Air Quality Data Fusion with Sensors, Satellites, and Models

Carl Malings

Morgan State University & GESTAR-II cooperative agreement

NASA Global Modeling and Assimilation Office

GESTAR II

GMAO

GODDARD EARTH SCIENCES

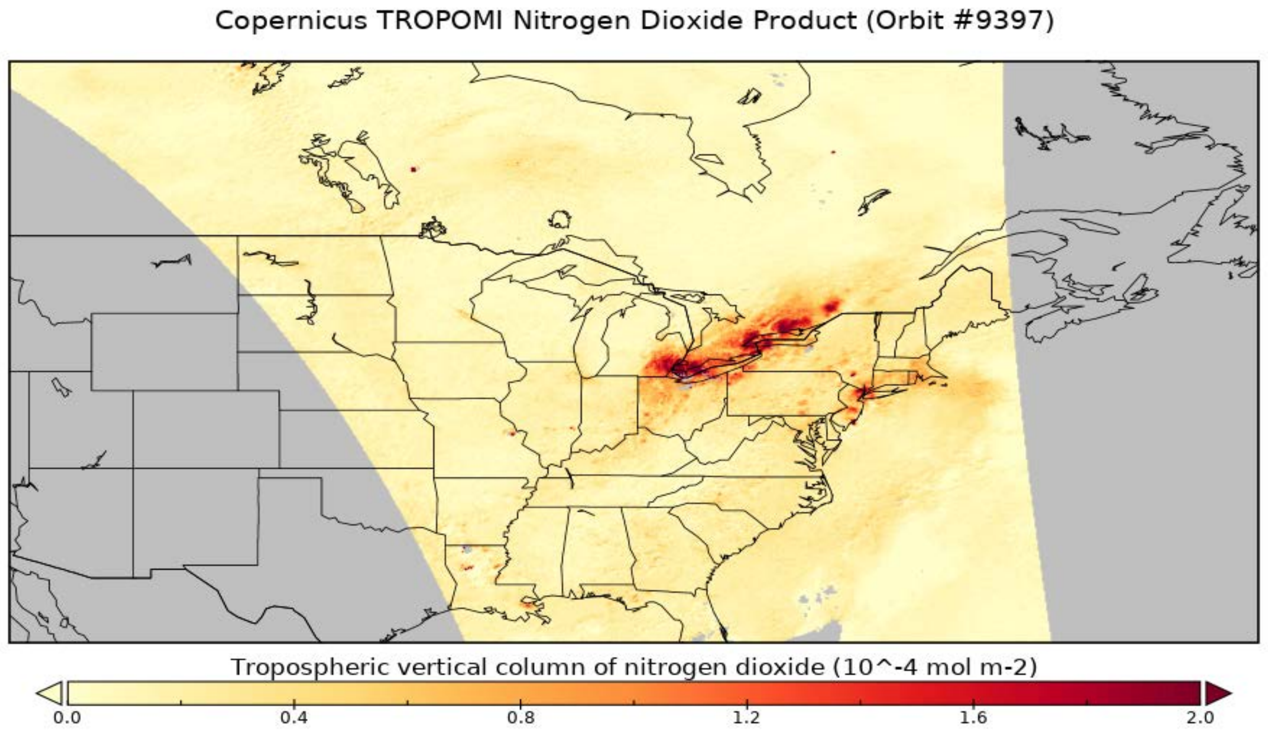
NASA Partner

- Motivation
  - Combining multiple sources of air quality data
  - NASA-funded project to support air quality managers
  - The advantages of uncertainty quantification
- Data Fusion Approach
  - Phase 1: model only
  - Phase 2: bring in satellite data
  - Phase 3: bring in historical ground monitor data
  - Phase 4: bring in near-real-time ground monitor data
  - Quantifying uncertainty and defining confidence intervals
- Case Study Results
  - Impacts of site-to-site differences
  - Impacts of different confidence intervals
  - Impacts of forecasting lead times
- Conclusions & Ongoing Work

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

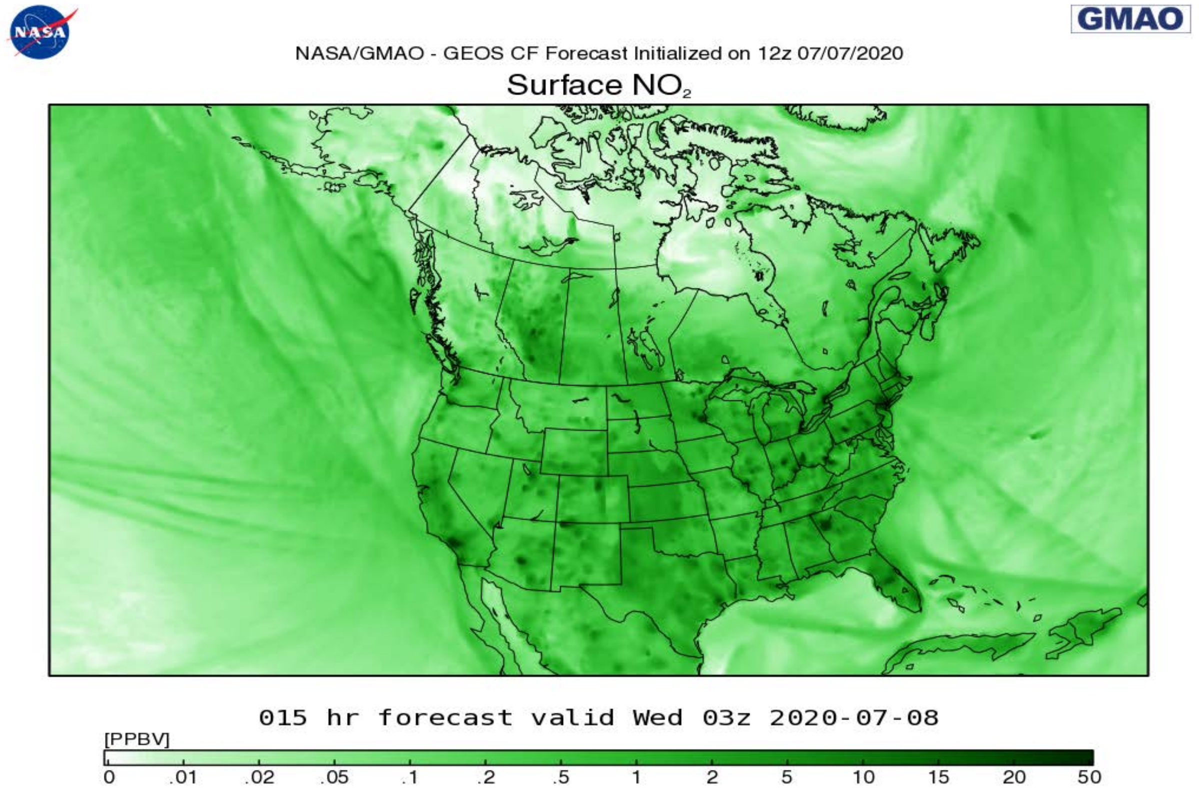
+ relatively inexpensive
+ dense/remote deployment
- greater noise and bias

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

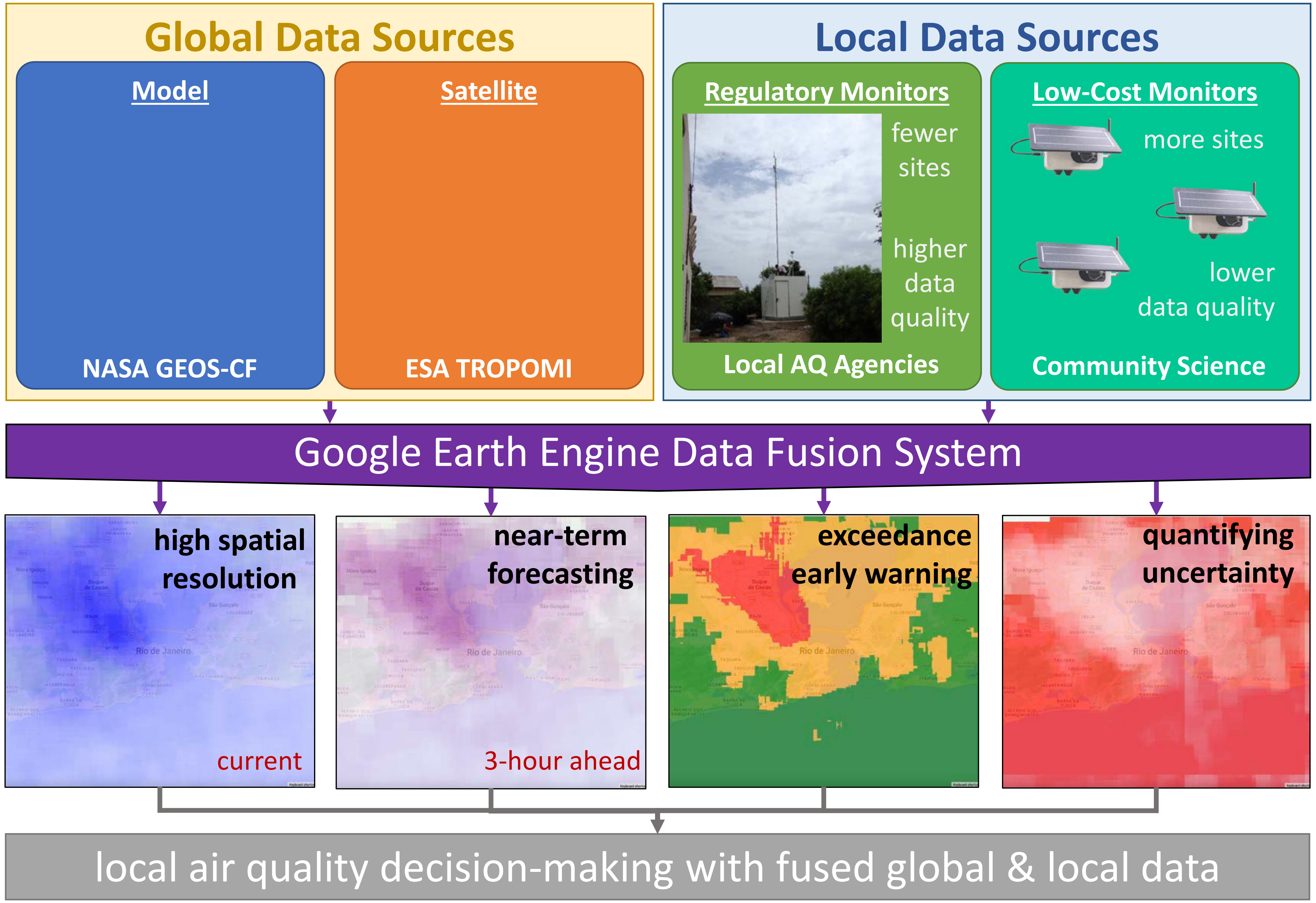


…integrate diverse **global** and **local** air quality data sources…

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

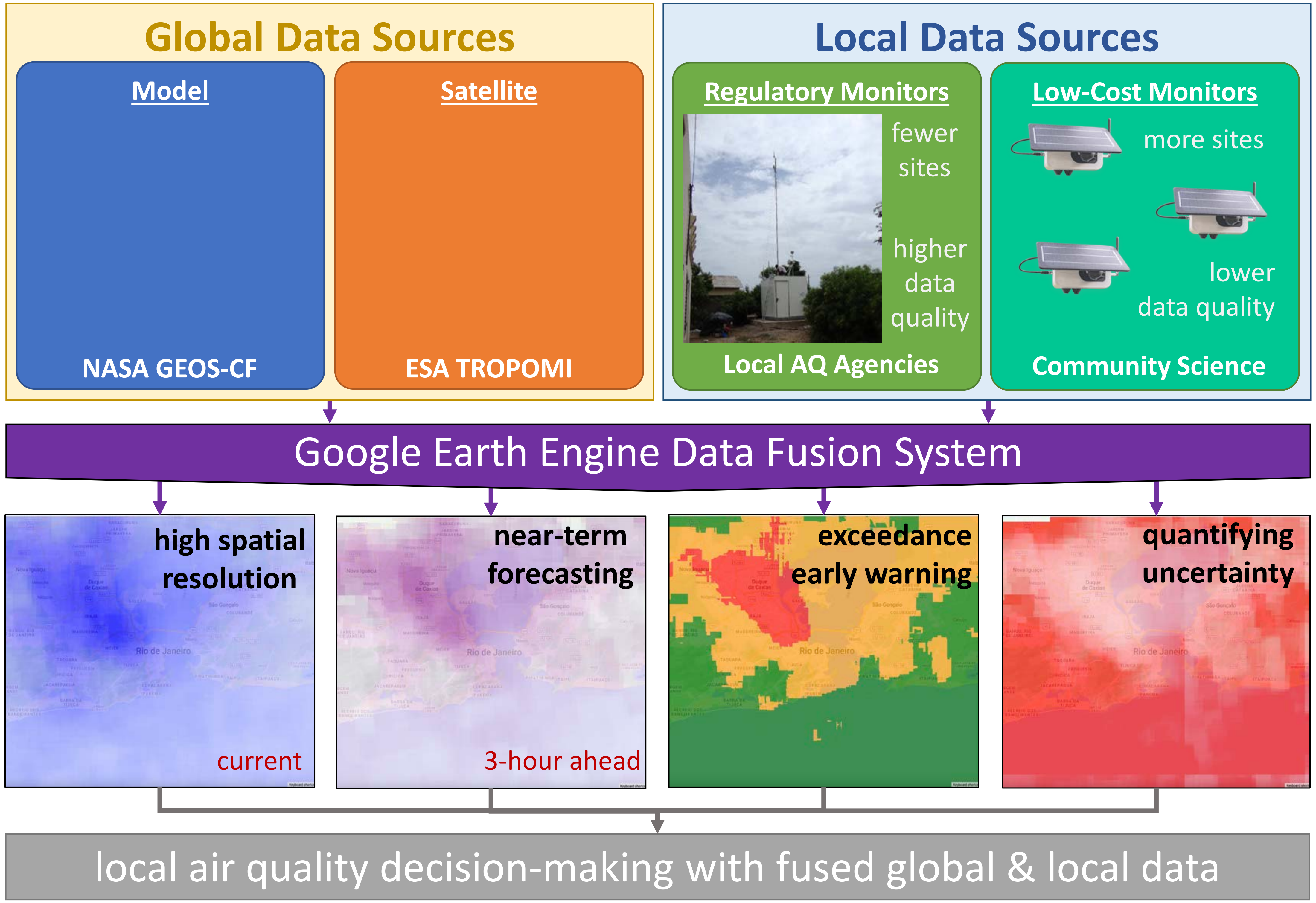
…to provide synthesized **estimates** and **forecasts** of air quality at a **local scale** but with a **global scope**…

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

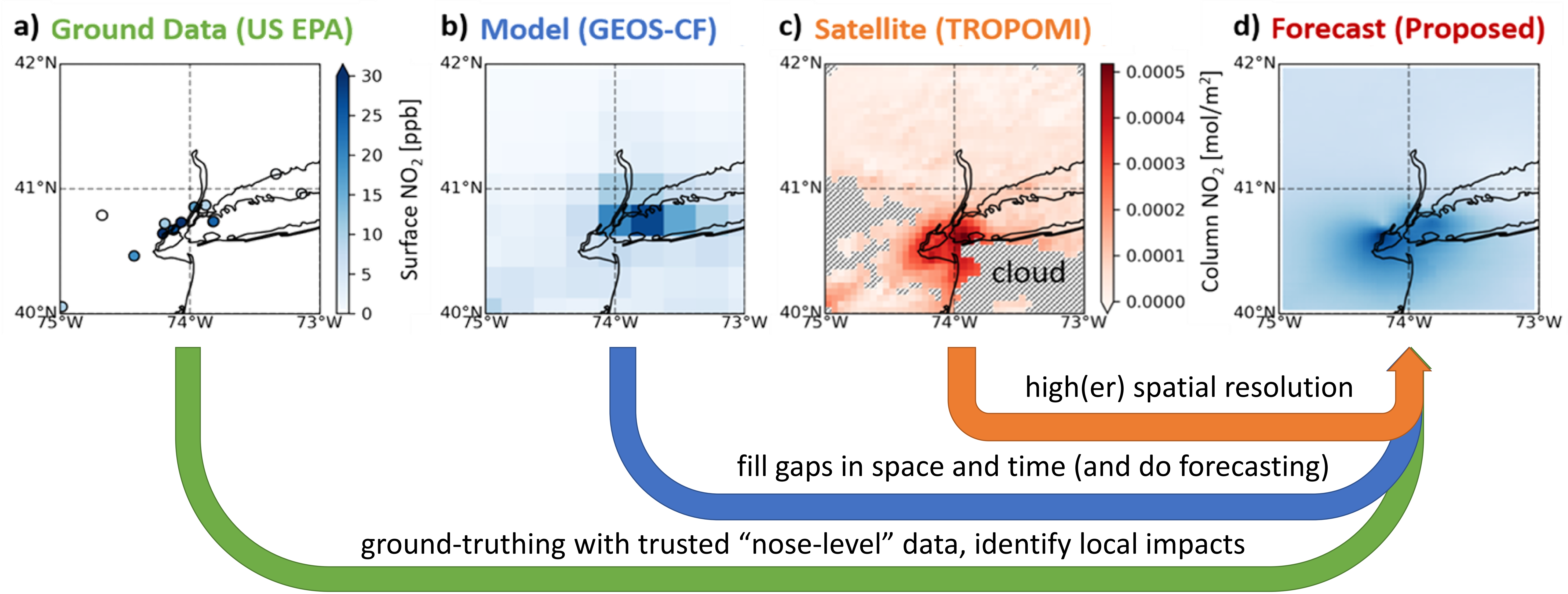


Source: NASA GMAO Science Snapshot "Google Earth Engine Data Fusion Tool to support Air Quality Managers"

- **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"



Source: 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

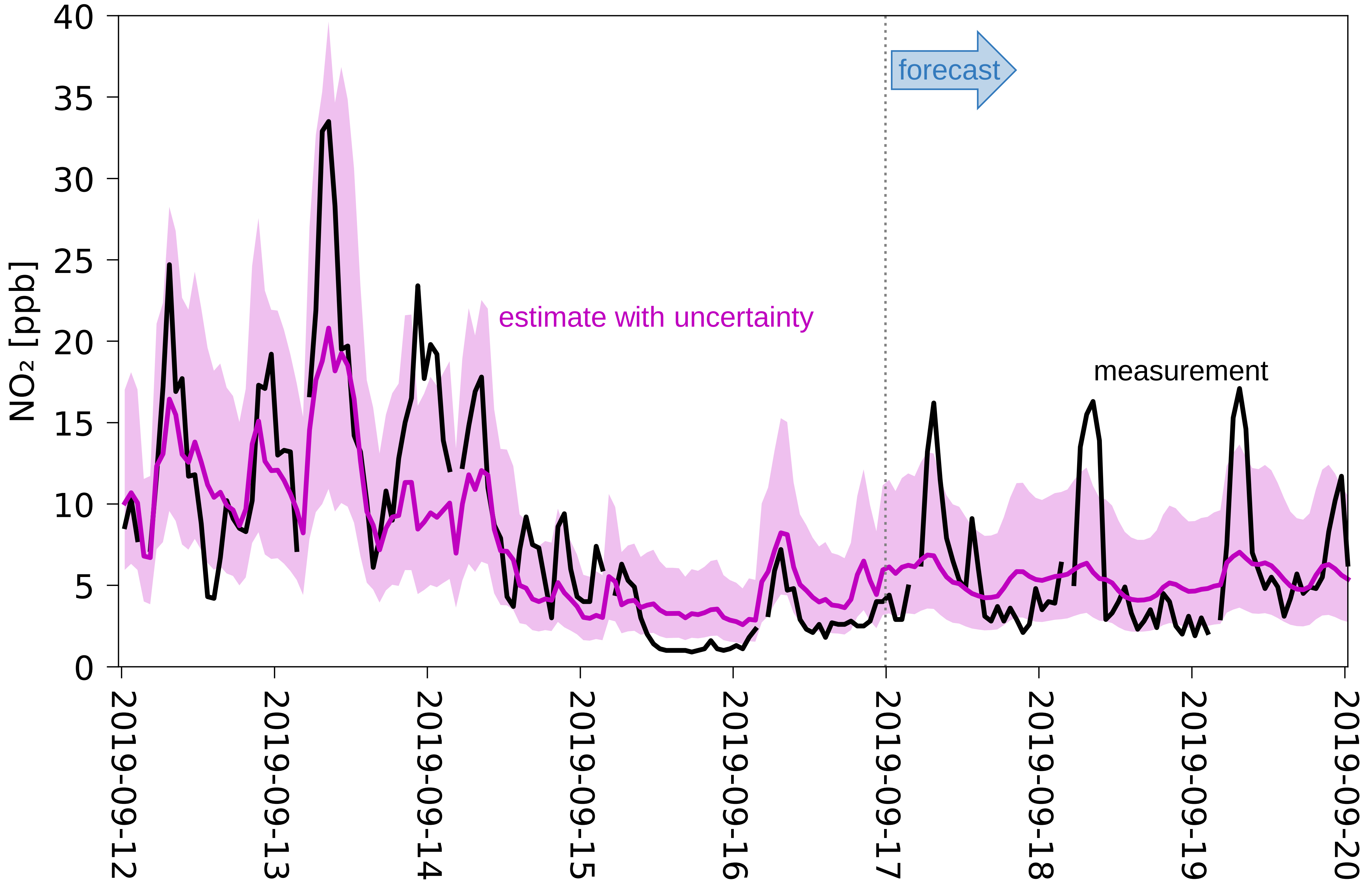
Provide a prior estimate of the relative confidence in a forecast

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

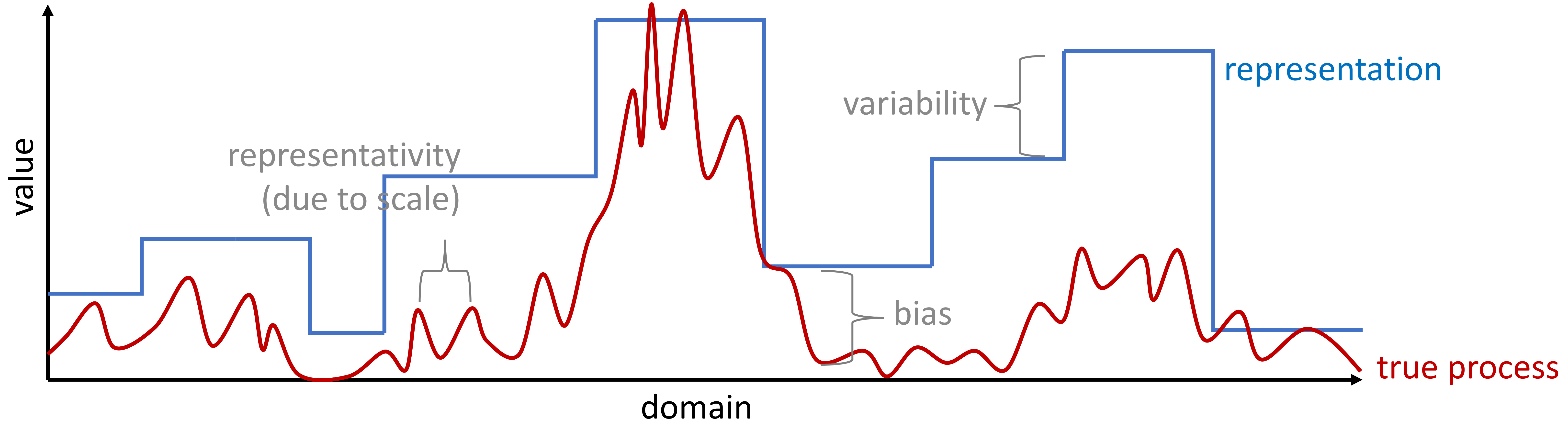
Identify a range of likely outcomes

Quantify the impacts of different data sources in reducing uncertainties

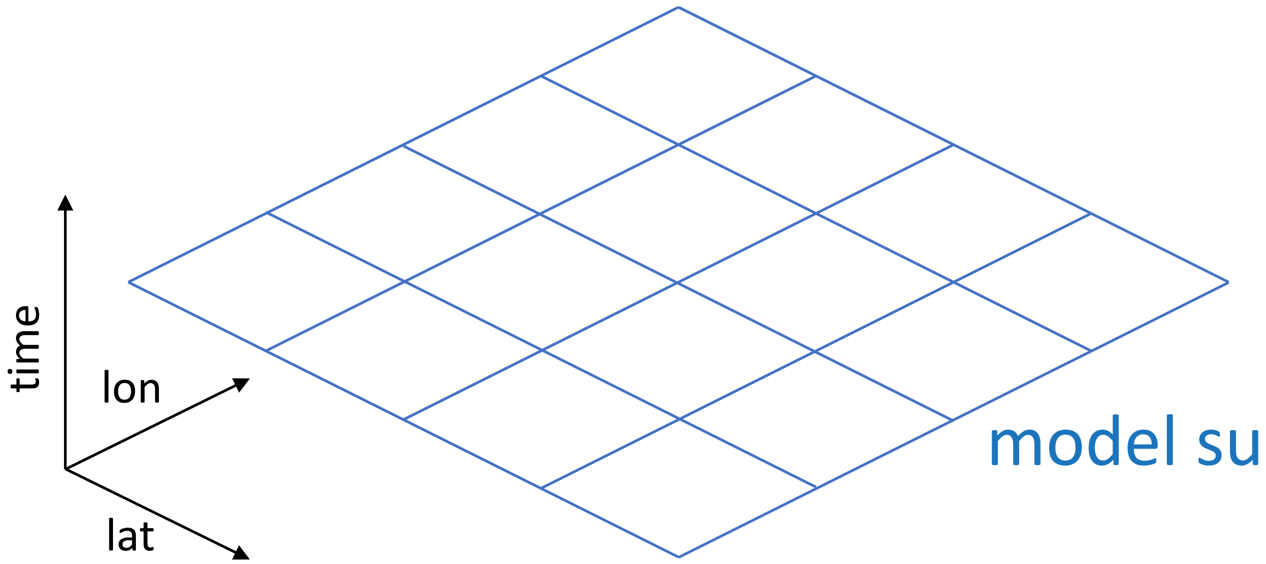
Identify the potential to reduce uncertainties through additional data collection



- Uncertainty – Overall characterization of potential errors in reproducing a process
  - Bias – Systematic errors in reproducing a process
  - Variability – Random errors in reproducing a process
  - Representativity – Errors in representing a process due to mismatched resolution

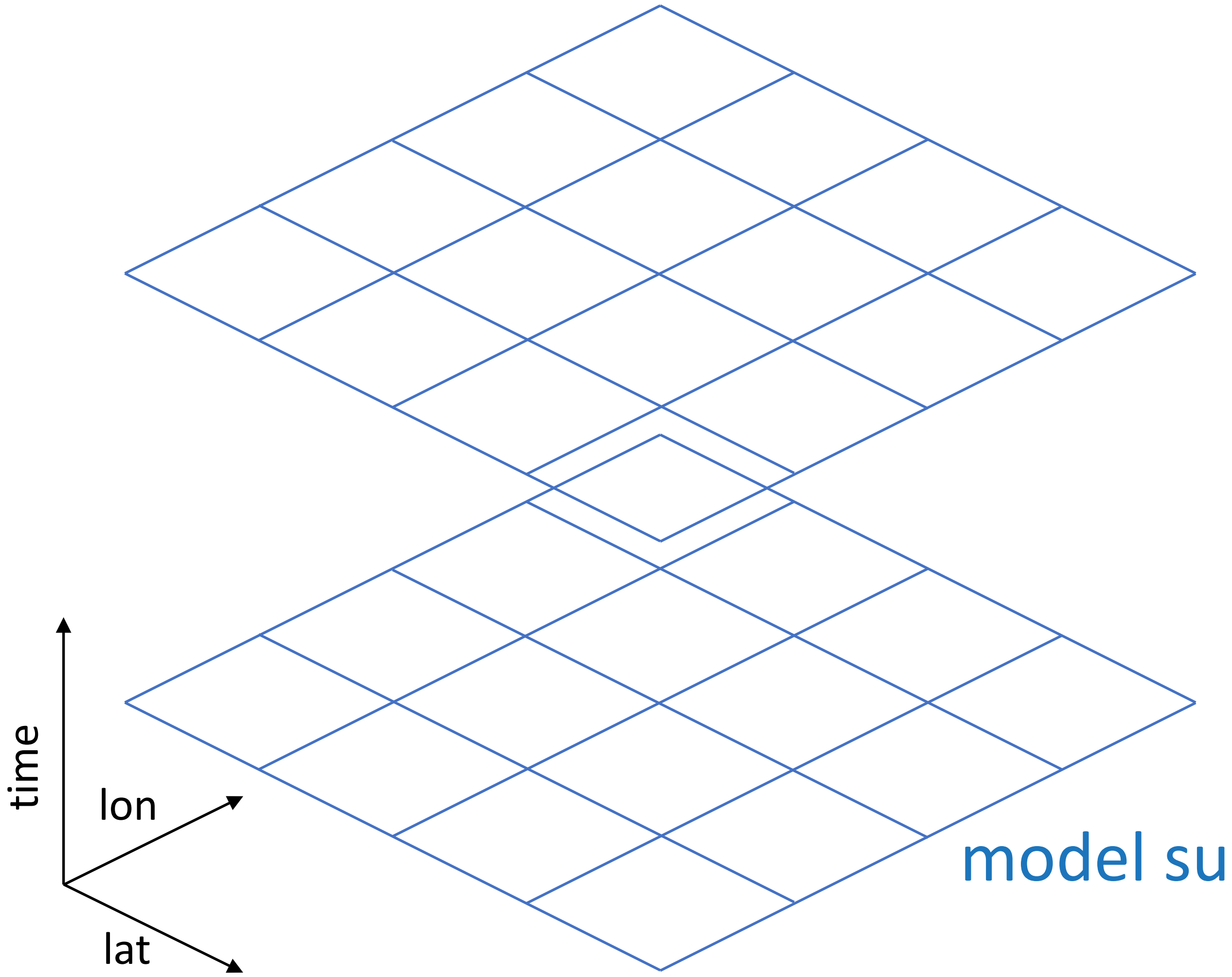


Phase	Estimate
1	forecast model (GEOS-CF)



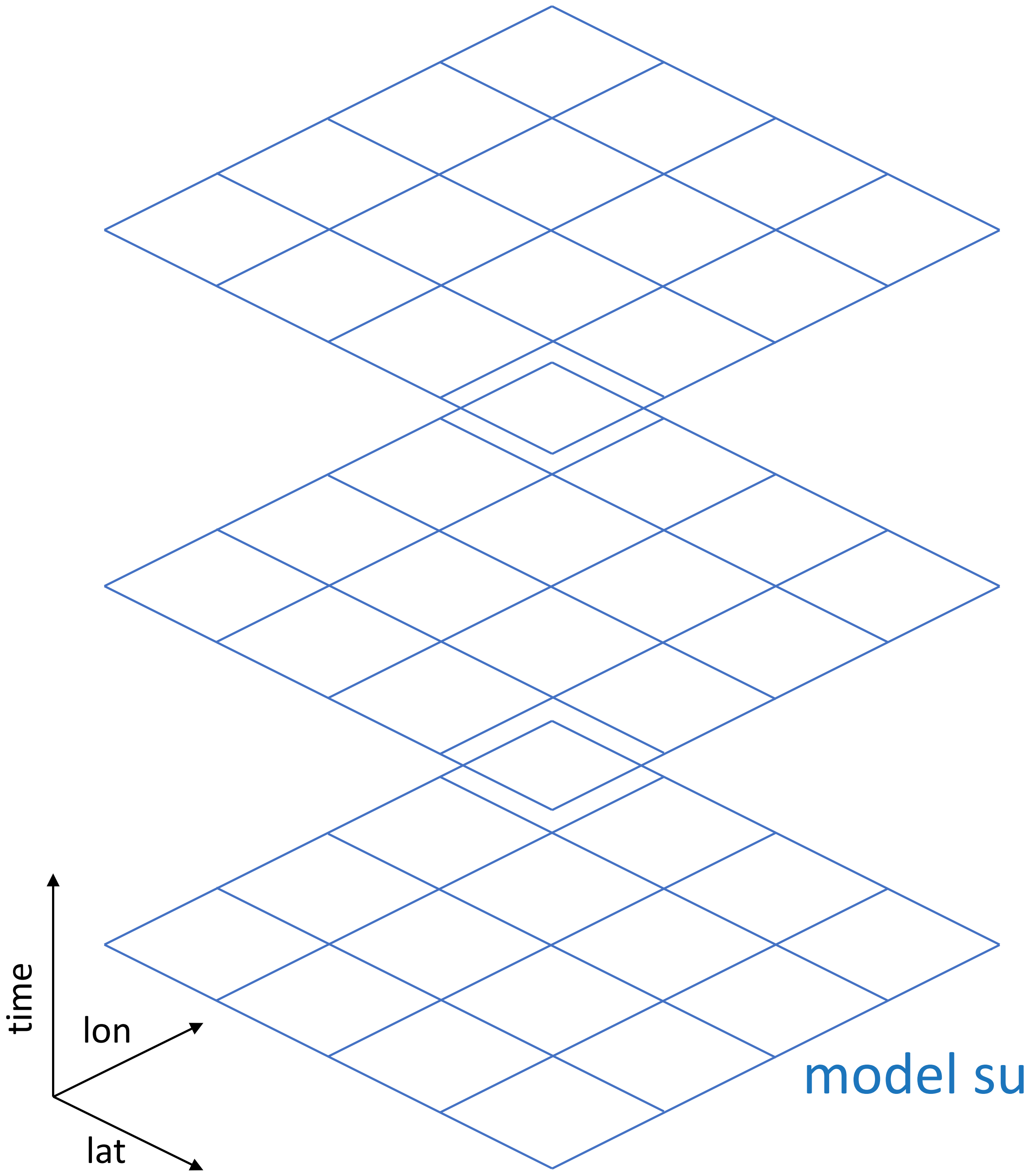
model surface concentration $M(x,t)$

Phase	Estimate
1	forecast model (GEOS-CF)



model surface concentration $M(x, t)$

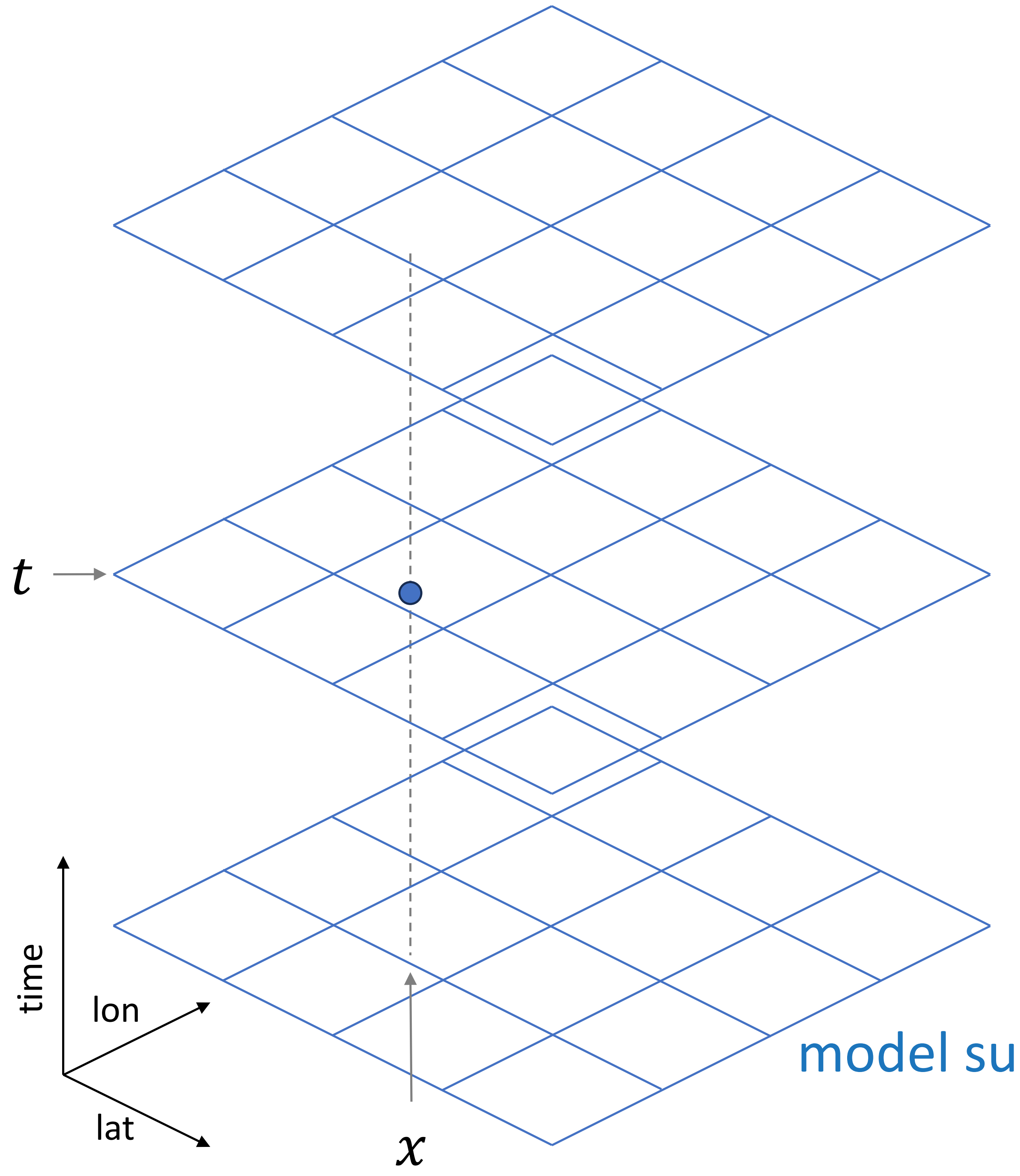
Phase	Estimate
1	forecast model (GEOS-CF)

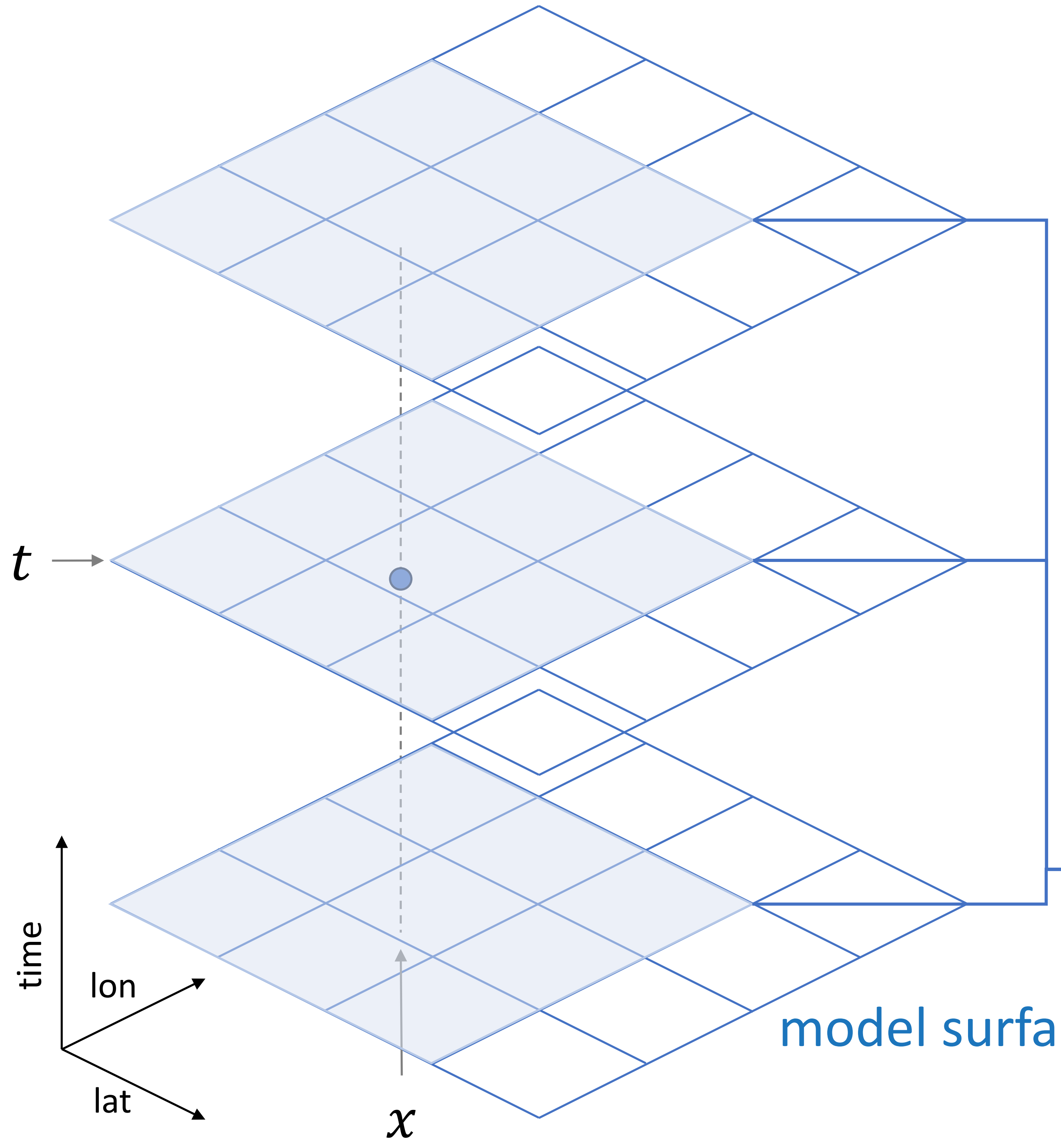


model surface concentration $M(x,t)$

Phase	Estimate
1	forecast model (GEOS-CF)

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





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

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

data fusion uncertainty (variance) at phase 1

$$V_1(x,t) = V_{F1}(x,t,\tau)$$

uncertainty due to forecasting by τ ahead (ignore this for now…)

$$+V_M(x,t)$$

uncertainty due to model internal variability

$$+V_{B1}(x,t)$$

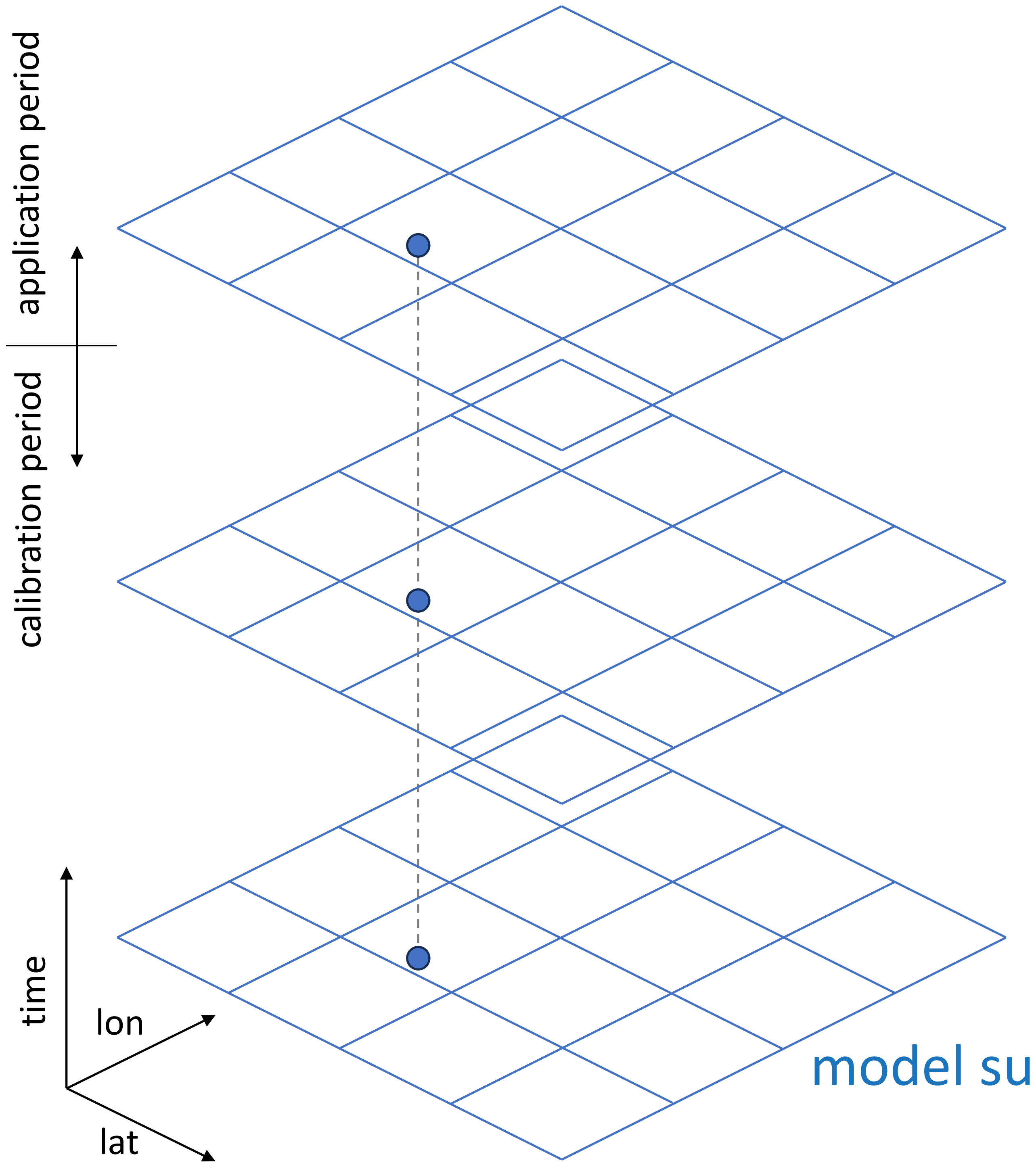
uncertainty due to model bias

$$+\,V_{R1}(x,t)$$

uncertainty due to spatial representativity (model scale)

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

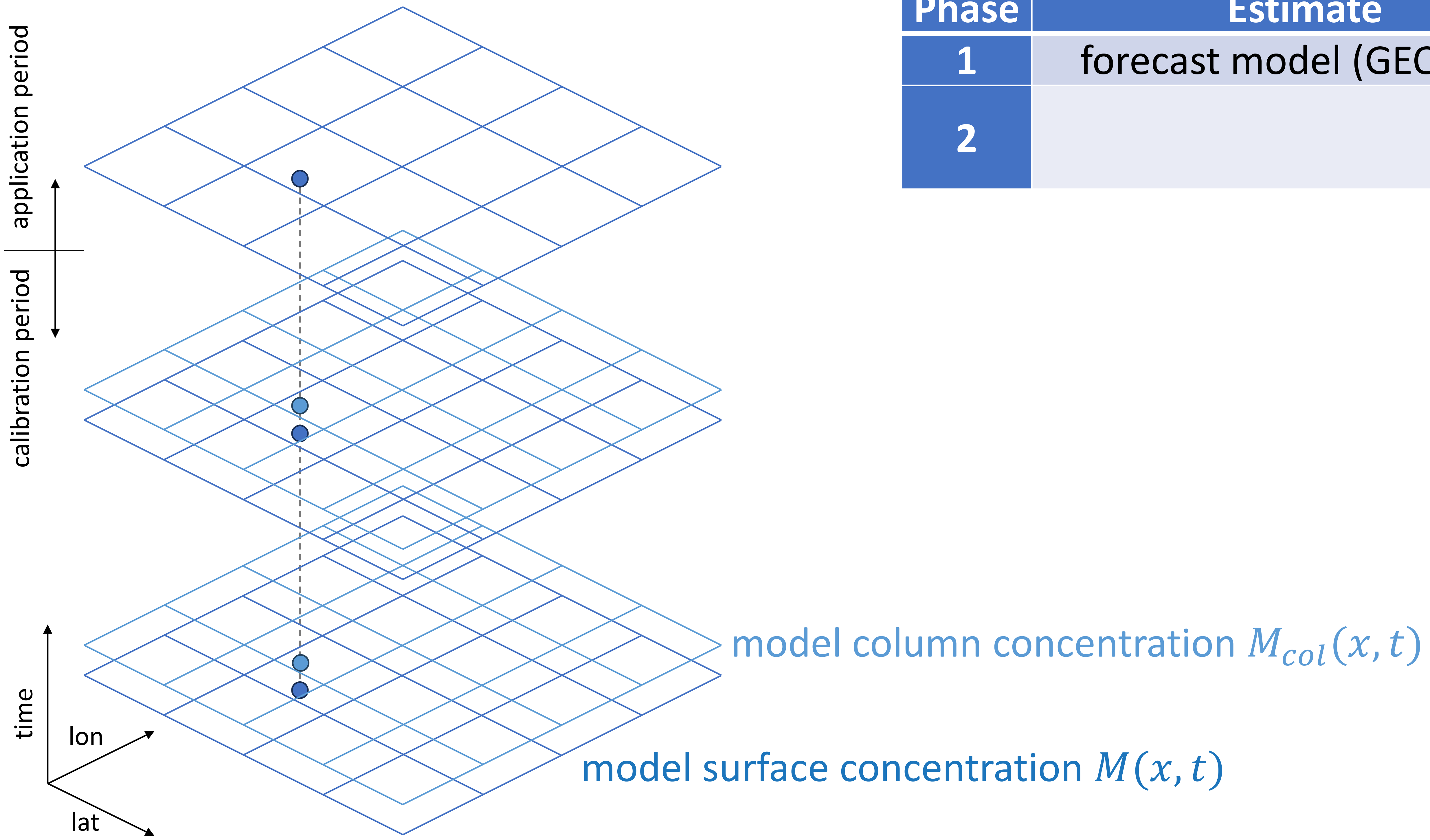
model surface concentration $M(x,t)$



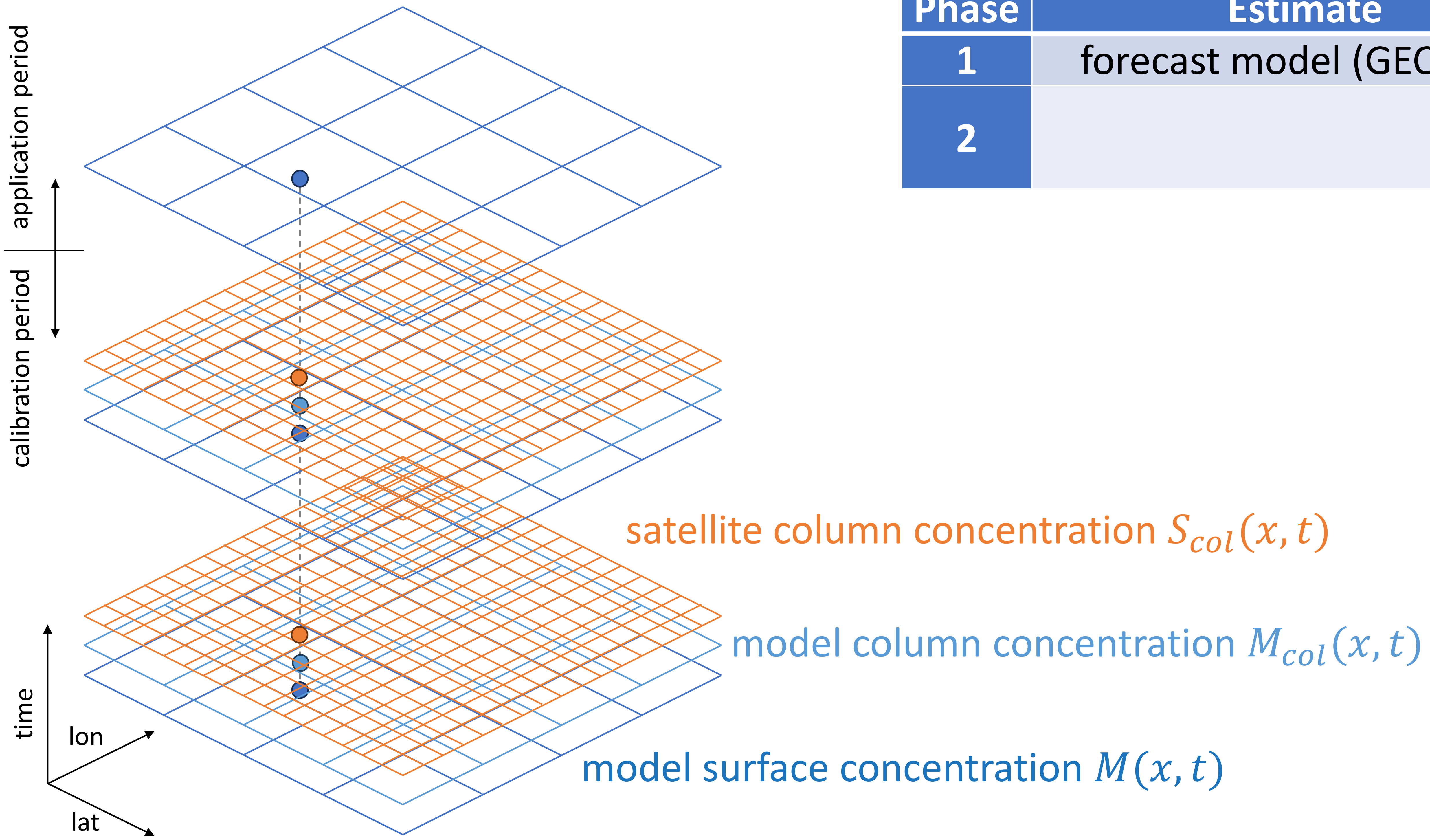
Phase	Estimate	Uncertainty
1	forecast model (GEOS-CF)	cell-to-cell variability of model
2		

model surface concentration $M(x, t)$

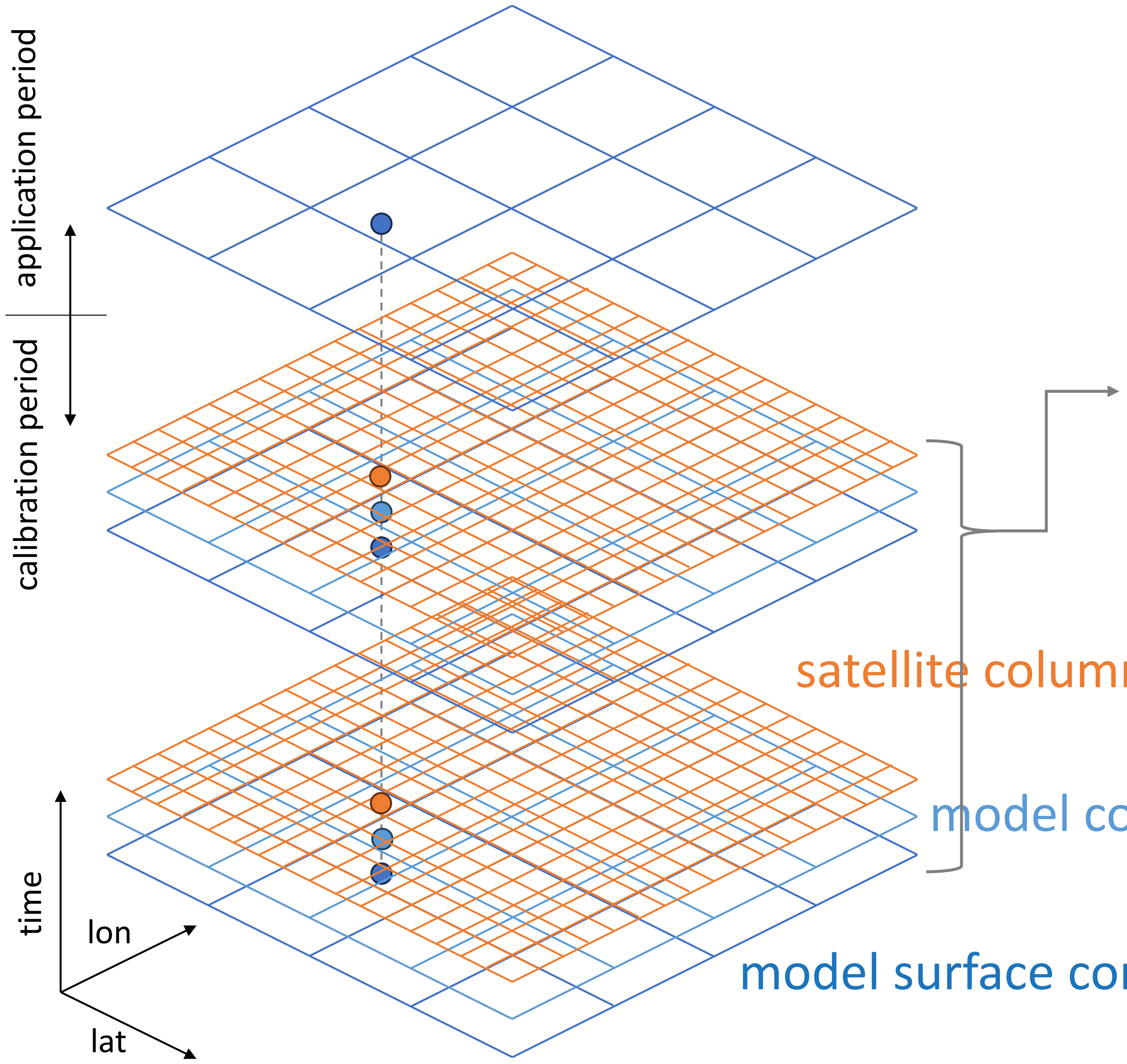
Phase	Estimate	Uncertainty
1	forecast model (GEOS-CF)	cell-to-cell variability of model
2		



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



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



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

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

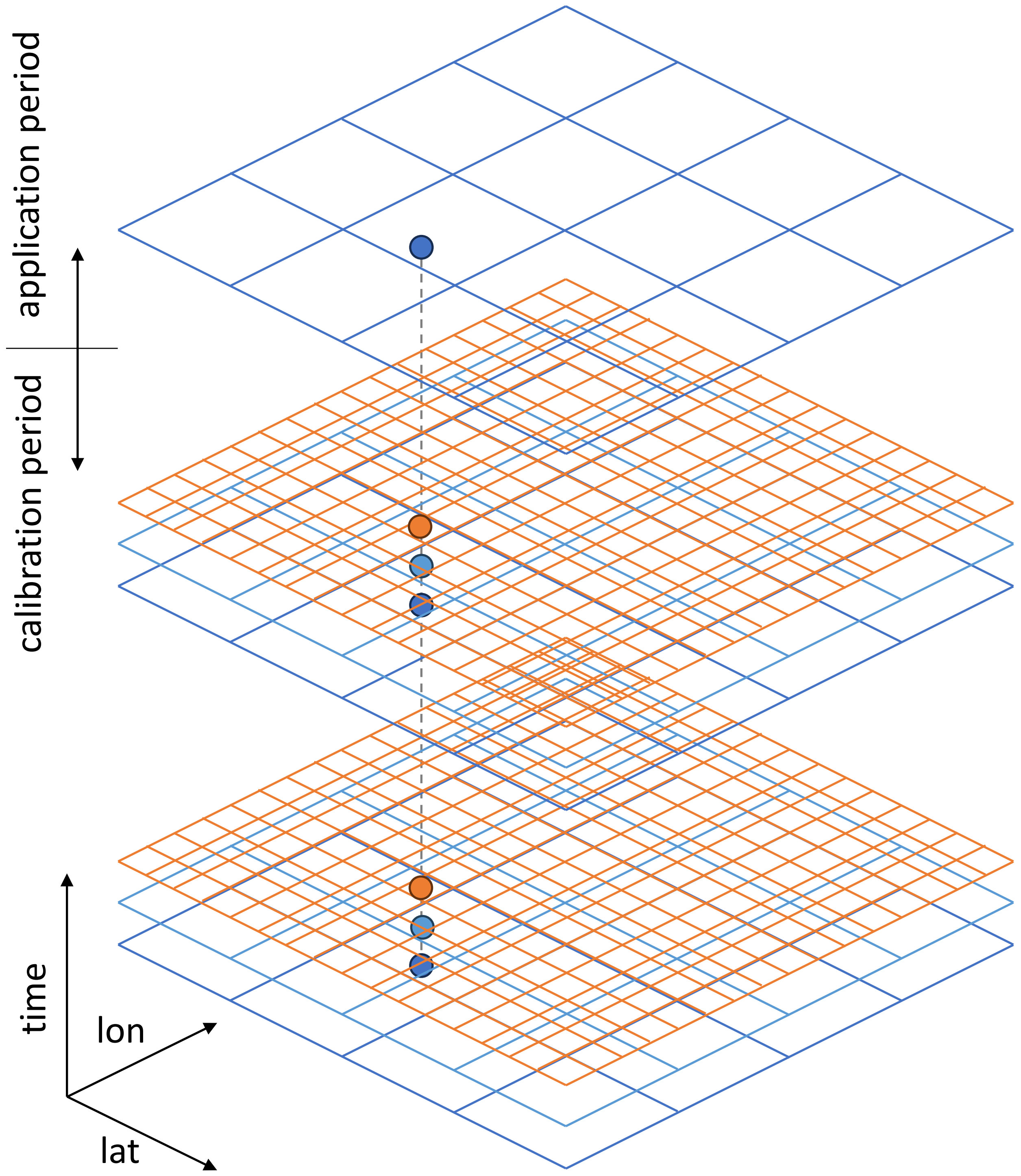
satellite column concentration $S_{col}(x,t)$

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	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

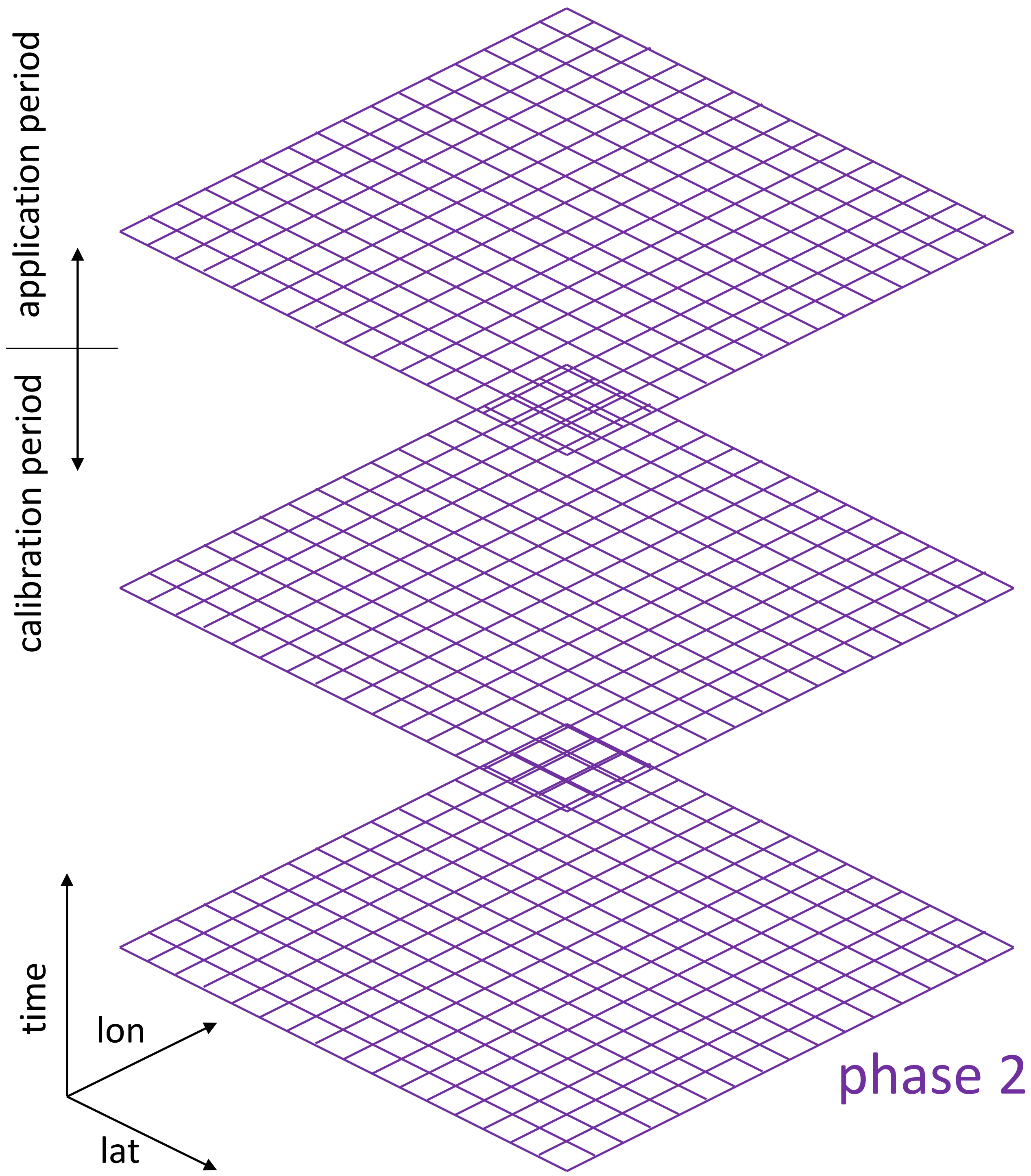
$+2V_{MD}(x,t)$ ⟵ co-variance of satellite-to-model differences with model outputs

$+V_{B2}(x,t)$ ⟵ uncertainty due to model & satellite bias

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

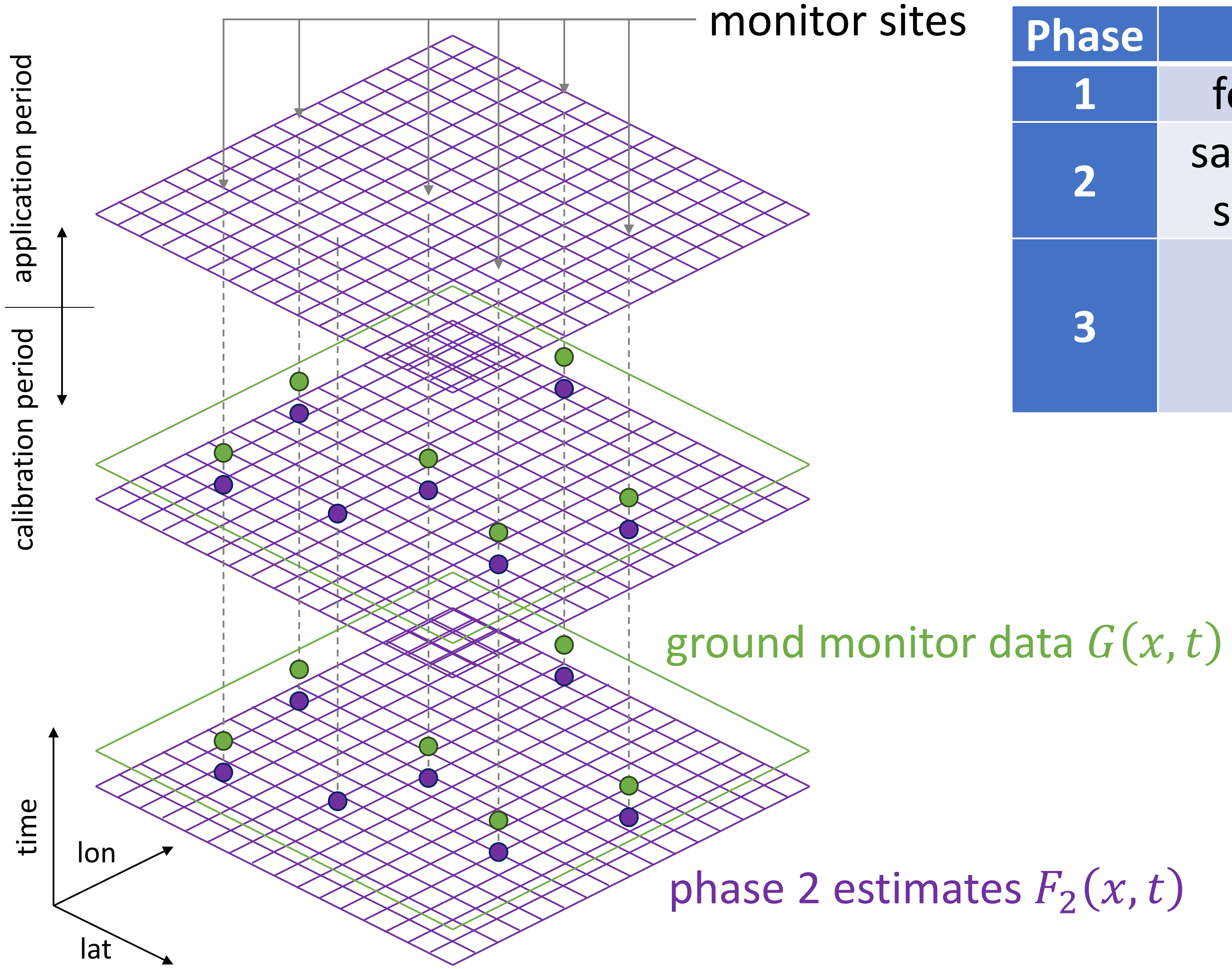
$$V_D(x,t) \approx \mathbb{V}_{t' \in T_{c,overpass}(t)}\left[\left(S_{col}(x,t') - M_{col}(x,t')\right)\phi(x,t')\,\psi(x,t,t')\right]$$

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

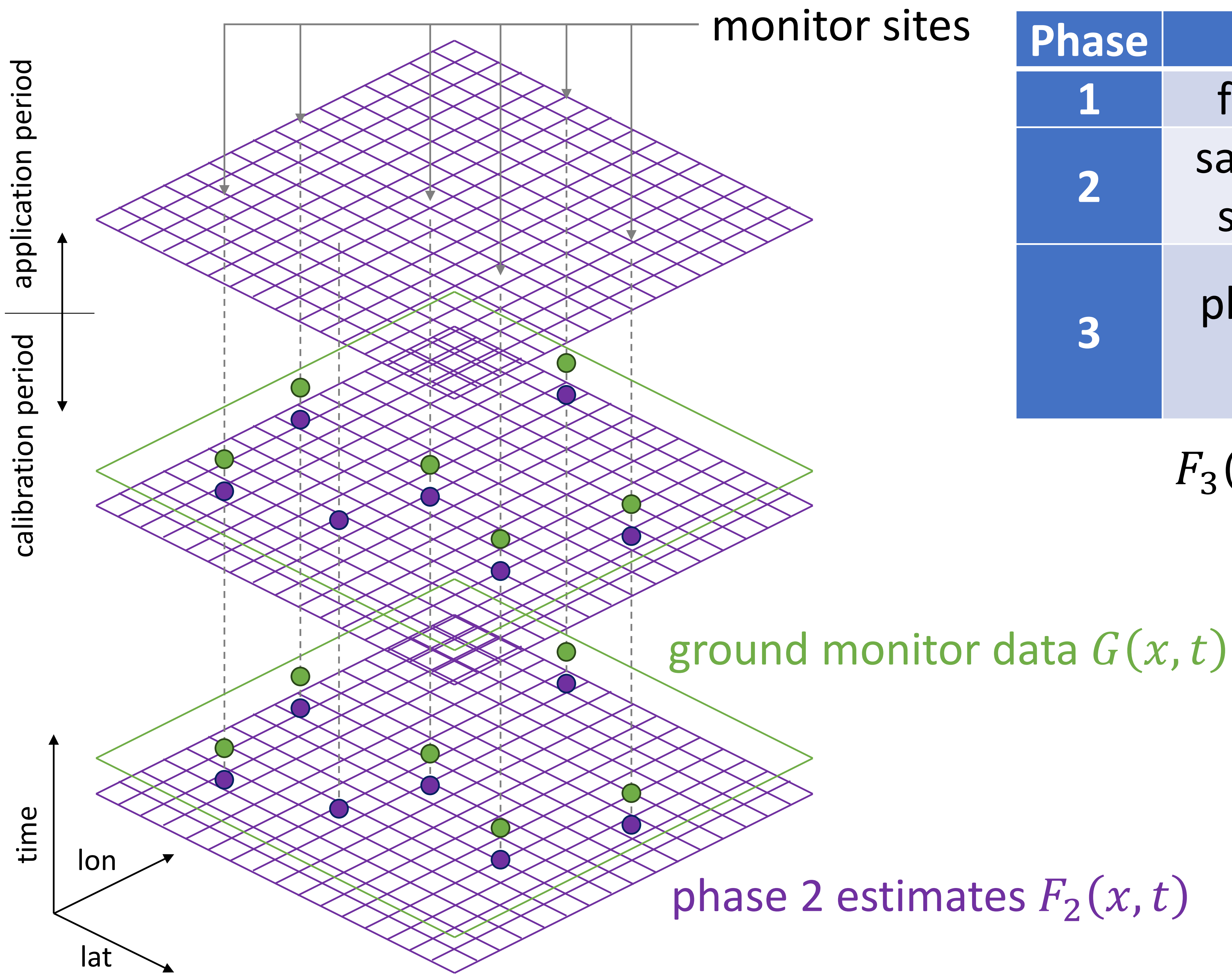


Phase	Estimate	Uncertainty
1	forecast model (GEOS-CF)	cell-to-cell variability of model
2	satellite (TROPOMI) informs sub-model-grid variability	satellite-to-model and surface-to-column ratios vary over time
3		

phase 2 estimates $F_2(x, t)$



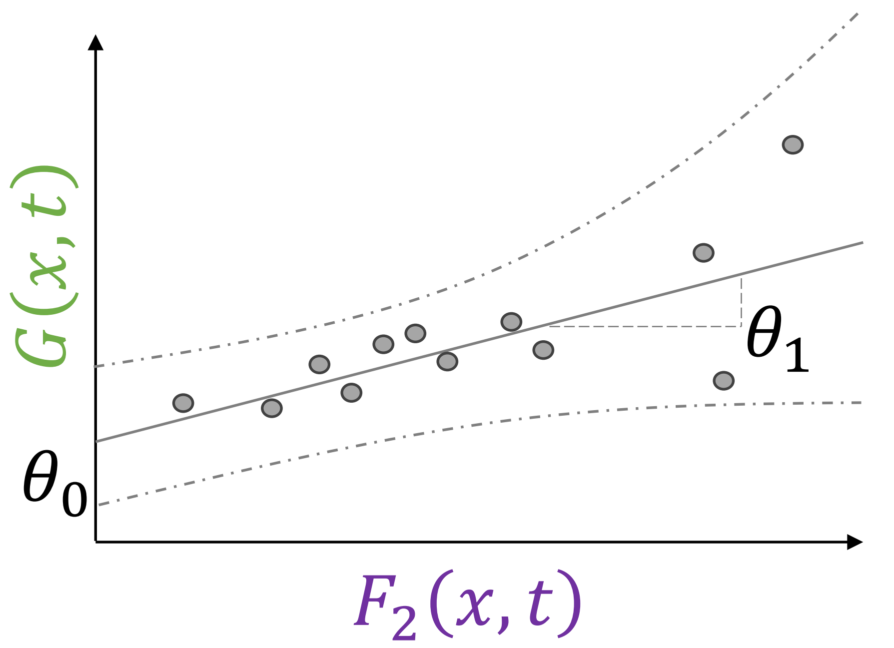
Phase	Estimate	Uncertainty
1	forecast model (GEOS-CF)	cell-to-cell variability of model
2	satellite (TROPOMI) informs sub-model-grid variability	satellite-to-model and surface-to-column ratios vary over time
3		

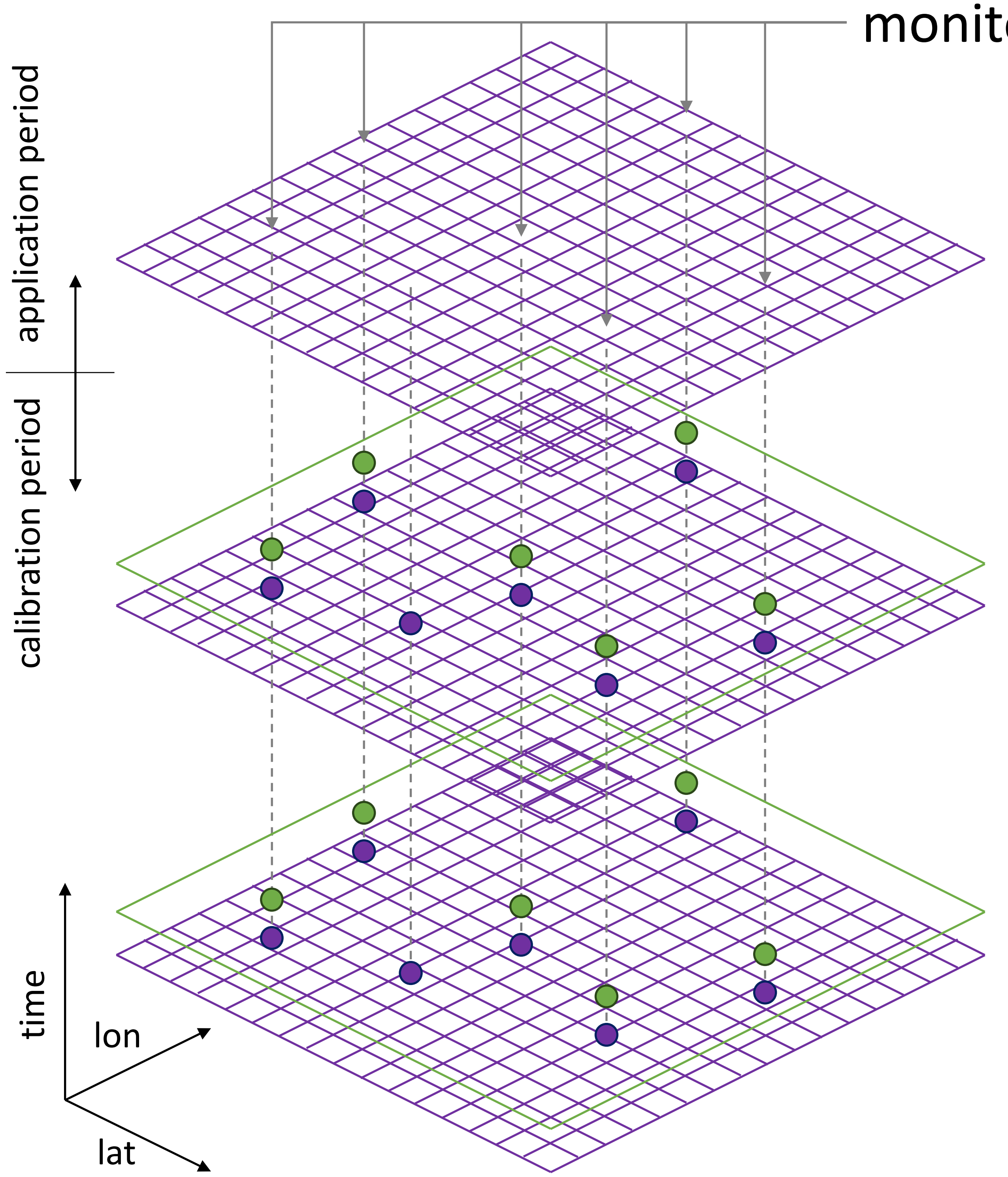


Phase	Estimate	Uncertainty
1	forecast model (GEOS-CF)	cell-to-cell variability of model
2	satellite (TROPOMI) informs sub-model-grid variability	satellite-to-model and surface-to-column ratios vary over time
3	phase 2 corrected to match surface monitor data	

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

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





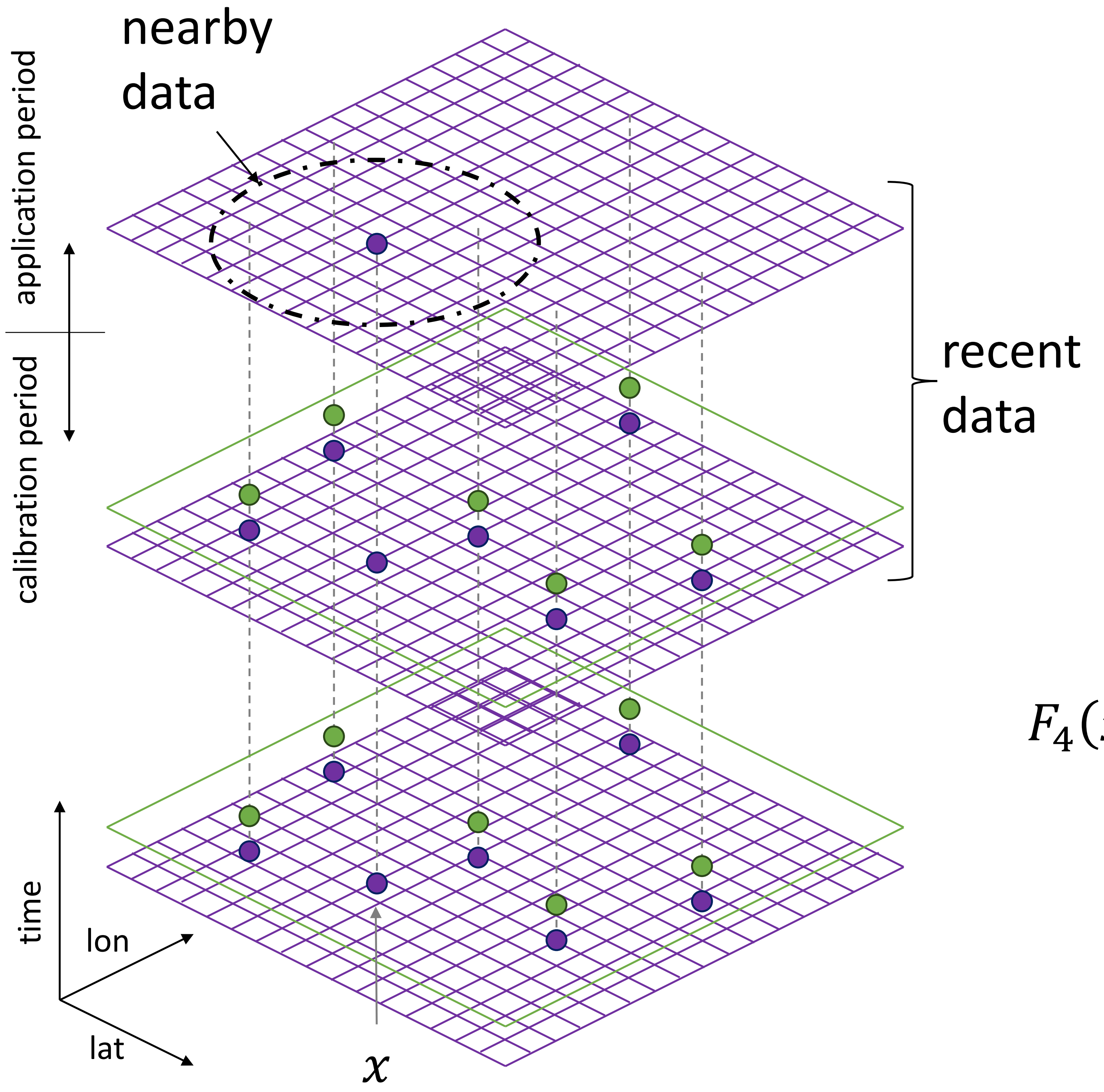
Phase	Estimate	Uncertainty
1	forecast model (GEOS-CF)	cell-to-cell variability of model
2	satellite (TROPOMI) informs sub-model-grid variability	satellite-to-model and surface-to-column ratios vary over time
3	phase 2 corrected to match surface monitor data	uncertain regression parameters between phase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

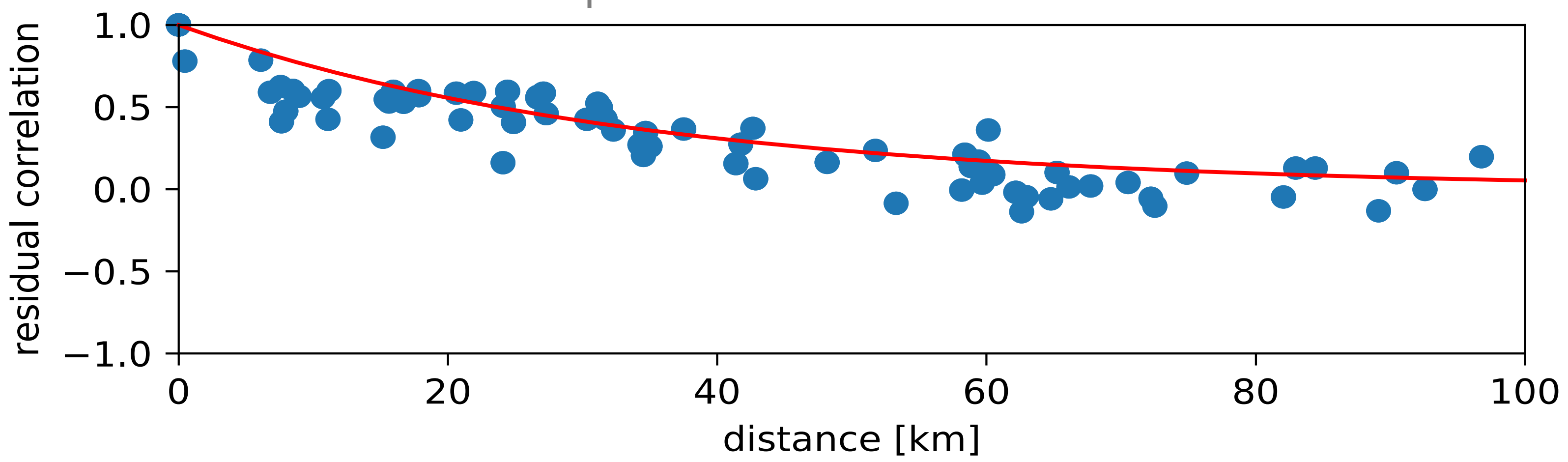
$$\left.\begin{aligned} &+\mathrm{var}[\theta_1]F_2(x,t)^2 \\ &+2\mathrm{cov}[\theta_0,\theta_1]F_2(x,t) \\ &+\mathrm{var}[\theta_0] \\ &+\sigma_{residual}^2 \end{aligned}\right\}$$

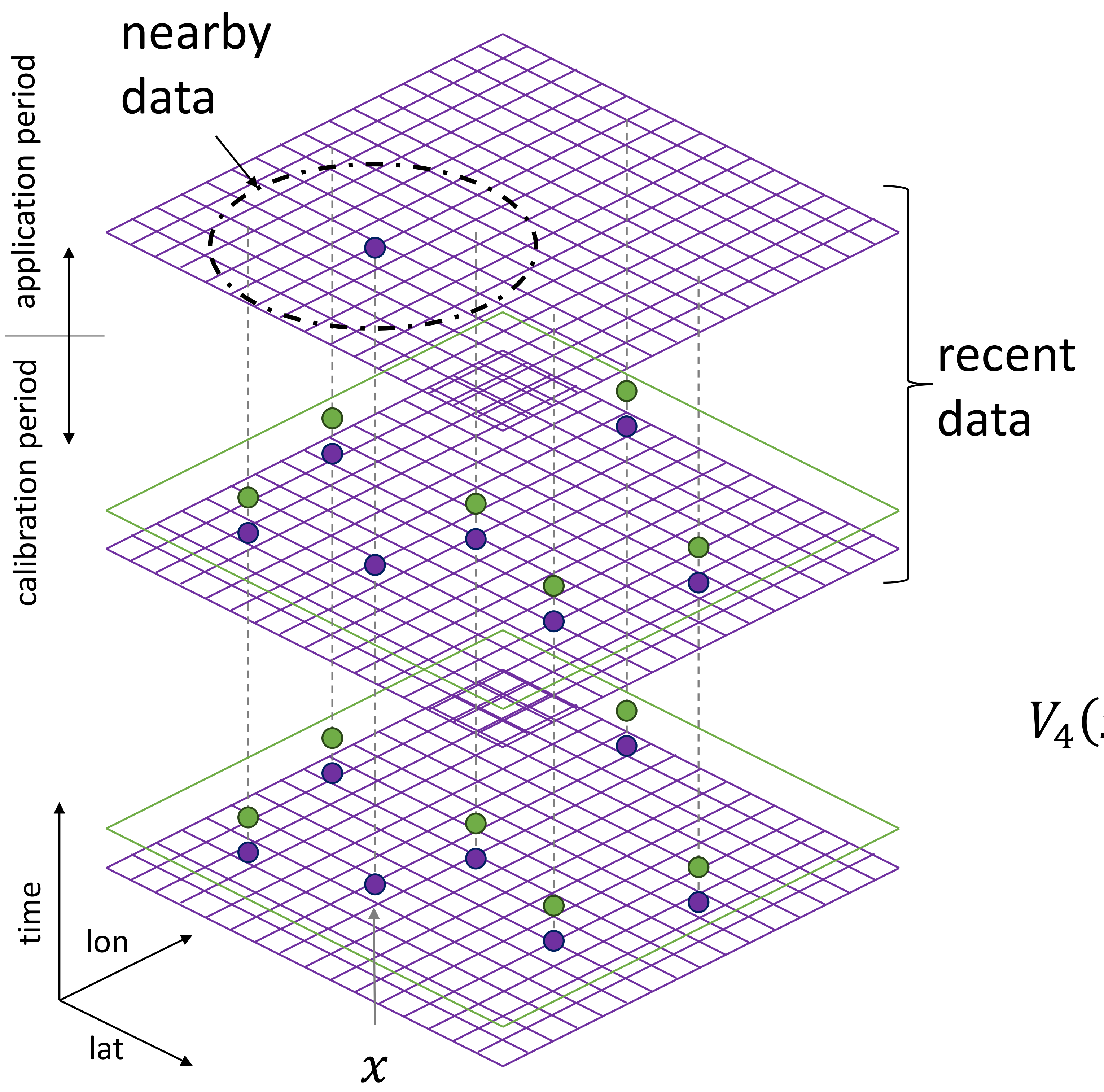
variance and co-variance of regression parameters as well as regression residual are known



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	

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



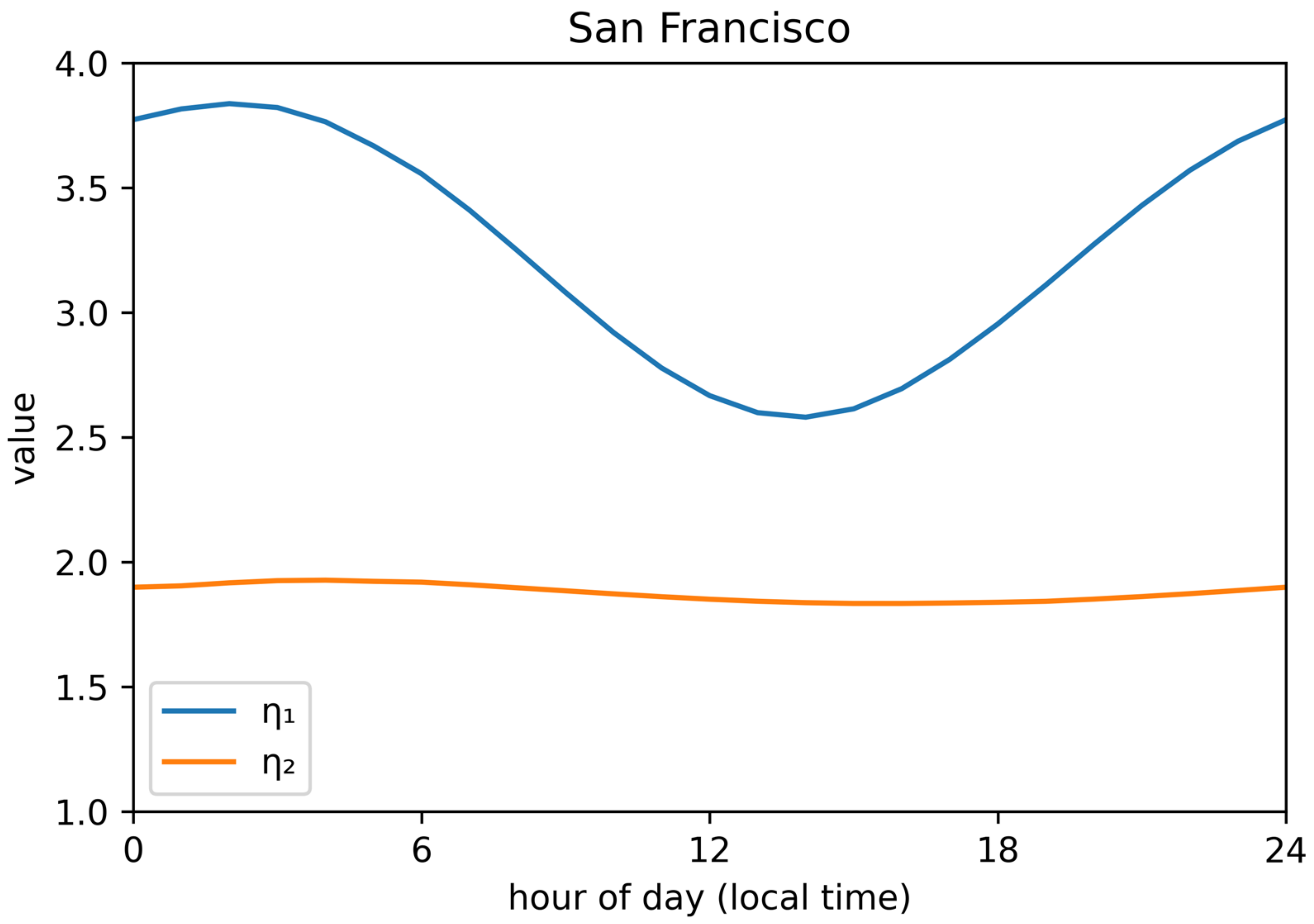


Phase	Estimate	Uncertainty
1	forecast model (GEOS-CF)	cell-to-cell variability of model
2	satellite (TROPOMI) informs sub-model-grid variability	satellite-to-model and surface-to-column ratios vary over time
3	phase 2 corrected to match surface monitor data	uncertain regression parameters between phase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

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

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

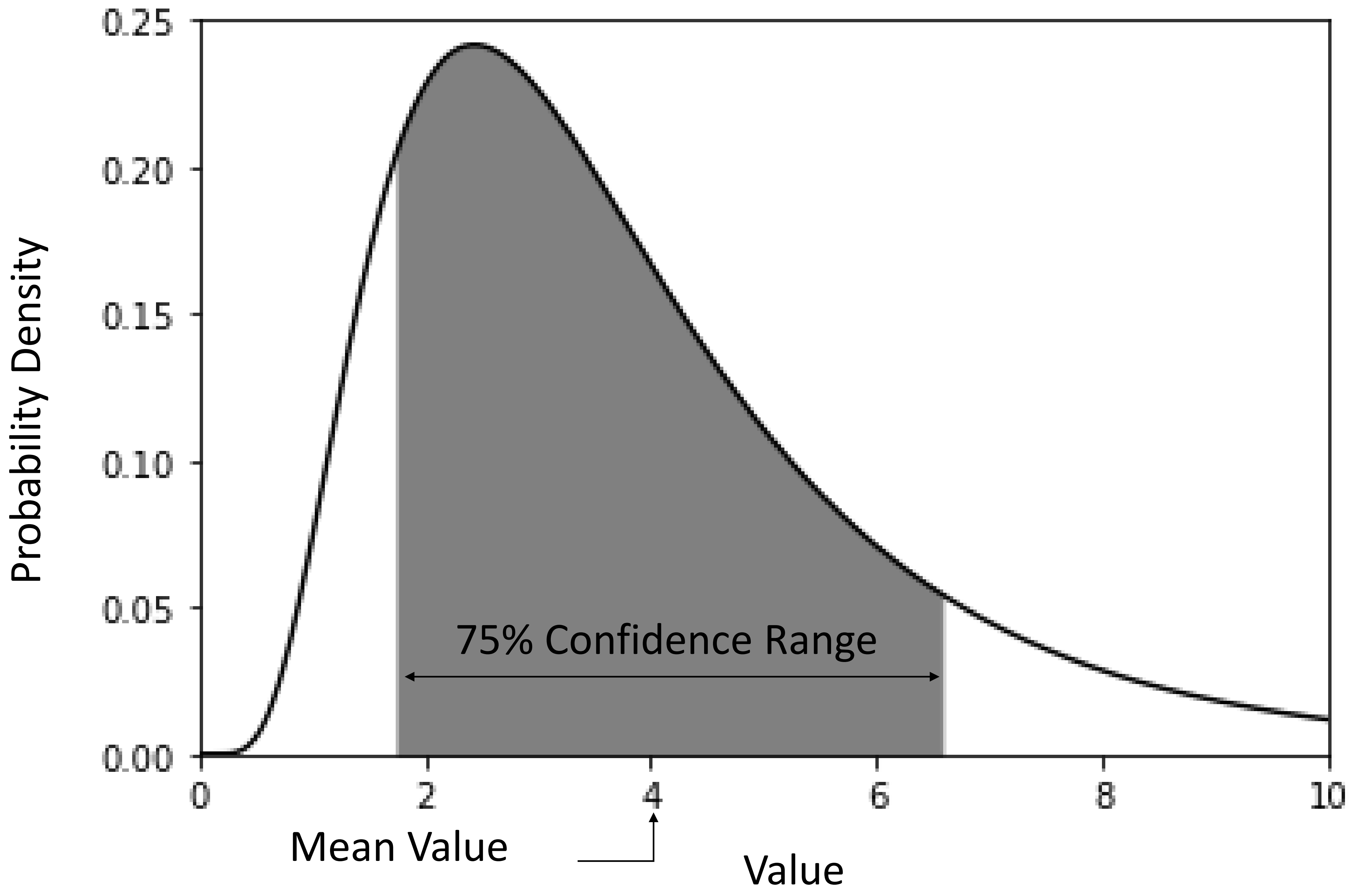
Phase		Estimate	Uncertainty			
			Bias	Model Variability	Model Scale Spatial Representativity	Satellite Scale Spatial Representativity
1	Model	$F_1(x,t) = M(x,t)$	$V_{B1}(x,t)$	$V_M(x,t)$	$V_{R1}(x,t)$	
2	Model & Satellite	$F_2(x,t)$ $= \text{avg}_{t' \in T_c(t)}\left[\left(S_{col}(x,t') - M_{col}(x,t')\right)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



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

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

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



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

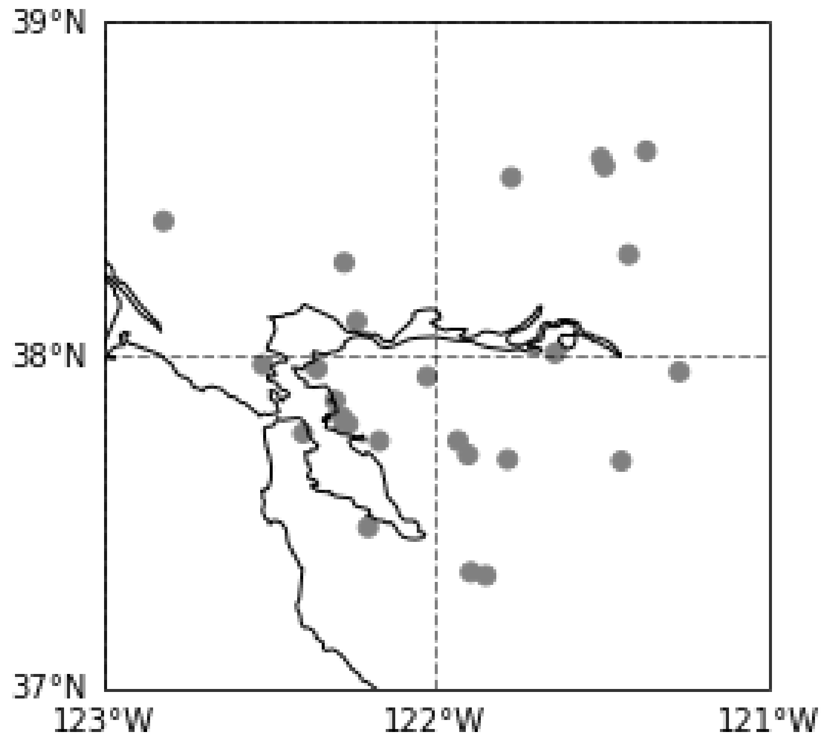
$$\mu(x,t) = \log\left[\frac{F(x,t)}{\sqrt{1+\frac{V(x,t)}{F(x,t)^2}}}\right]$$

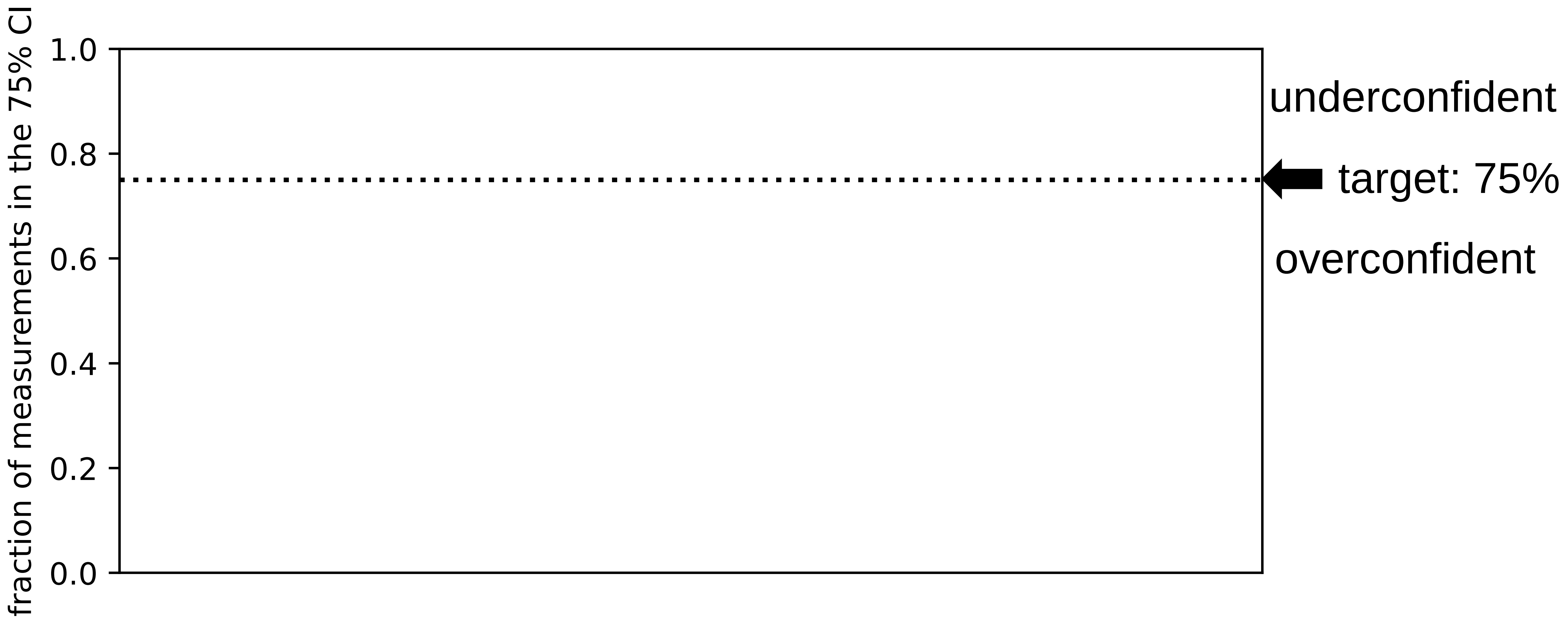
$$\sigma(x,t) = \sqrt{\log\left[1+\frac{V(x,t)}{F(x,t)^2}\right]}$$

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

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

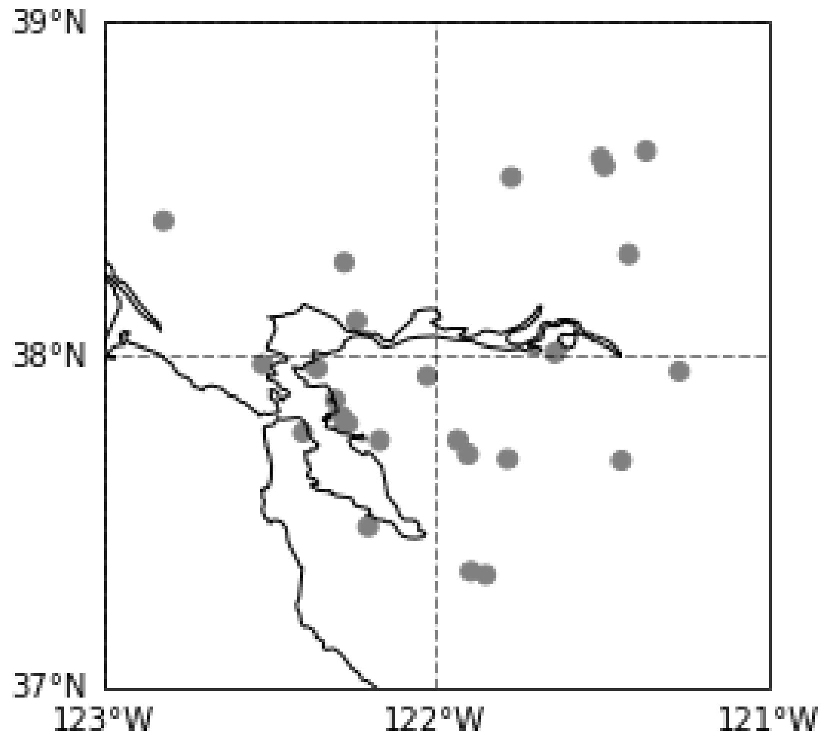
Ground Sites

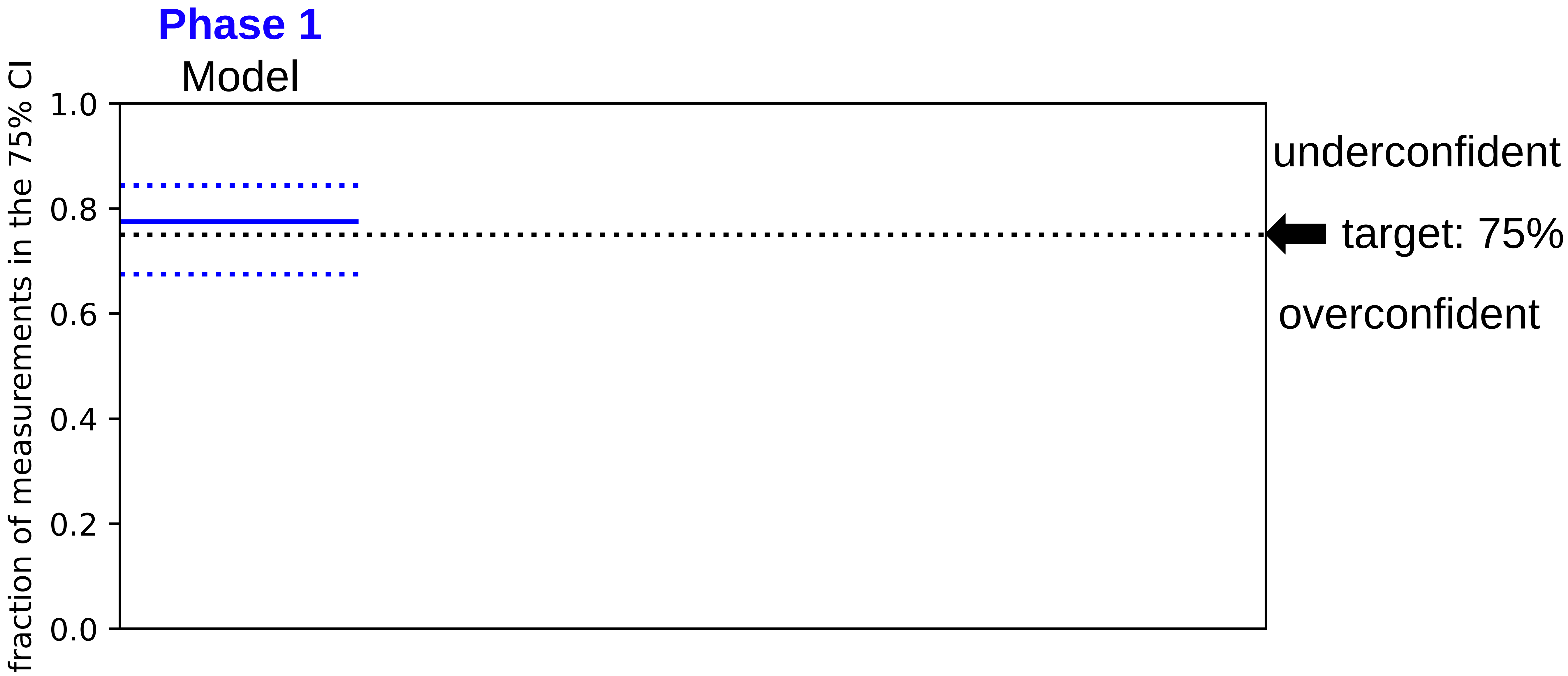




Case Study Details
San Francisco
September 2019
Surface NO_2
Lognormal distribution
Cross-validation test
25 ground monitors

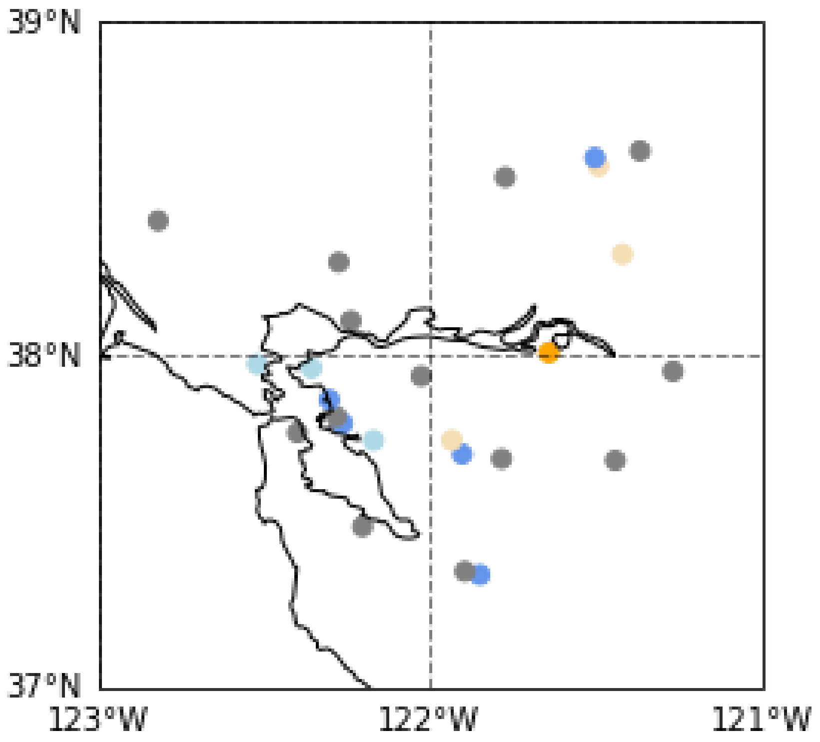
Ground Sites



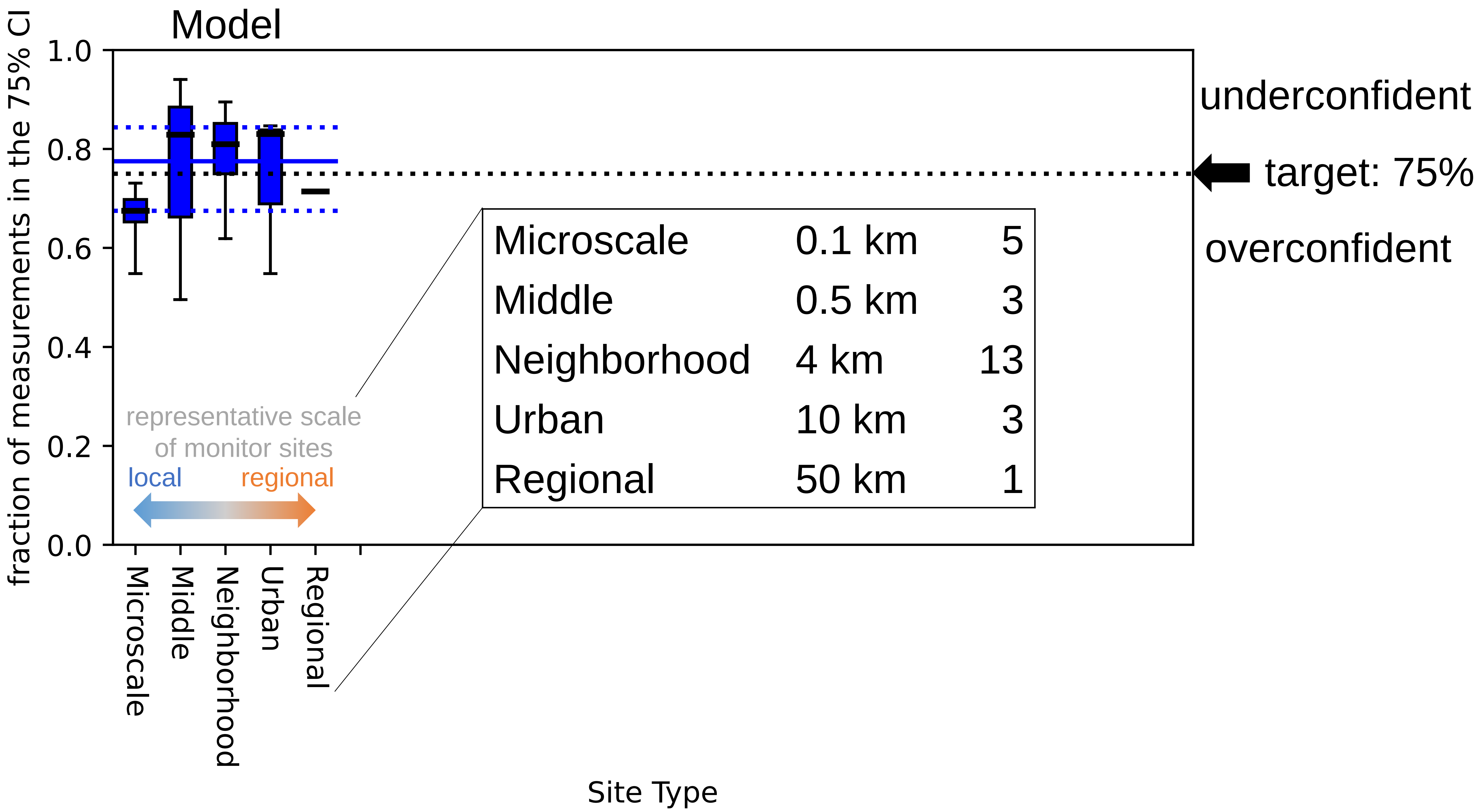


Case Study Details
San Francisco
September 2019
Surface NO_2
Lognormal distribution
Cross-validation test
25 ground monitors

Ground Sites

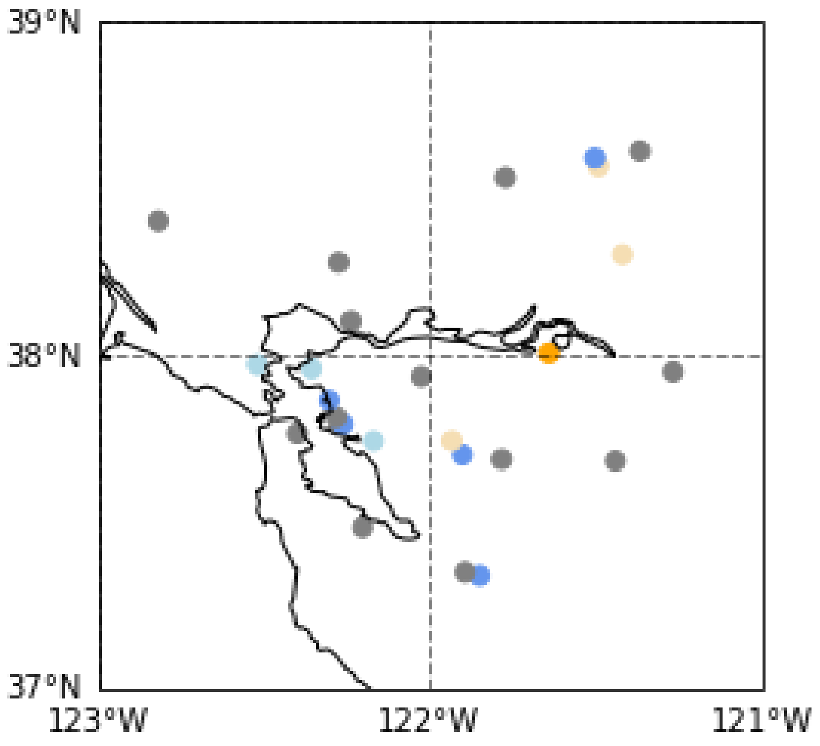


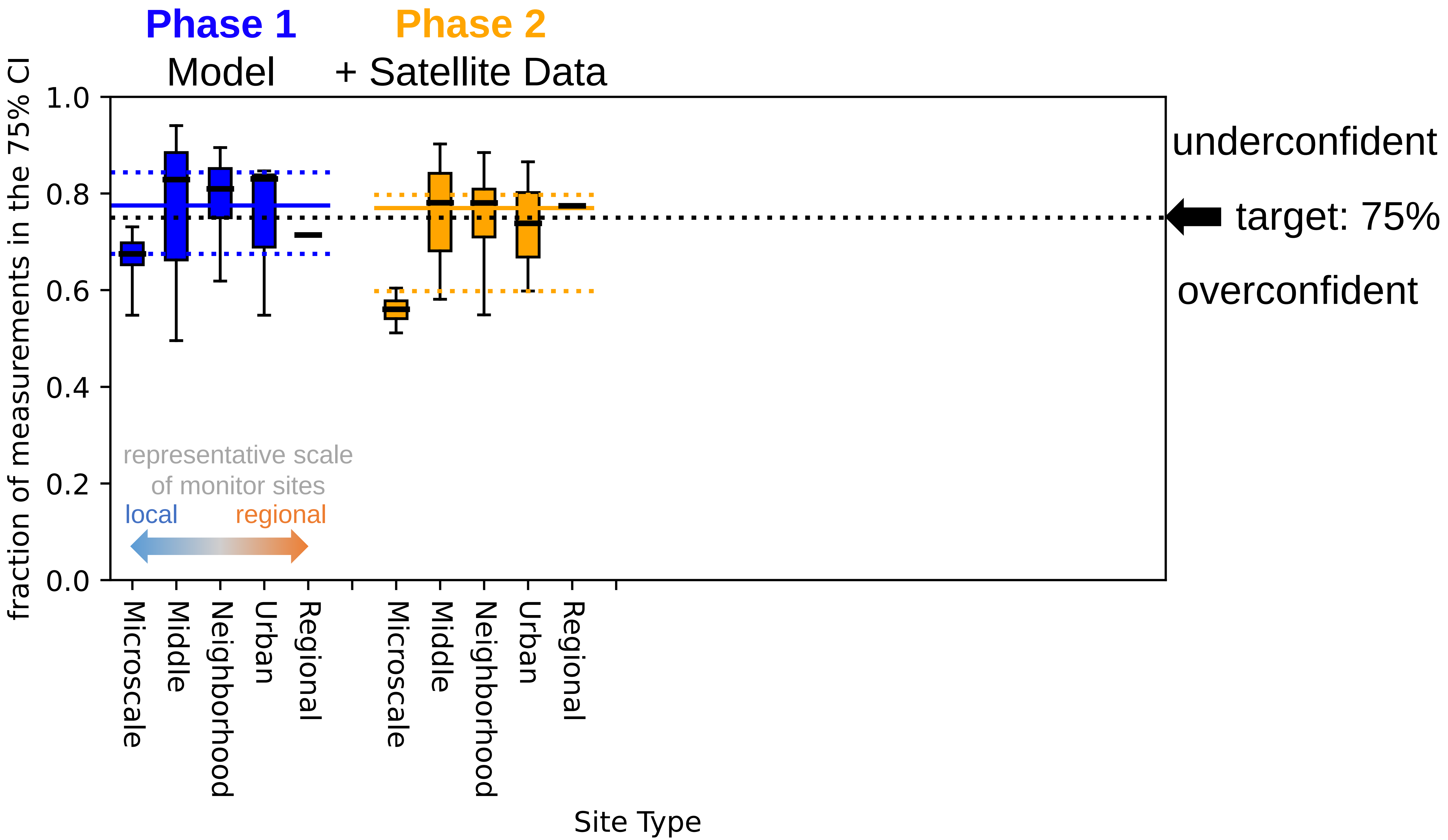
Phase 1



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

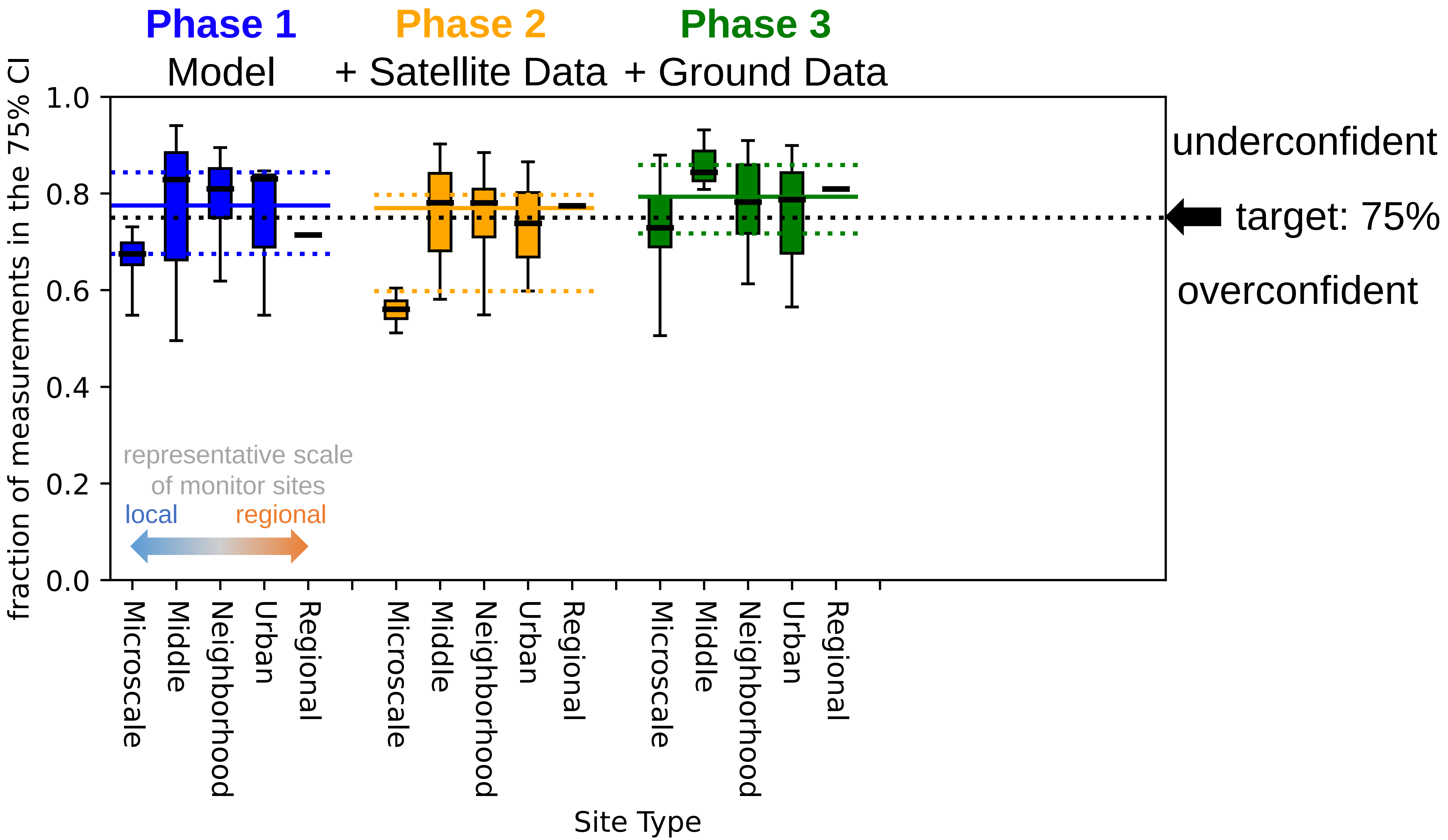
Ground Sites



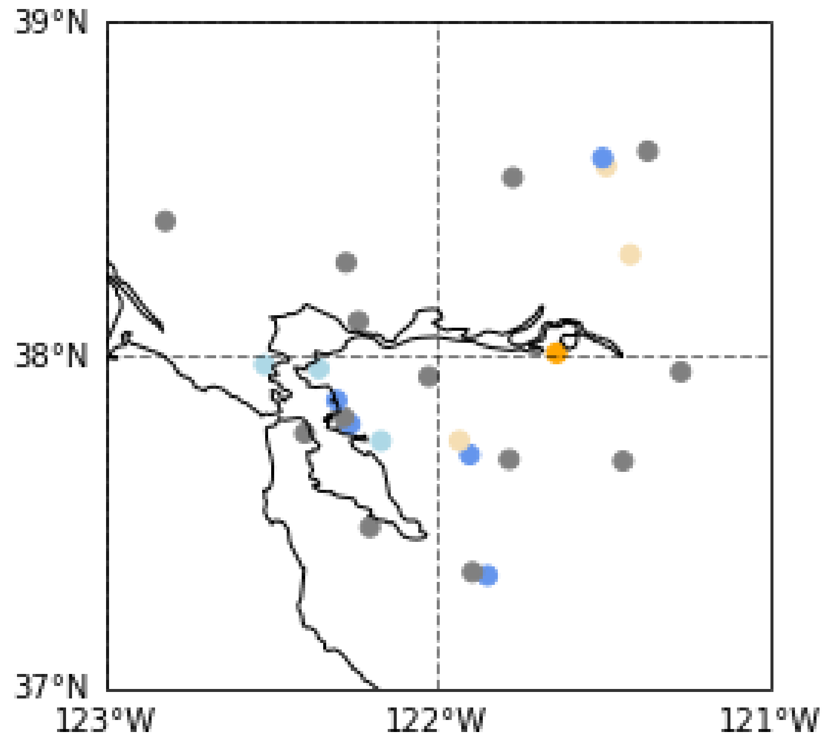


Case Study Details

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



Ground Sites



Case Study Details

San Francisco

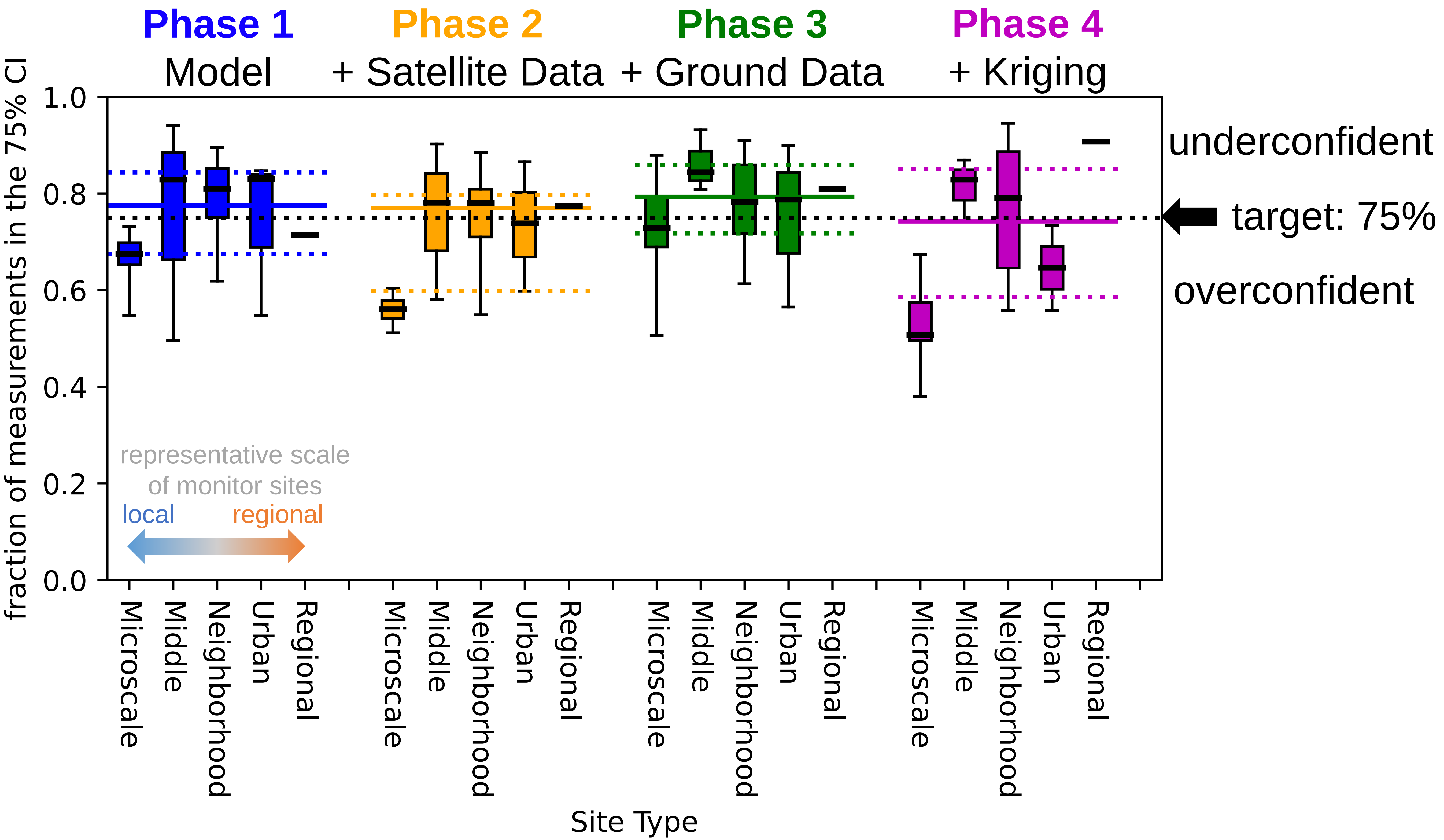
September 2019

Surface NO_2

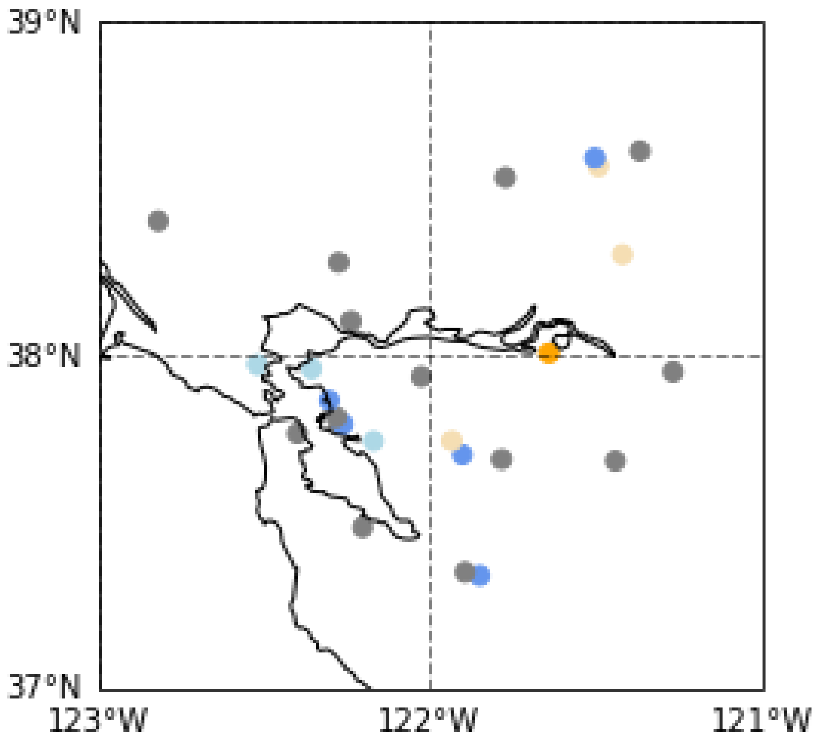
Lognormal distribution

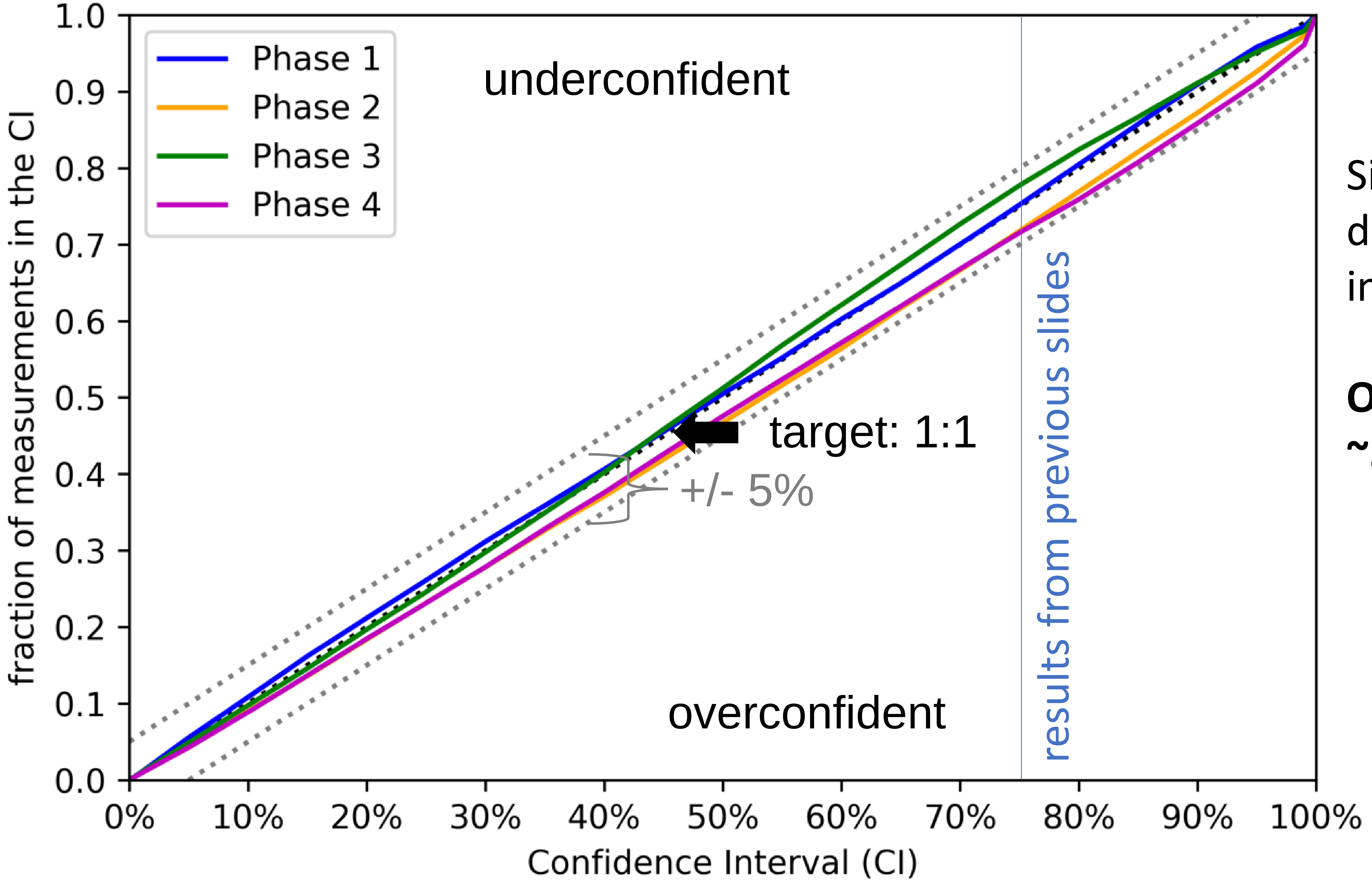
Cross-validation test

25 ground monitors



Ground Sites



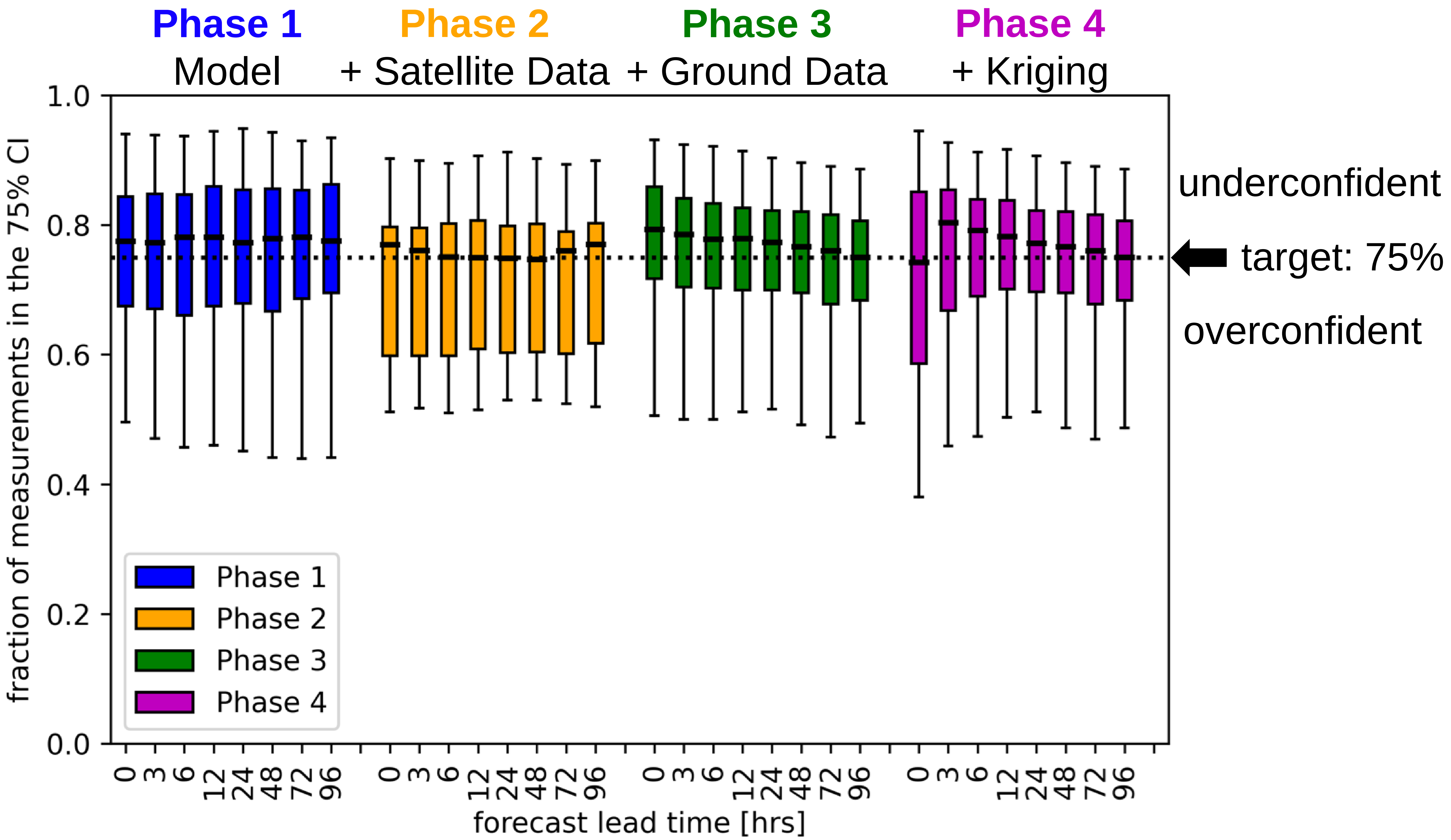


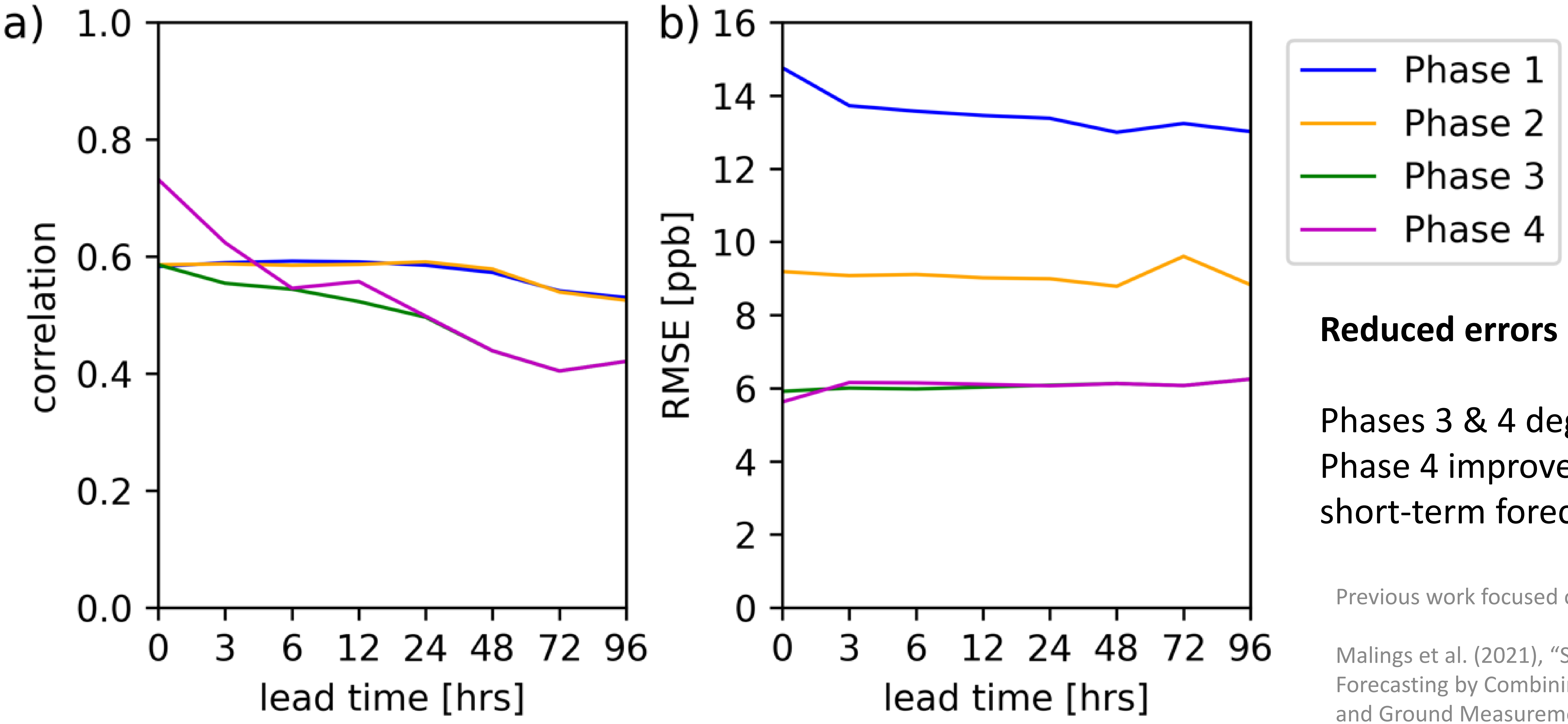
Similar performance for different confidence interval definitions

Overall coverage within ~5% of target

Similar performance for different forecast lead times

Widest spread in phase 4 coverage at 0 lead time





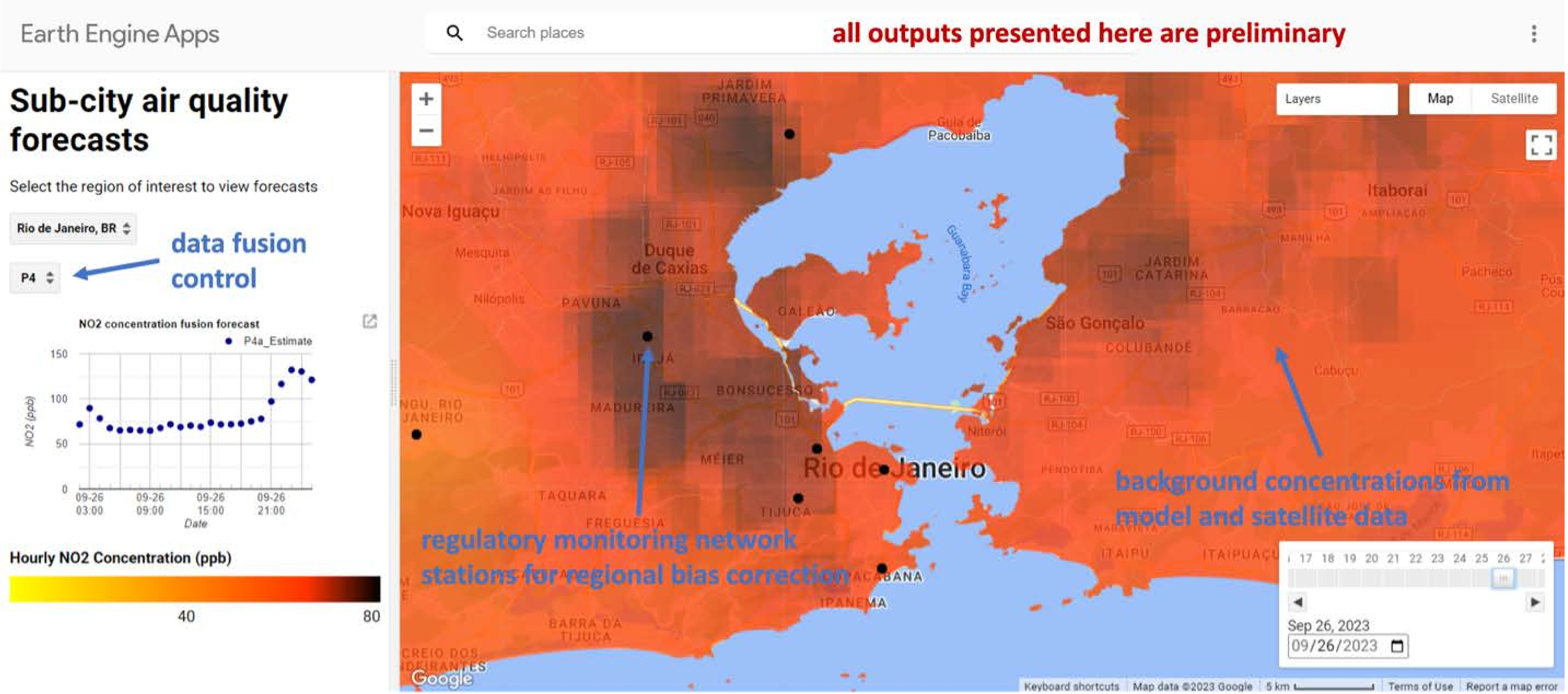
Reduced errors as phase increases

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

- 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?



Source: NASA GMAO Science Snapshot [“Google Earth Engine Data Fusion Tool to support Air Quality Managers”]

Global Modeling and Assimilation Office
GESTAR II Cooperative Agreement

Partner

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.