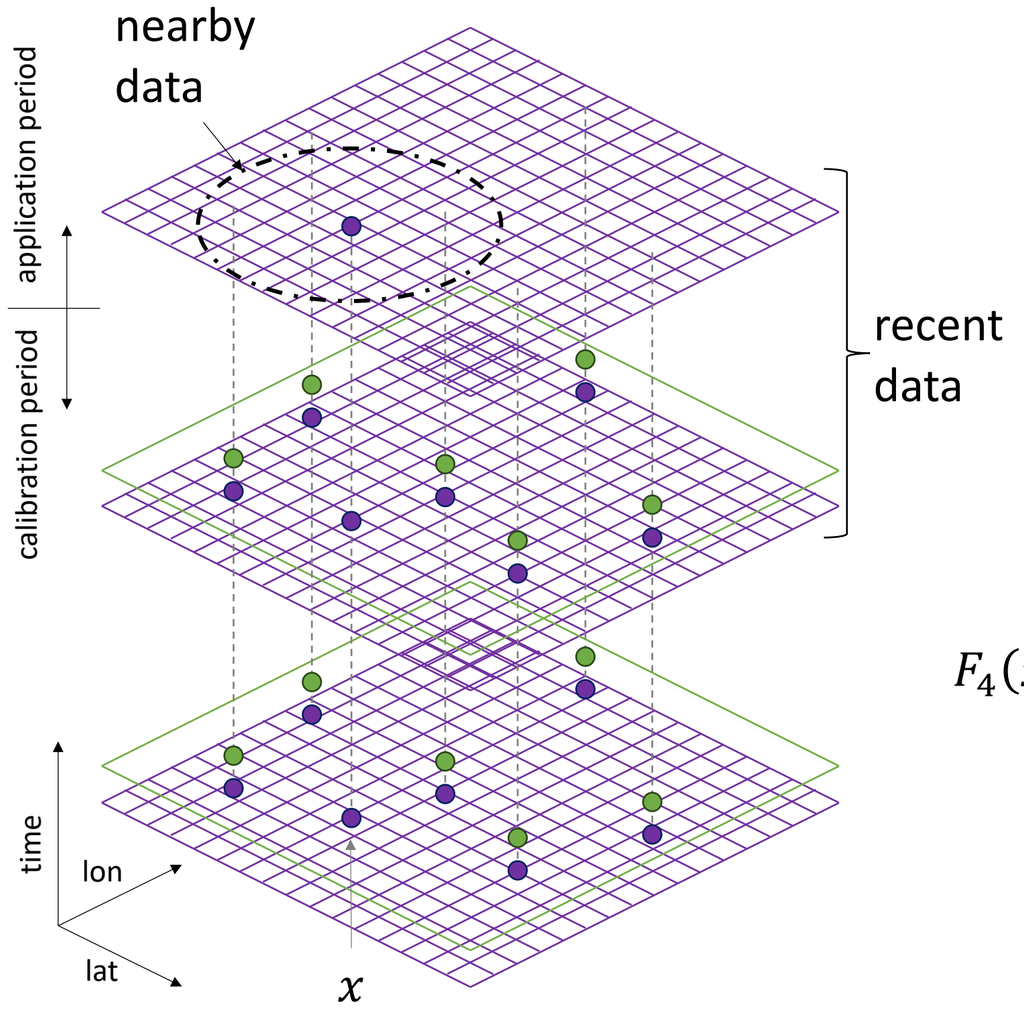
$$+2\mathrm{cov}[\theta_0,\theta_1]F_2(x,t)$$

$$+\mathrm{var}[\theta_0]$$

$$+\sigma^2_{residual}$$

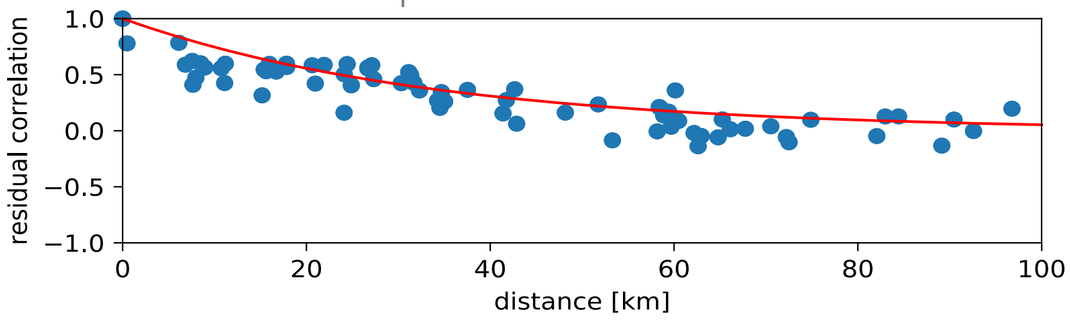
} variance and co-variance of regression parameters as well as regression residual are known

Phase 4: Residual Kriging

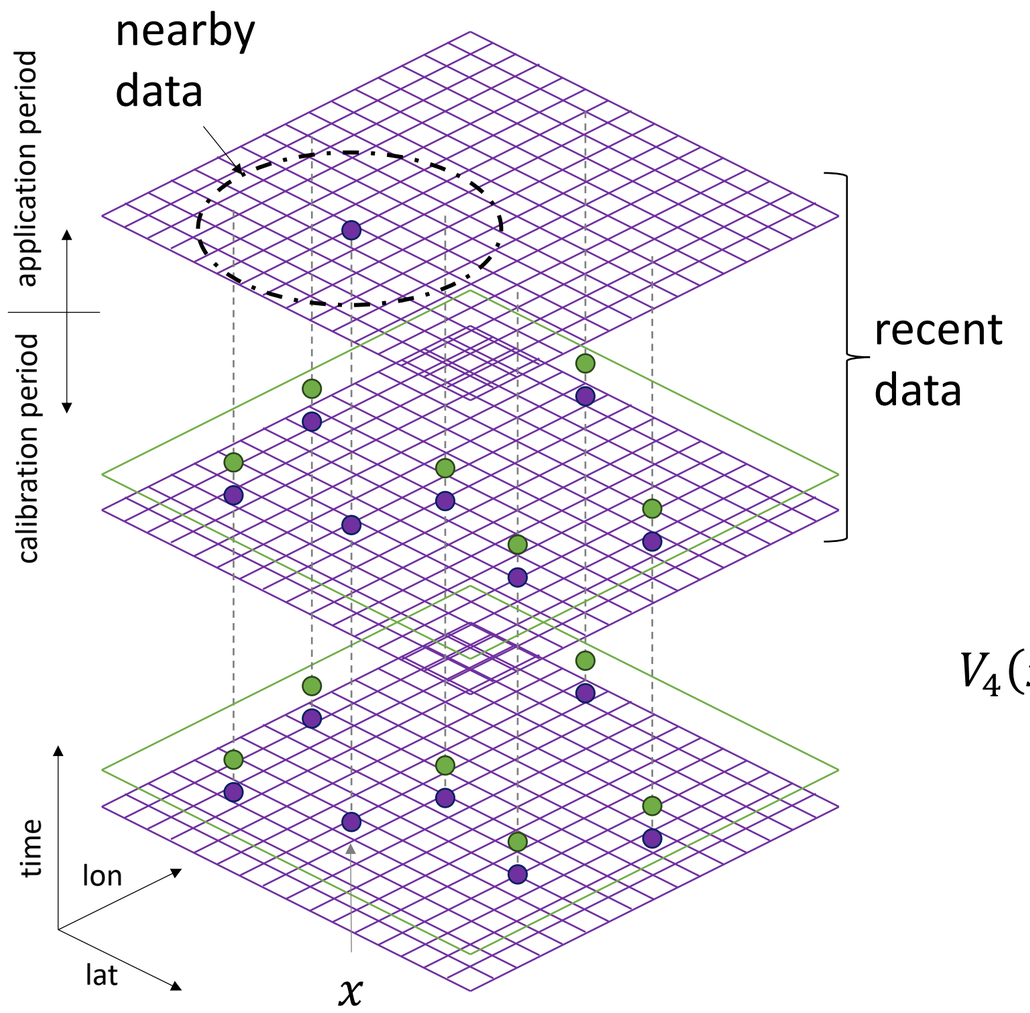


Phase	Estimate	Uncertainty
1	forecast model (GEOS-CF)	cell-to-cell variability of model
2	satellite (TROPOMI) informs sub-model-grid variability	satellite-to-model and surface-to-column ratios vary over time
3	phase 2 corrected to match surface monitor data	uncertain regression parameters between phase 2 output and surface monitor data
4	update phase 3 based on recent surface monitor data	

$$F_4(x,t) = F_3(x,t) + \sum_{x' \in X_n(x), t' \in T_n(t)} K(x,x',t,t') \left[G(x',t') - F_3(x',t') \right]$$



Phase 4: Residual Kriging



Phase	Estimate	Uncertainty
1	forecast model (GEOS-CF)	cell-to-cell variability of model
2	satellite (TROPOMI) informs sub-model-grid variability	satellite-to-model and surface-to-column ratios vary over time
3	phase 2 corrected to match surface monitor data	uncertain regression parameters between phase 2 output and surface monitor data
4	update phase 3 based on recent surface monitor data	uncertainty reduction via updating with nearby & recent data (kriging)

$$V_4(x,t) = V_3(x,t) - \sum_{x' \in X_n(x), t' \in T_n(t)} K(x,x',t,t') \operatorname{cov}[G(x',t'), F_3(x',t')]$$

Summary of Data Fusion Estimates & Uncertainties

Partner

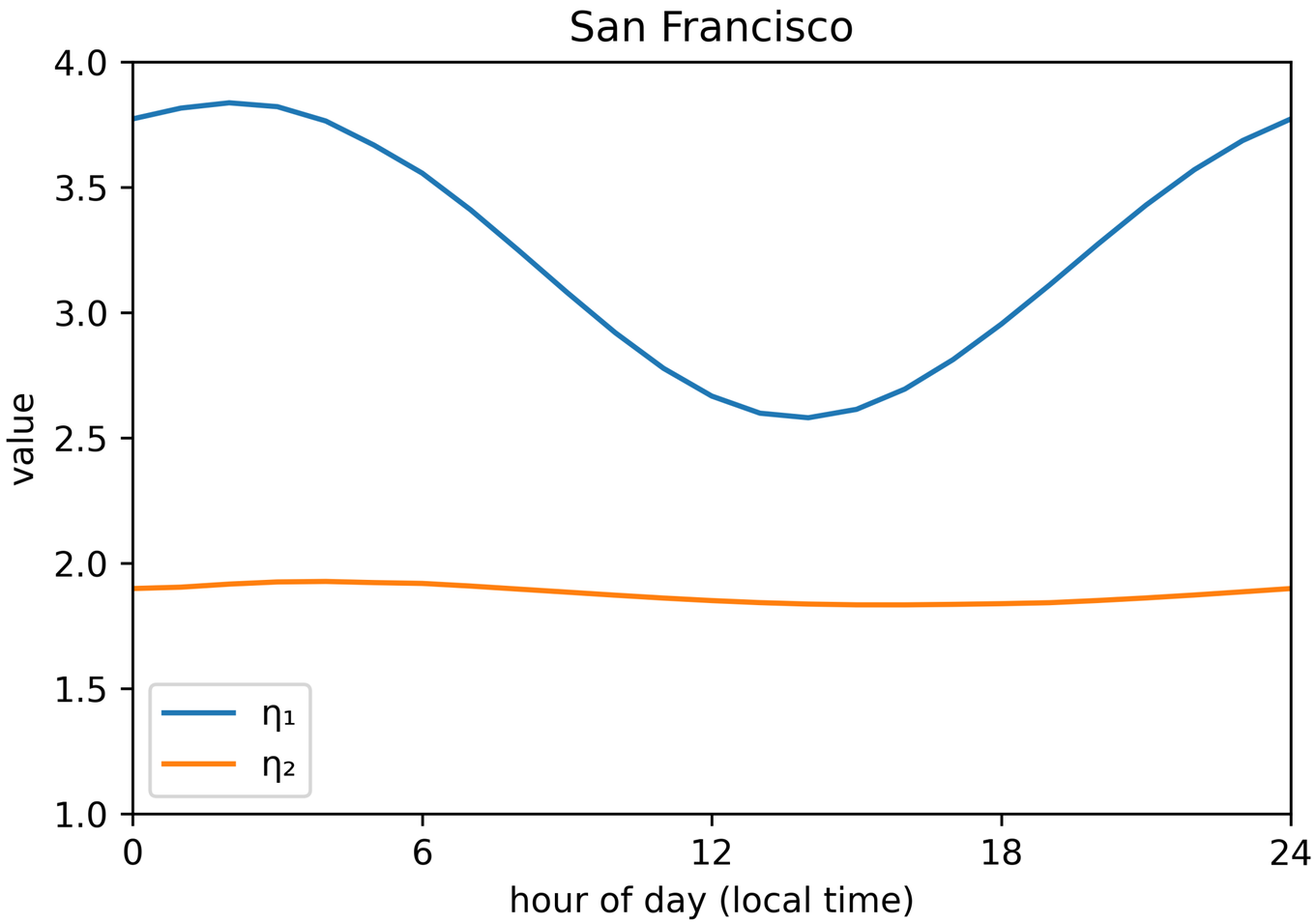
Phase		Estimate	Uncertainty			
			Bias	Model Variability	Model Scale Spatial Representativity	Satellite Scale Spatial Representativity
1	Model	$F_1(x,t) = M(x,t)$	$V_{B1}(x,t)$	$V_M(x,t)$	$V_{R1}(x,t)$	
2	Model & Satellite	$F_2(x,t) = \text{avg}_{t' \in T_c(t)}\left[\left(S_{col}(x,t') - M_{col}(x,t')\right)\phi(x,t')\,\psi(x,t,t')\right] + F_1(x,t) = D(x,t) + F_1(x,t)$	$V_{B2}(x,t)$	$V_M(x,t)$	$V_D(x,t) + 2V_{MD}(x,t)$	$V_{R2}(x,t)$
3	Model & Satellite & Ground	$F_3(x,t) = \theta_1 F_2(x,t) + \theta_0$ $with \quad \theta_0, \theta_1 = \mathbb{LR}_{t' \in T_c(t), x' \in X_c(x)}[G(x',t') \sim F_2(x',t')]$	0*	$\theta_1^2 V_M(x,t)$	$\theta_1^2[V_D(x,t) + 2V_{MD}(x,t)]$	$\text{var}[\theta_1]F_2(x,t)^2 + 2\text{cov}[\theta_0,\theta_1]F_2(x,t) + \text{var}[\theta_0] + \sigma^2_{residual}$
4	Model & Satellite & Ground & Kriging	$F_4(x,t) = \text{F}_3(x,t) + \sum_{x' \in X_n(x), t' \in T_n(t)} K(x,x',t,t')\,[G(x',t') - F_3(x',t')]$	0*	$\theta_1^2 V_M(x,t)$	$\theta_1^2[V_D(x,t) + 2V_{MD}(x,t)]$	$\text{var}[\theta_1]F_2(x,t)^2 + 2\text{cov}[\theta_0,\theta_1]F_2(x,t) + \text{var}[\theta_0] + \sigma^2_{residual}$
			$-\sum_{x' \in X_n(x), t' \in T_n(t)} K(x,x',t,t')\,\text{cov}[G(x',t'), F_3(x,t)]$			

Summary of Data Fusion Estimates & Uncertainties

Partner

Phase		Estimate	Uncertainty			
			Bias	Model Variability	Model Scale Spatial Representativity	Satellite Scale Spatial Representativity
1	Model	$F_1(x,t) = M(x,t)$	$V_{B1}(x,t)$	$V_M(x,t)$	$V_{R1}(x,t)$	
2	Model & Satellite	$F_2(x,t)$ $= \text{avg}_{t' \in T_c(t)}\left[\left(S_{col}(x,t') - M_{col}(x,t')\right)\phi(x,t')\,\psi(x,t,t')\right]$ $+ F_1(x,t) = D(x,t) + F_1(x,t)$	$V_{B2}(x,t)$	$V_M(x,t)$	$V_D(x,t)$ $+ 2V_{MD}(x,t)$	$V_{R2}($
3	Model & Satellite & Ground	$F_3(x,t) = \theta_1 F_2(x,t) + \theta_0$ $with \quad \theta_0, \theta_1 = \mathbb{LR}_{t' \in T_c(t), x' \in X_c(x)}[G(x',t') \sim F_2(x',t')]$	0*	$\theta_1^2 V_M(x,t)$	$\theta_1^2[V_D(x,t)$ $+ 2V_{MD}(x,t)]$	$\text{var}[\theta_1]F_2(x,t)^2$ $+ 2\text{cov}[\theta_0,\theta_1]F_2($ $+ \text{var}[\theta_0$ $+ \sigma^2_{residual}$
4	Model & Satellite & Ground & Kriging	$F_4(x,t)$ $= \text{F}_3(x,t)$ $+ \sum_{x' \in X_n(x), t' \in T_n(t)} K(x,x',t,t')\,[G(x',t') - F_3(x',t')]$	0*	$\theta_1^2 V_M(x,t)$	$\theta_1^2[V_D(x,t)$ $+ 2V_{MD}(x,t)]$	$\text{var}[\theta_1]F_2(x,t)^2$ $+ 2\text{cov}[\theta_0,\theta_1]F_2(x,t)$ $+ \text{var}[\theta_0]$ $+ \sigma^2_{residual}$
			$-\sum_{x' \in X_n(x), t' \in T_n(t)} K(x,x',t,t')\,\text{cov}[G(x',t'), F_3(x,t)]$			

“Backfilling” Uncertainty Estimates

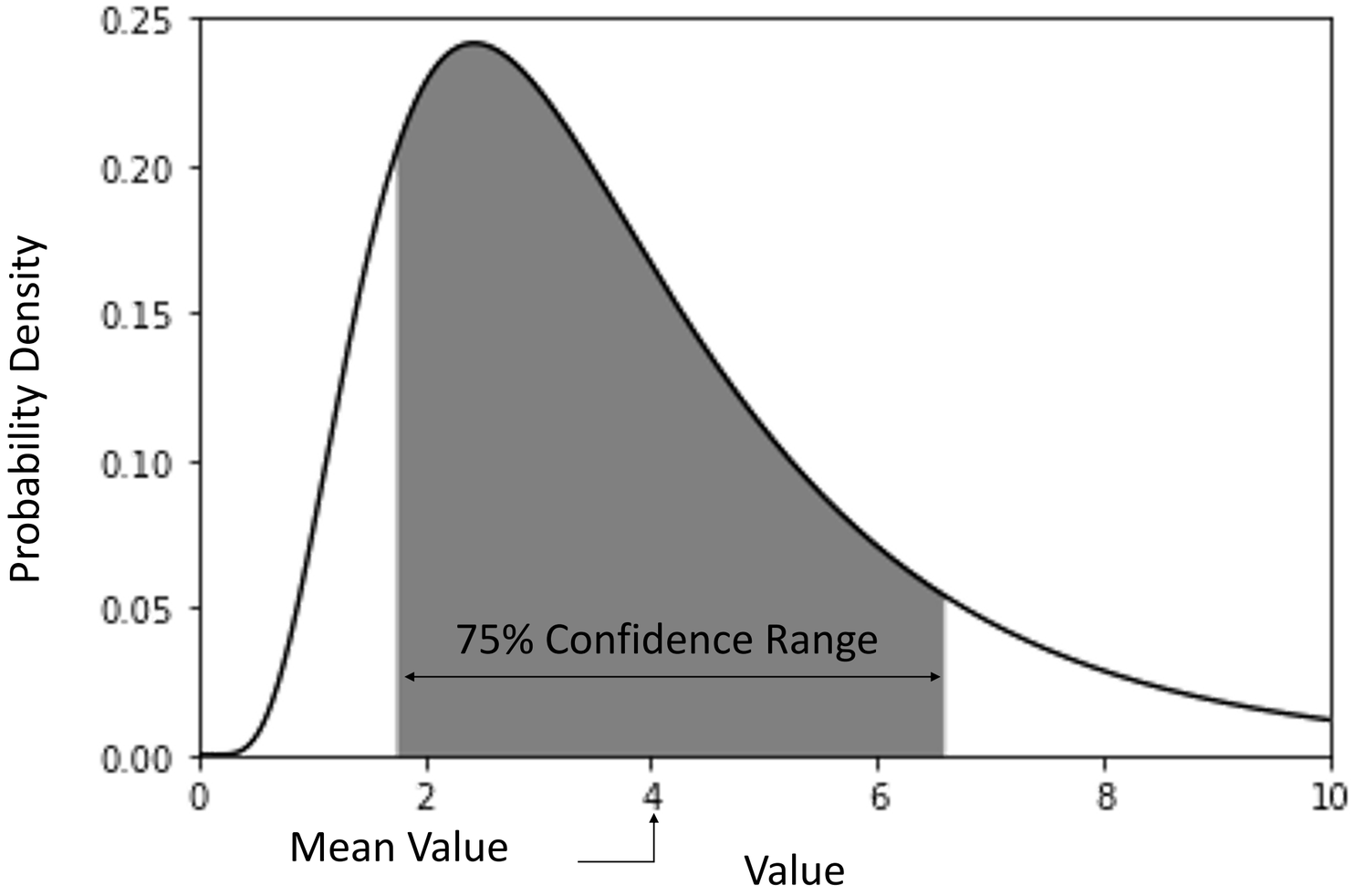


Assume an empirical relationship between bias and representation errors and the quantifiable component of the uncertainty in phases 1 and 2.

$$V_{B1}(x,t) + V_{R1}(x,t) \approx \eta_1^2 \, V_M(x,t)$$

$$V_{B2}(x,t) + V_{R2}(x,t) \approx \eta_2^2 \left(V_M(x,t) + V_D(x,t) + 2V_{MD}(x,t)\right)$$

Defining Confidence Intervals



Assuming a distribution for the values being estimated (a lognormal distribution is assumed in this case), confidence intervals can be estimated.

$$\mu(x,t) = \log\left[\frac{F(x,t)}{\sqrt{1+\frac{V(x,t)}{F(x,t)^2}}}\right]$$

$$\sigma(x,t) = \sqrt{\log\left[1+\frac{V(x,t)}{F(x,t)^2}\right]}$$

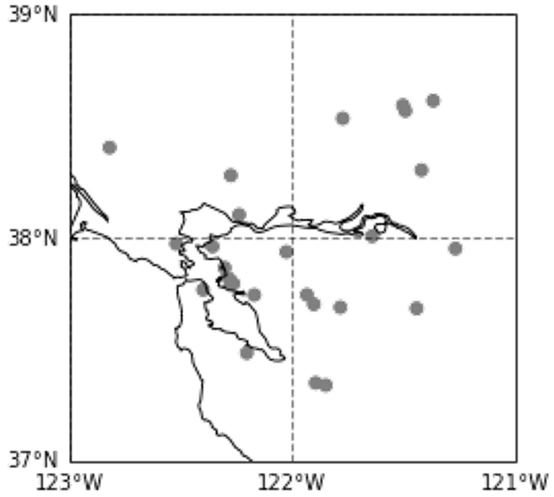
$$f(x,t) \sim LN\big(\mu(x,t), \sigma(x,t)\big)$$

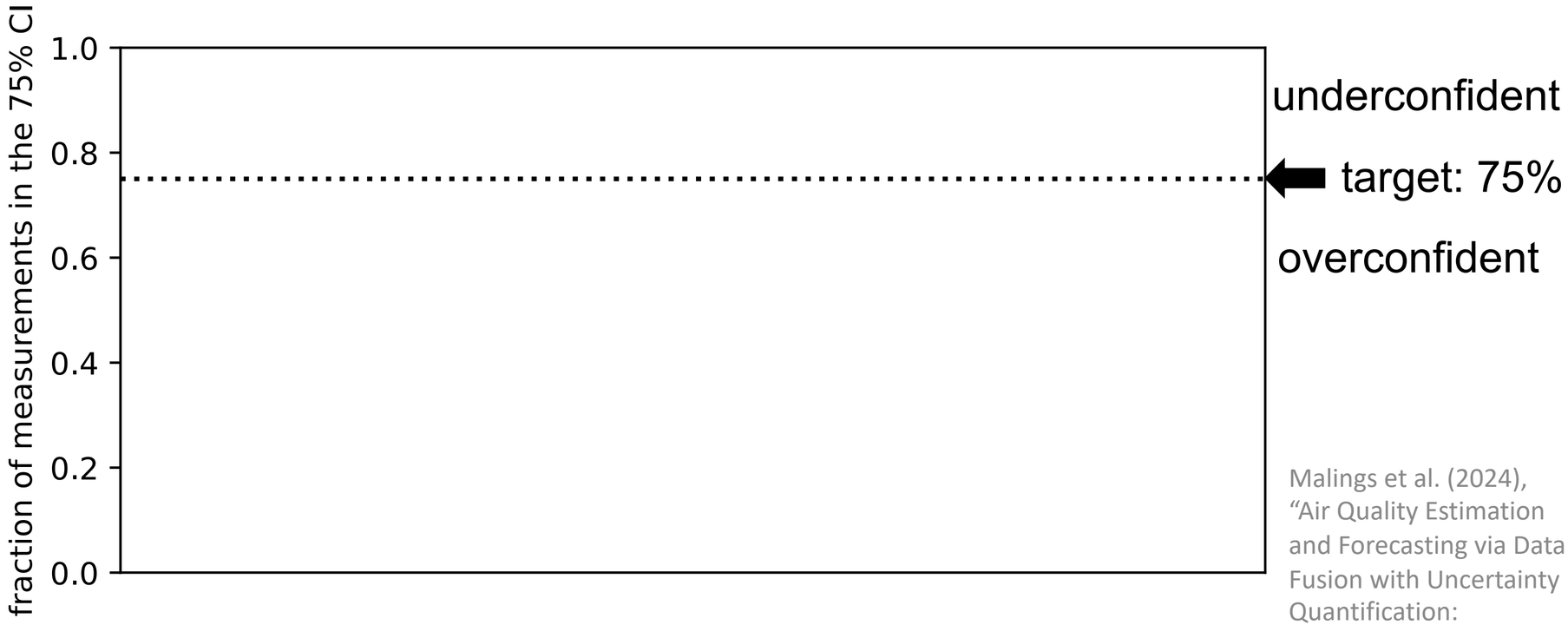
Case Study Results

Case Study Details

San Francisco
September 2019
Surface NO_2
Lognormal distribution
Cross-validation test
25 ground monitors

Ground Sites





Malings et al. (2024), "Air Quality Estimation and Forecasting via Data Fusion with Uncertainty Quantification: Theoretical Framework and Preliminary Results" *JGR: Machine Learning & Computation (under review).* PREPRINT DOI: 10.22541/essoar.171052565.52477494/v1

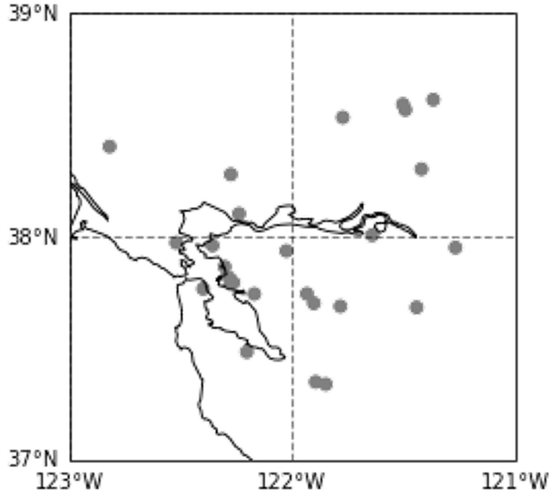
Case Study Results

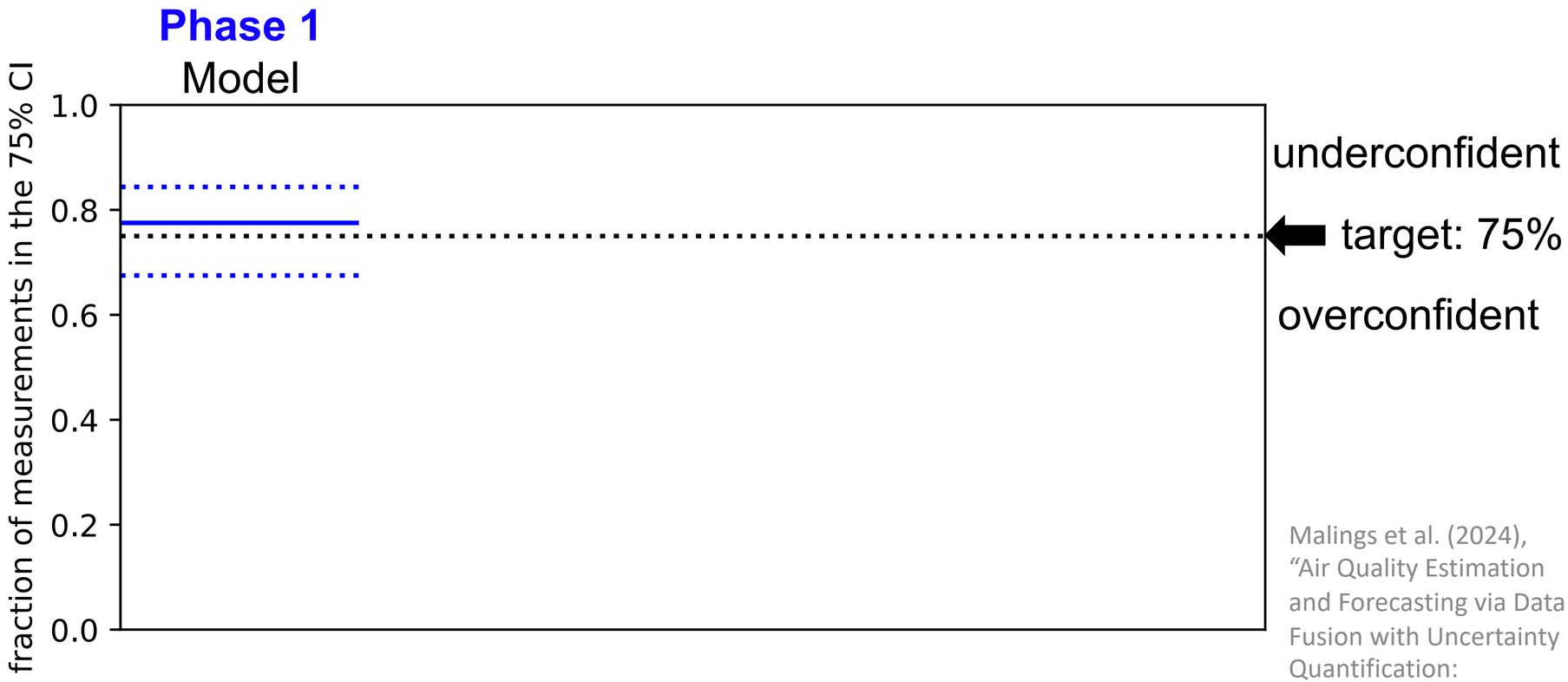
Partner

Case Study Details

San Francisco
September 2019
Surface NO_2
Lognormal distribution
Cross-validation test
25 ground monitors

Ground Sites





Malings et al. (2024), "Air Quality Estimation and Forecasting via Data Fusion with Uncertainty Quantification: Theoretical Framework and Preliminary Results" *JGR: Machine Learning & Computation (under review).* PREPRINT DOI: 10.22541/essoar.171052565.52477494/v1

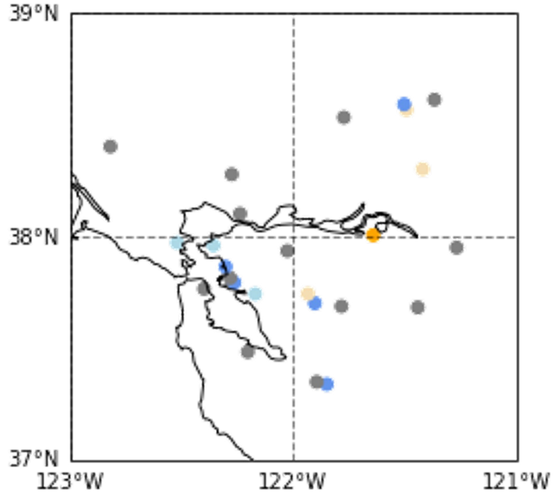
Case Study Results

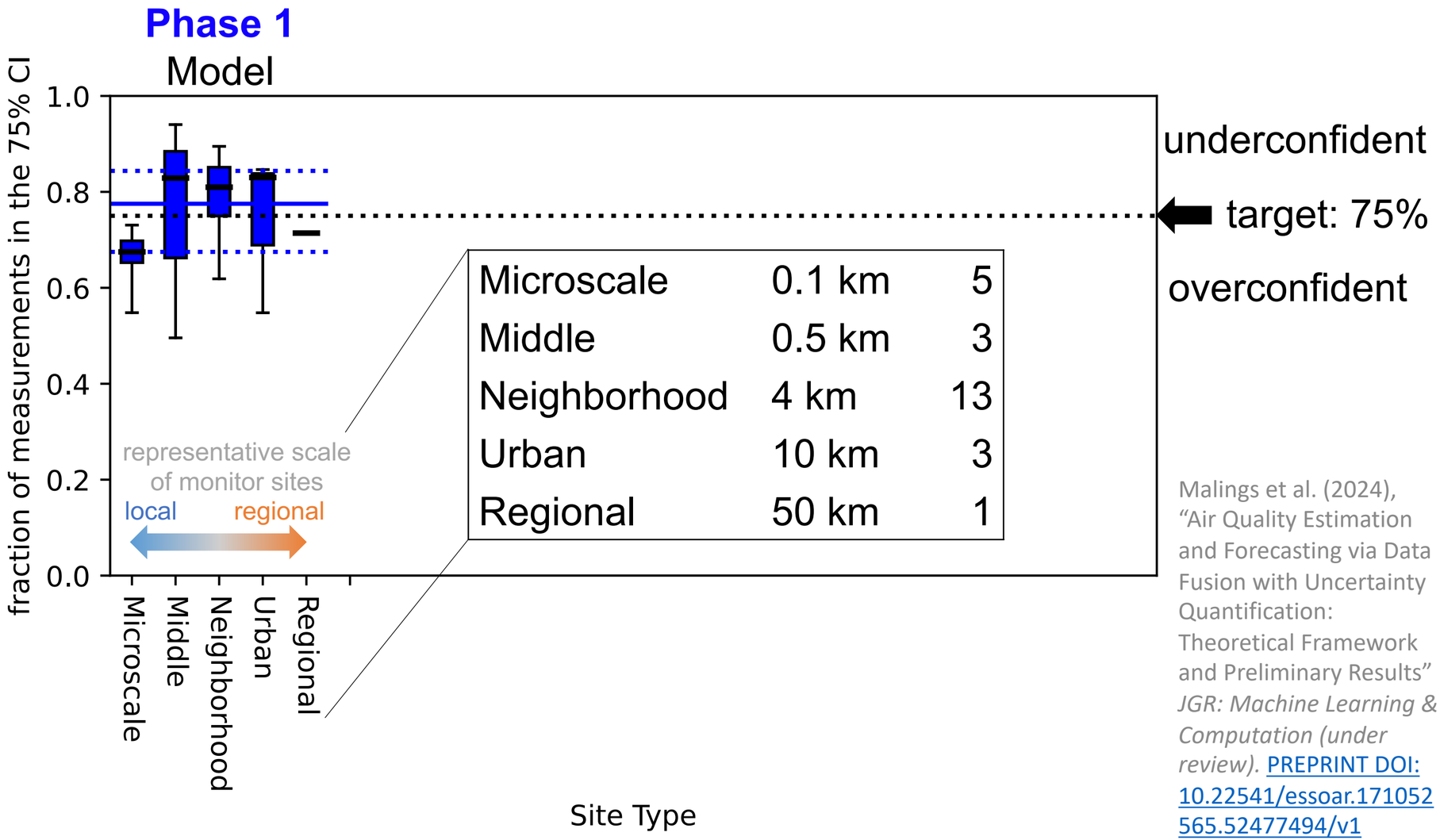
Partner

Case Study Details

San Francisco
September 2019
Surface NO_2
Lognormal distribution
Cross-validation test
25 ground monitors

Ground Sites





Malings et al. (2024), "Air Quality Estimation and Forecasting via Data Fusion with Uncertainty Quantification: Theoretical Framework and Preliminary Results" *JGR: Machine Learning & Computation (under review).* PREPRINT DOI: 10.22541/essoar.171052565.52477494/v1

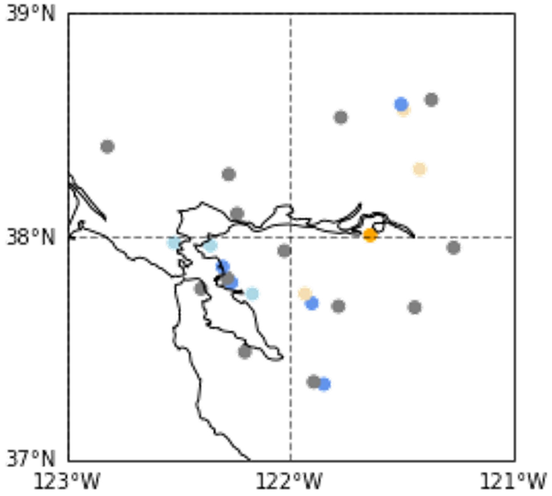
Case Study Results

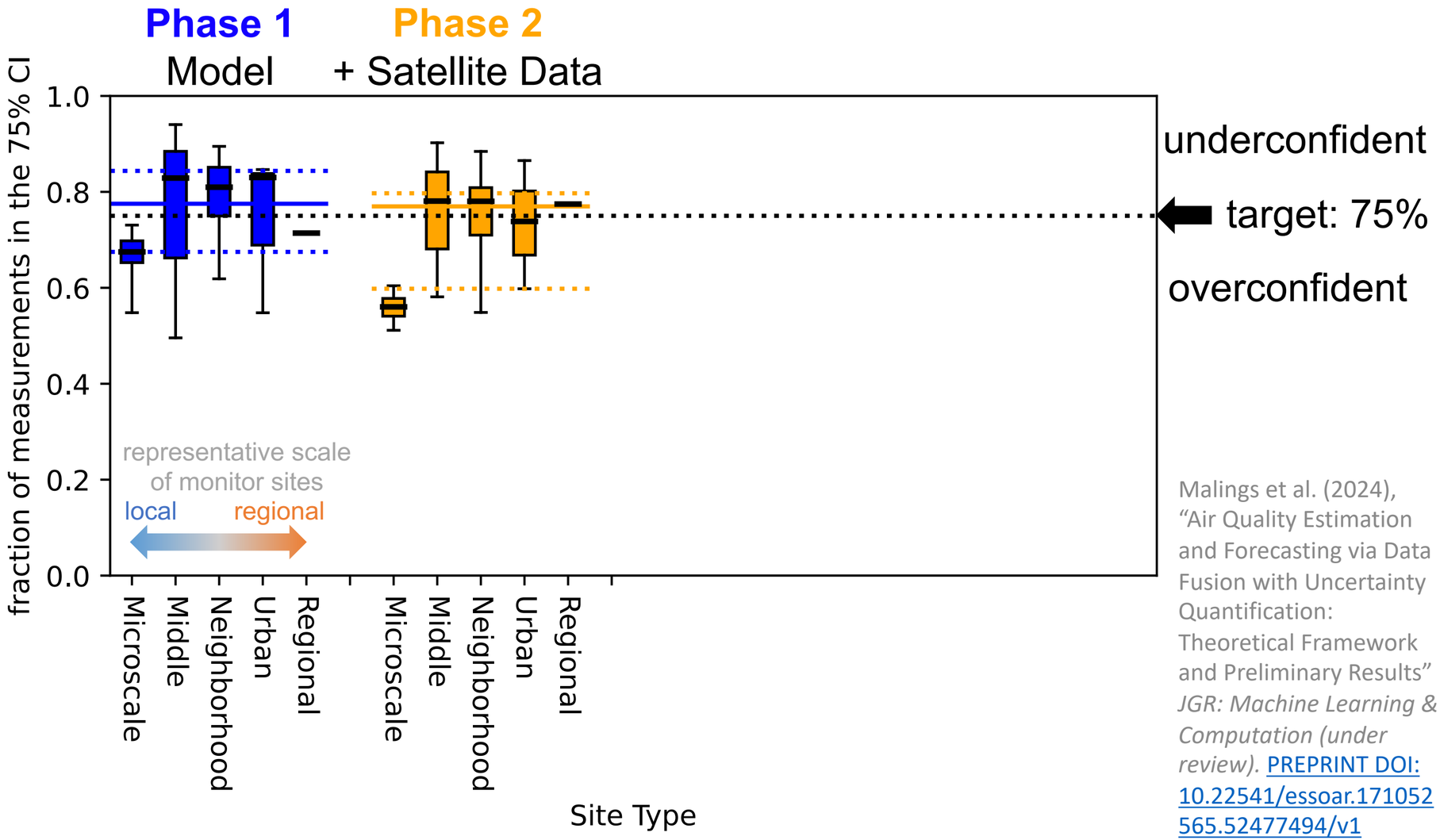
Partner

Case Study Details

San Francisco
September 2019
Surface NO_2
Lognormal distribution
Cross-validation test
25 ground monitors

Ground Sites





Malings et al. (2024), "Air Quality Estimation and Forecasting via Data Fusion with Uncertainty Quantification: Theoretical Framework and Preliminary Results" *JGR: Machine Learning & Computation (under review).* PREPRINT DOI: 10.22541/essoar.171052565.52477494/v1

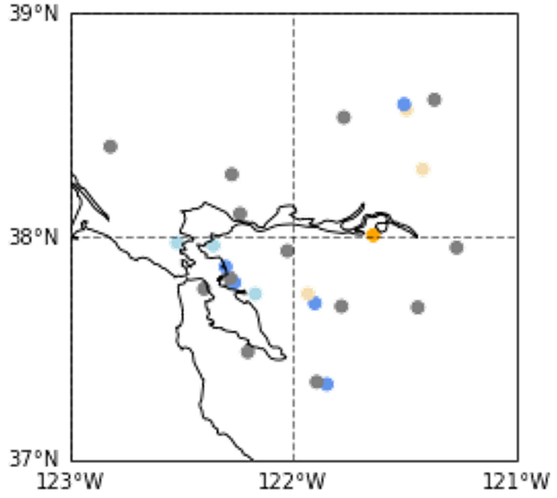
Case Study Results

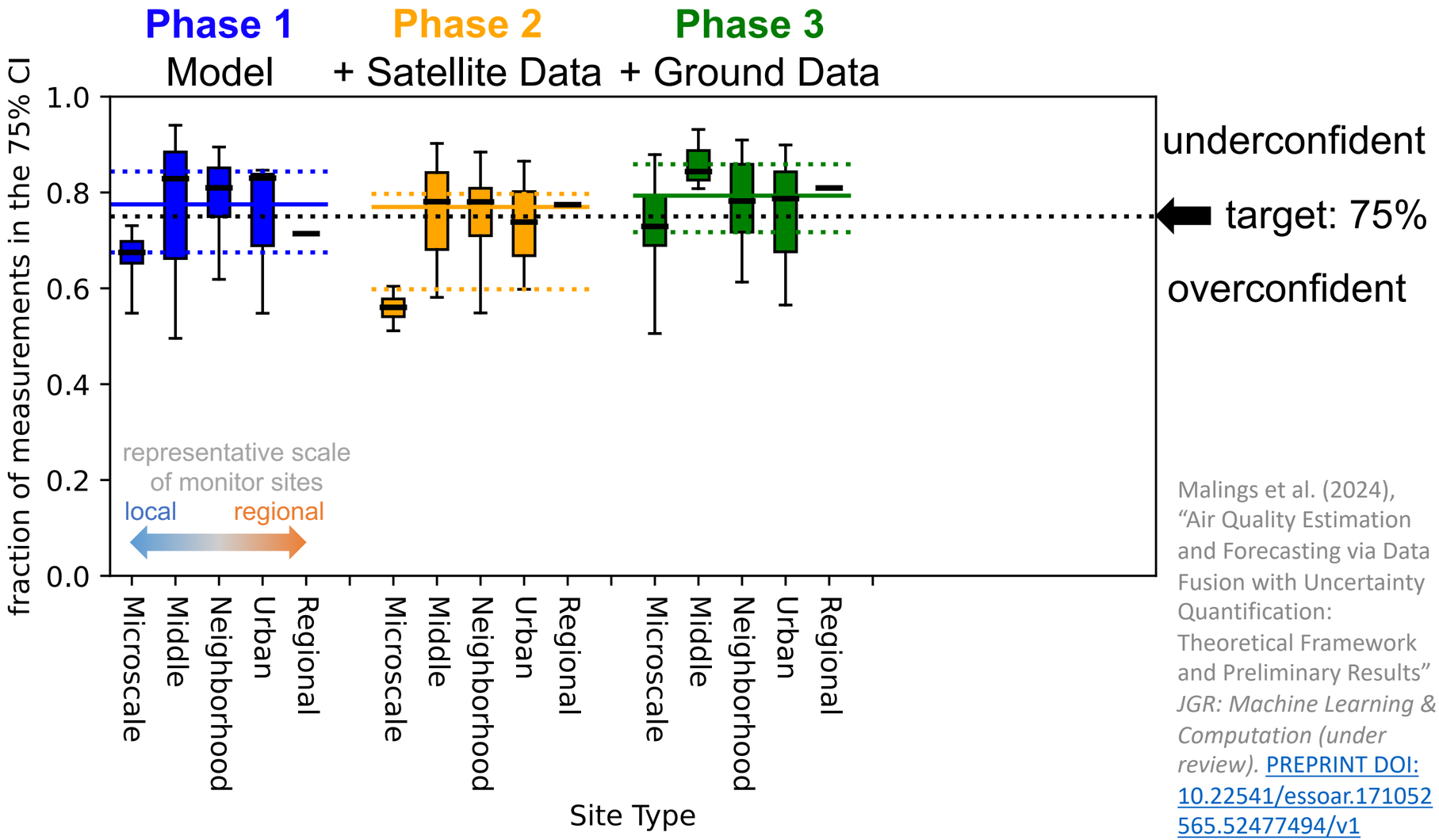
Partner

Case Study Details

San Francisco
September 2019
Surface NO_2
Lognormal distribution
Cross-validation test
25 ground monitors

Ground Sites





Malings et al. (2024), "Air Quality Estimation and Forecasting via Data Fusion with Uncertainty Quantification: Theoretical Framework and Preliminary Results" *JGR: Machine Learning & Computation (under review).* PREPRINT DOI: 10.22541/essoar.171052565.52477494/v1

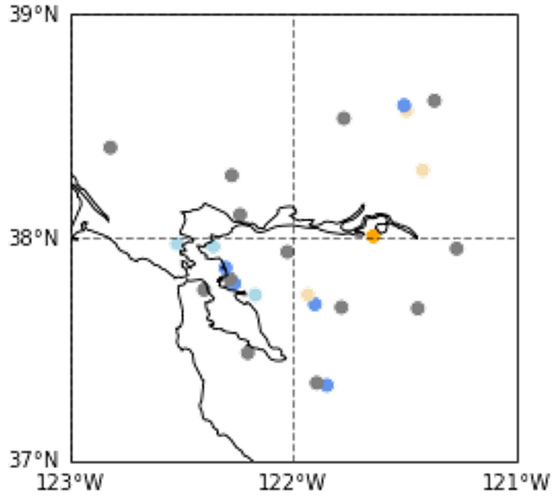
Case Study Results

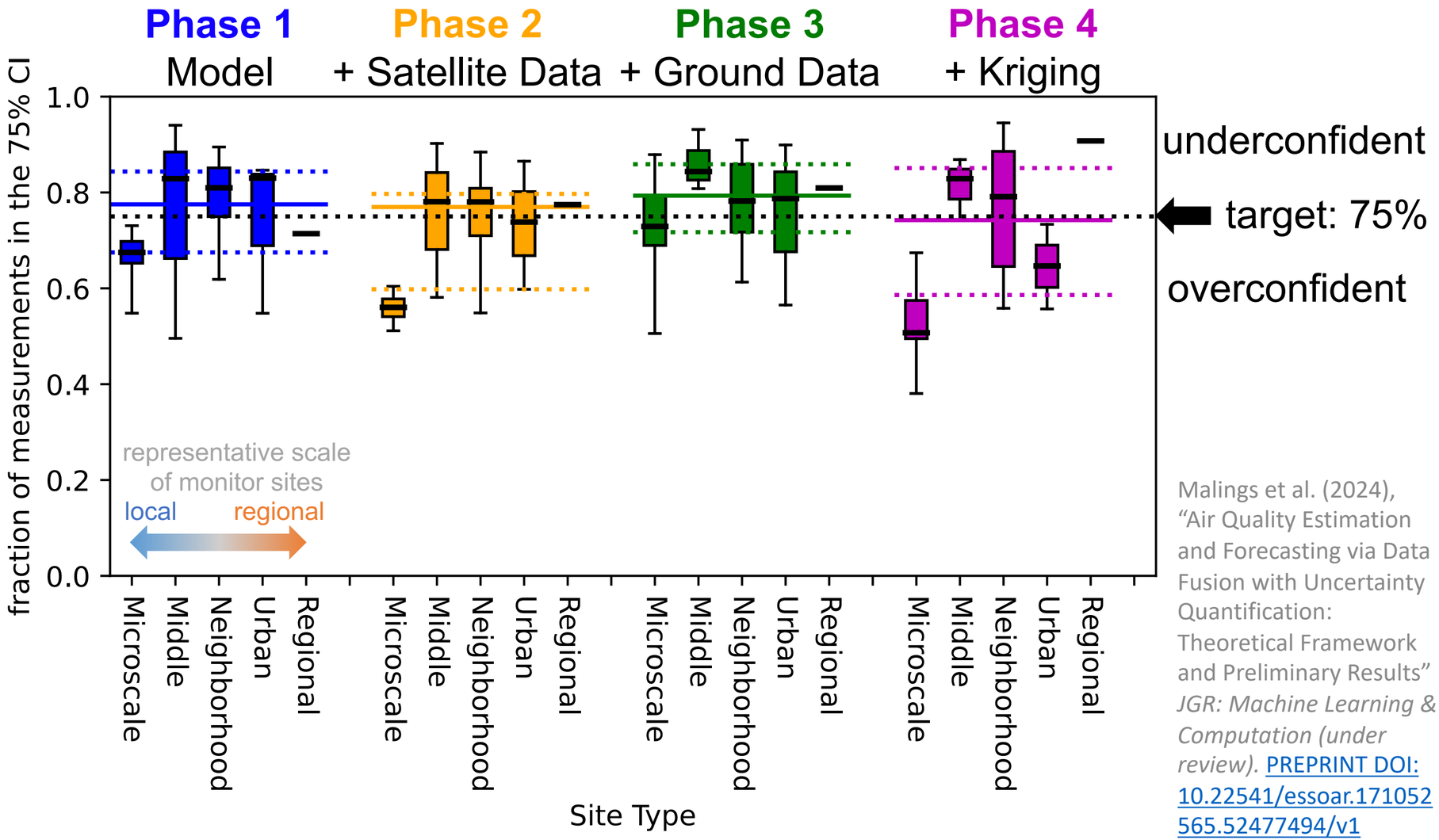
Partner

Case Study Details

San Francisco
September 2019
Surface NO_2
Lognormal distribution
Cross-validation test
25 ground monitors

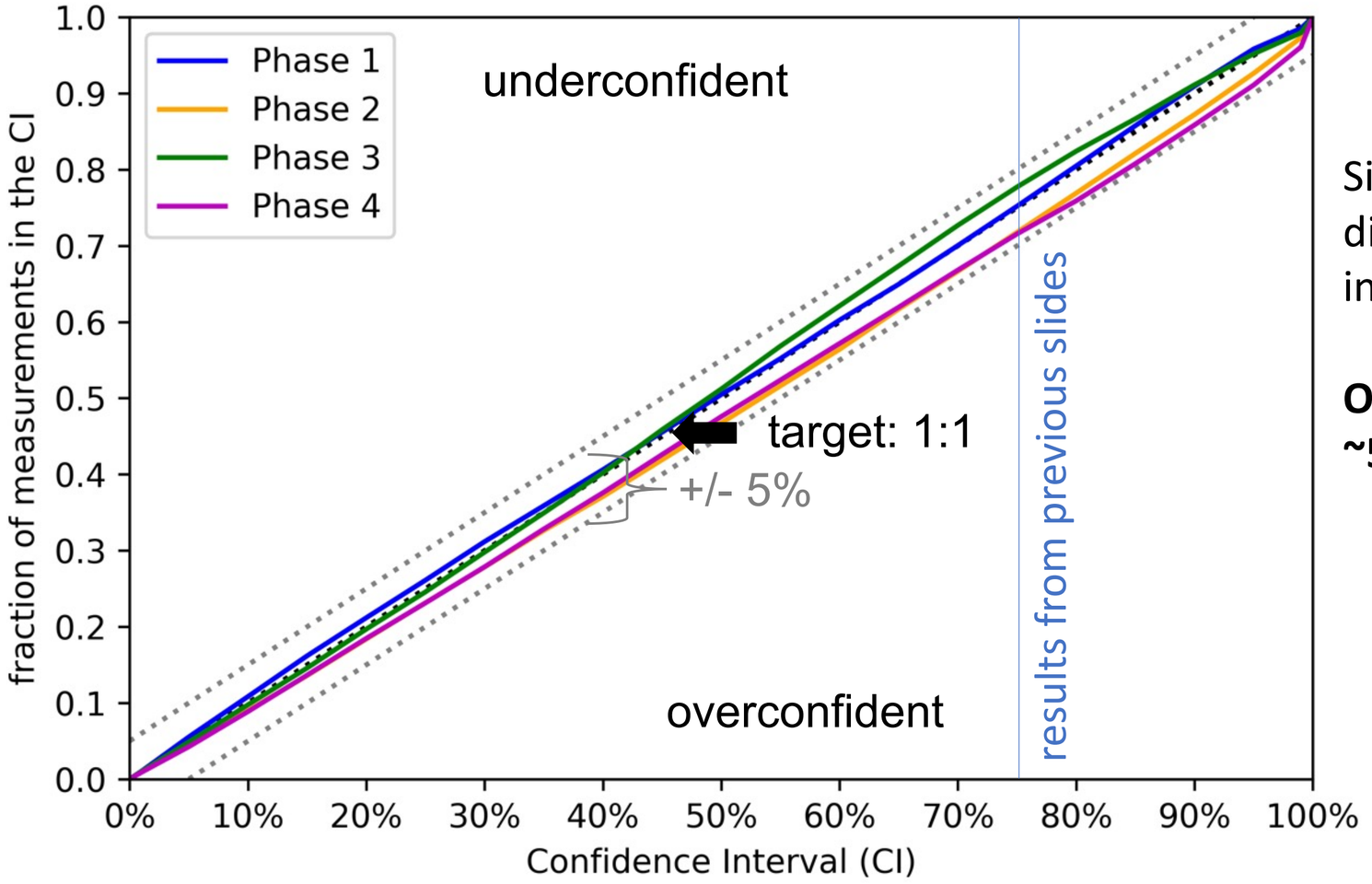
Ground Sites





Malings et al. (2024), "Air Quality Estimation and Forecasting via Data Fusion with Uncertainty Quantification: Theoretical Framework and Preliminary Results" *JGR: Machine Learning & Computation (under review).* PREPRINT DOI: 10.22541/essoar.171052565.52477494/v1

Effect of Different Confidence Intervals



Similar performance for different confidence interval definitions

Overall coverage within ~5% of target

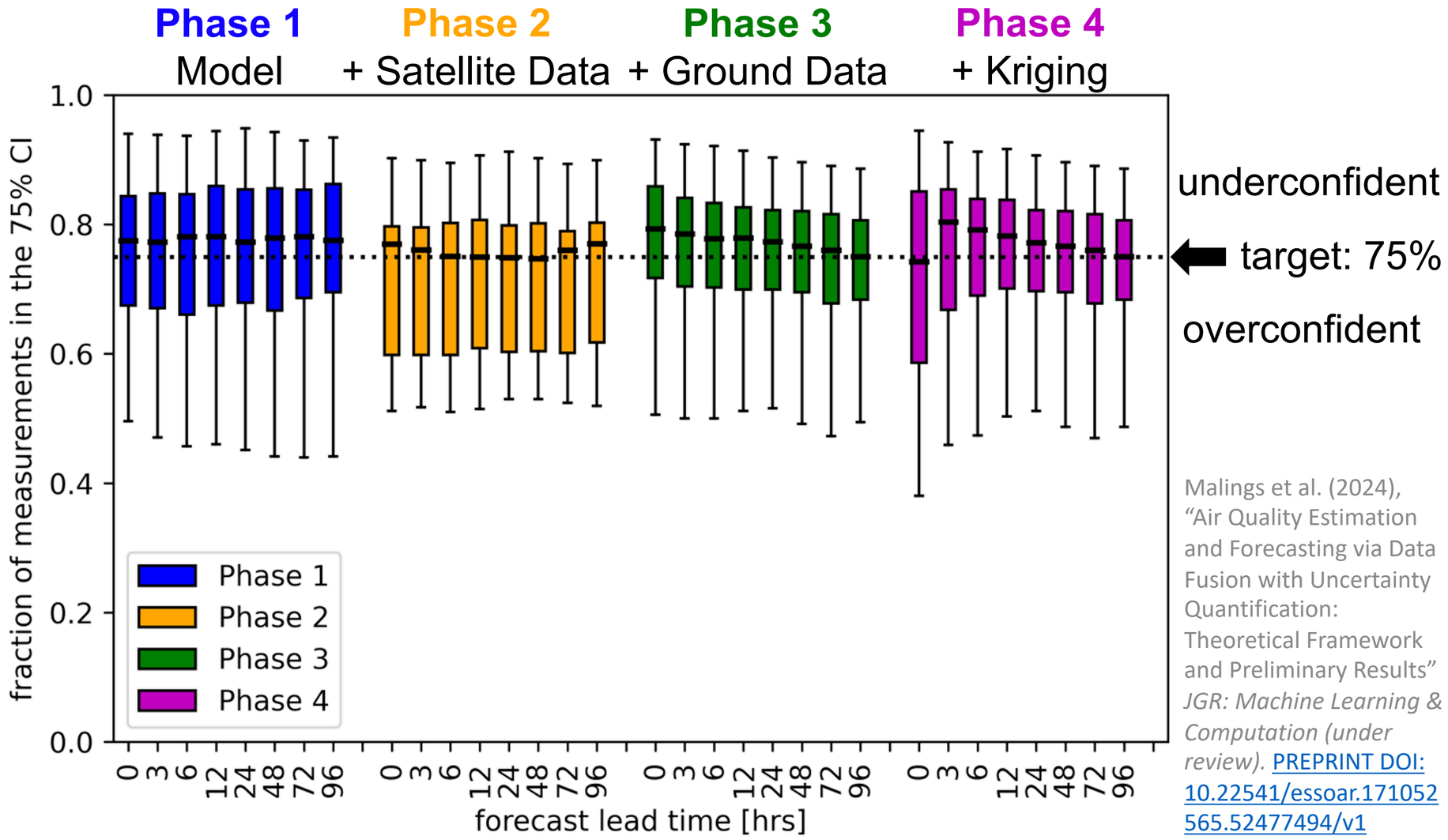
Malings et al. (2024), "Air Quality Estimation and Forecasting via Data Fusion with Uncertainty Quantification: Theoretical Framework and Preliminary Results" *JGR: Machine Learning & Computation (under review).* PREPRINT DOI: 10.22541/essoar.171052565.52477494/v1

Effect of Forecast Lead Time

Partner

Similar performance for different forecast lead times

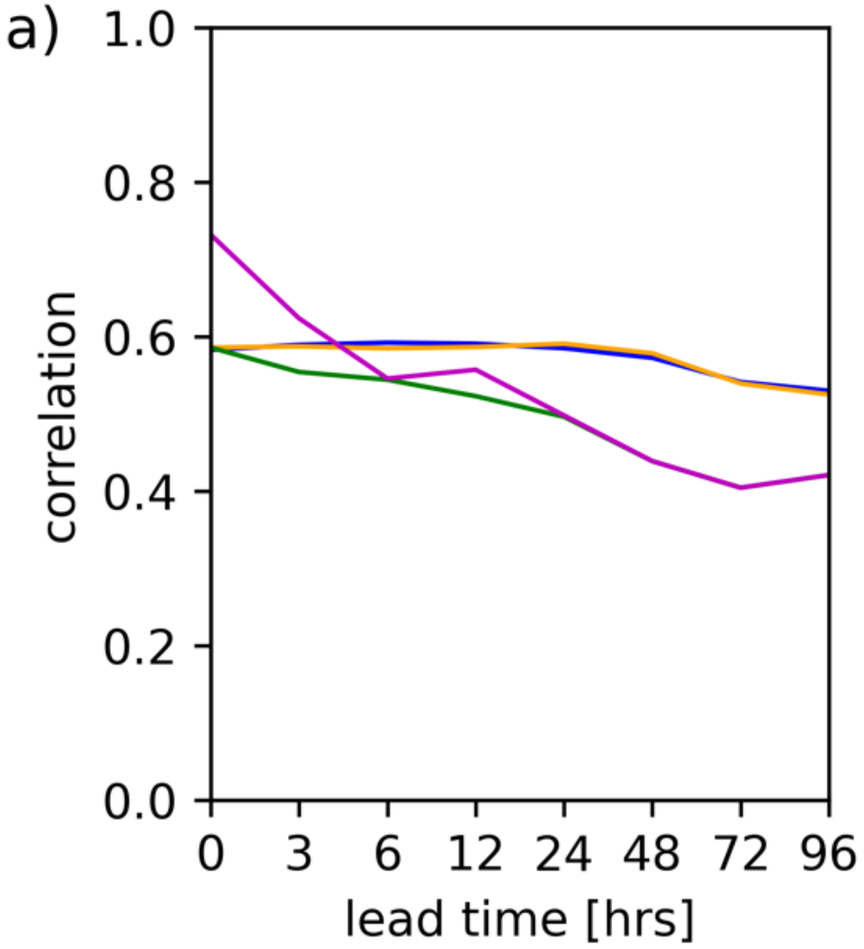
Widest spread in phase 4 coverage at 0 lead time

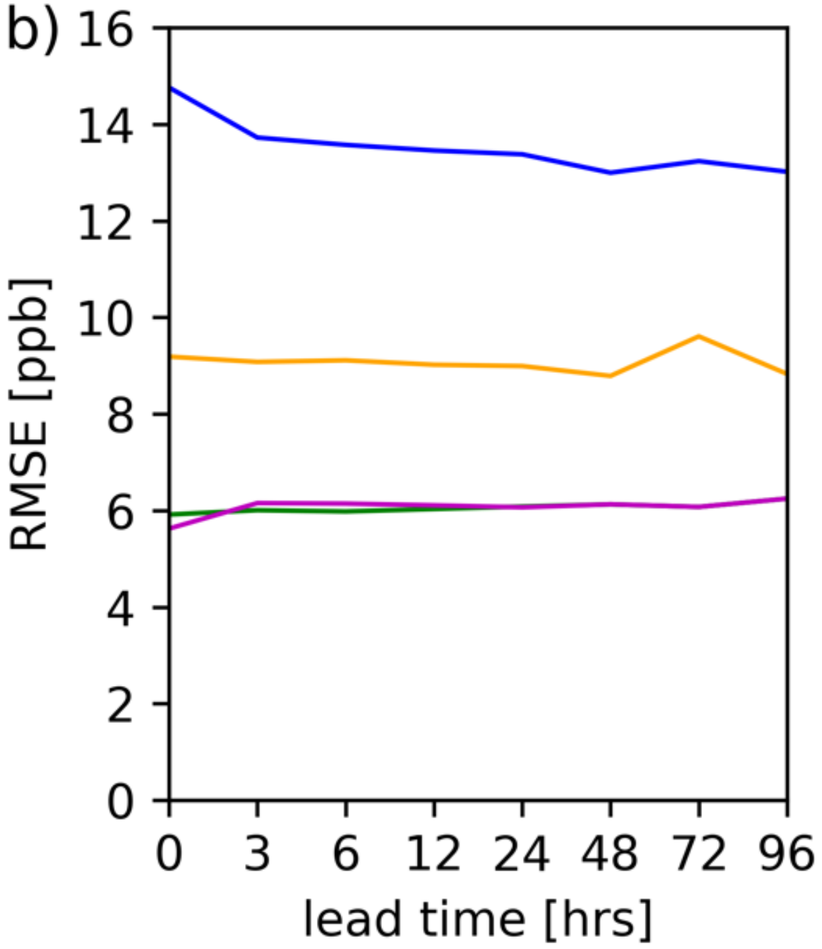


Malings et al. (2024), "Air Quality Estimation and Forecasting via Data Fusion with Uncertainty Quantification: Theoretical Framework and Preliminary Results" *JGR: Machine Learning & Computation (under review).* PREPRINT DOI: 10.22541/essoar.171052565.52477494/v1

Data Fusion Performance Metrics

Partner







Reduced errors as phase increases

Phases 3 & 4 degrade correlation, Phase 4 improves correlations for short-term forecasts

Previous work focused on performance assessment:

Malings et al. (2021), "Sub-City Scale Hourly Air Quality Forecasting by Combining Models, Satellite Observations, and Ground Measurements" *Earth & Space Science.* DOI: 10.1029/2021EA001743

Prototype User Interface

Partner

Earth Engine Apps

Search places

Sub-city air quality forecasts

This application visualizes hourly NO_2 concentration forecasts for different locations around the world. Forecasts are generated daily for the next day using the NASA GEOS-CF forecast, satellite retrievals from the TROPOMI instrument, and ground-level air quality measurements from reference-grade monitors and low-cost sensors. Reference-grade monitor data are shown with purple squares, while low-cost sensors are shown using green circles. Currently, all times are displayed in UTC/GMT.

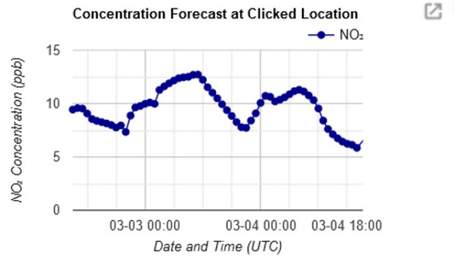
Select the region, input data, and time of interest to view forecasts

Rio de Janeiro, BR

Model, satellite, and interpolated ground monitors

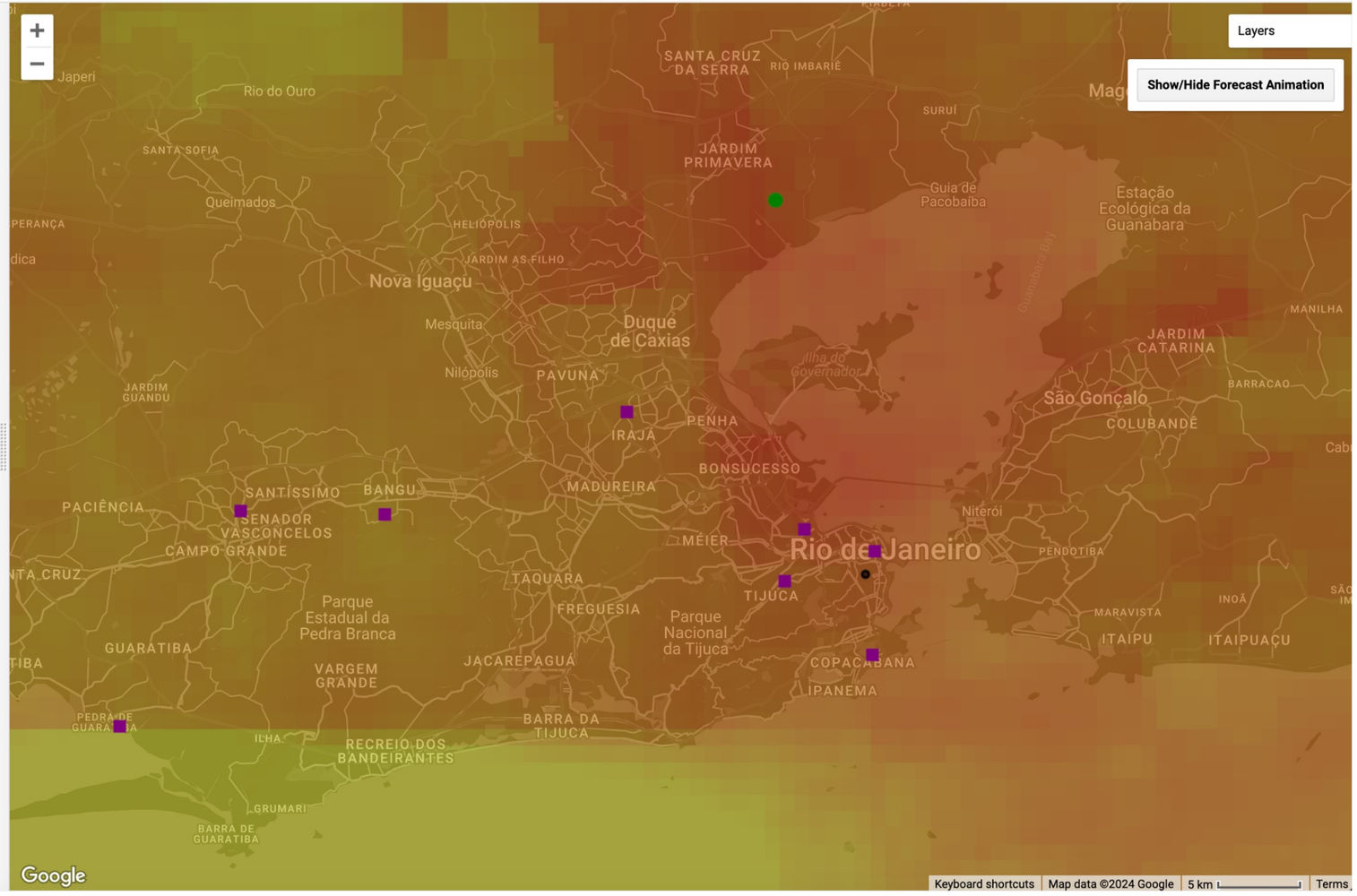
2024-03-04 08:00

lat: -22.92 lon: -43.18



NO_2 Concentration (ppb)

0 10 20



Malings et al. (2024), "Air Quality Estimation and Forecasting via Data Fusion with Uncertainty Quantification: Theoretical Framework and Preliminary Results" *JGR: Machine Learning & Computation (under review).* PREPRINT DOI: 10.22541/essoar.171052565.52477494/v1

Conclusions & Ongoing Work

Partner

- Theoretical
  - Better approach to uncertainty quantification near sources
    - Include ancillary data, experiment with non-linear (machine learning) methods
  - Better approach to uncertainty quantification at Phase 4
    - Non-isotropic correlation functions?
  - Incorporating low-cost air quality sensors
    - Possibility to regionally re-calibrate sensors based on Phase 3 outputs

- Practical
  - Implement data fusion system in Google Earth Engine
    - Efficiency improvements needed!
  - Design the user interface
    - **How to display uncertainty in an intuitive way?**

Thank you!

Questions?

This material is based upon work supported by the National Aeronautics and Space Administration (NASA) under Grants 80NSSC22K1473 and WBS 389018.02.09.02.72 issued through the NASA Health and Air Quality Applied Sciences Program.

Proposal 1: Expanding Data Fusion Work

Partner

- Expand to new regions
- Estimate new pollutants
  - PM10
  - Methane
- Include new satellite missions
  - TEMPO, GEMS
  - MAIA
- Develop new data visualization & analysis tools
  - Better ways to convey forecast uncertainty
  - Use uncertainty to guide monitor network expansions
- Funding Mechanism: NASA HAQAST
  - NOI: June 4
  - Proposal due: July 9
  - Budget: ~$100K
  - Limited funding for technical development, but access to a wider network of future users

Proposal 2: Assessing Air Sensor Representativity

Partner

- Problem: Difficult to assess data quality of air sensors while they are deployed
  - Especially in areas without a lot of monitoring density to begin with
- Solution(?): Use satellites (with support from models?) as another standard of comparison
  - Comparing patterns and trends can provide corroboration for in-situ data
  - Quantitative assessment of data quality is probably not possible, but qualitative checks might be sufficient
  - Satellites can provide evidence of specific conditions (e.g., dust events) when sensors might perform poorly
- Objective: develop “screening tools” for sensors by comparing with satellite data
  - Especially important to determine under what circumstances such comparisons are and are not valid.
- Deliverables
  - Prototype data comparison and screening tools
  - Training for AQ managers in how to use the tools
  - General guidance about how to appropriately compare satellite and in-situ data
- Funding Mechanism: NASA HAQ
  - NOI: July 8
  - Proposal due: August 6
  - Budget: ~$250K
  - Larger scope for technical development of new methods, but still must focus on specific end-user needs

Proposal 3: Supporting Exceptional Event Demos

Partner

- NASA systems (GEOS-CF) have supported exceptional event demonstrations
  - e.g., stratospheric ozone intrusion
- Can this process be standardized/systematized? What other capabilities might be developed?
  - Are existing visualization and data access tools sufficient? Are new portals needed for the existing model?
  - As a global model, GEOS-CF might better represent international transport
  - Enabling/disabling different emission sources in counterfactual model runs
  - Tracers have recently been implemented in GEOS-CF
  - Data assimilation has recently been implemented; what can the increments tell us about exceptional events?
- Funding Mechanism: NASA MAP
  - PI: K. Emma Knowland
  - NOI: May 31
  - Proposal due: TBD
  - Budget: ~$200K
  - Targeting “Earth System Modeling for Earth Science to Action” research theme