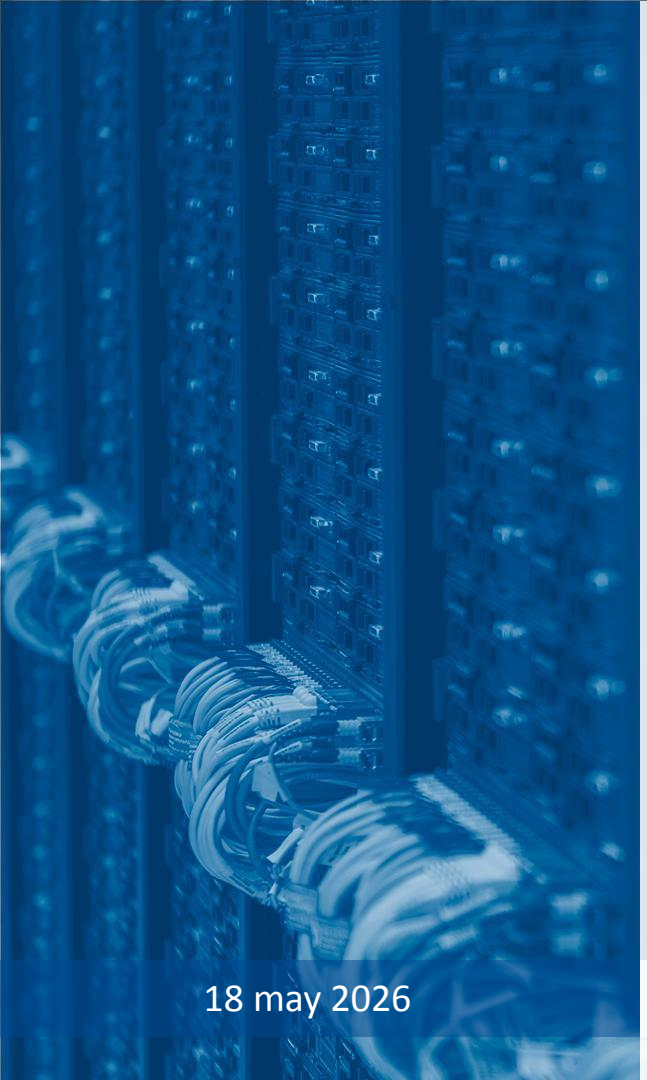


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

FAIRMODE-WG5: Contribution to the bias projection exercise - Phase 2

Ada Barrantes, Cristina Campos, Cristina Carnerero and Jan Mateu

18 may 2026

Air Quality Team, Earth Sciences Department from BSC

Methodology

- Bias correction trained with Base Case 2015 applied to future Scenarios (2022, 2023, and 2024) for NO_2, O_3, $PM_{2.5}$ using:
  - **Random Forest** (RF)
  - **Light Gradient Boosting Machine** (LGB)
- Considering only **background stations** and Urban, Suburban and Rural
- Corrections: **with and without including** the regridded Phase 1 **topography** as a predictor
- **Validated using the training-test** (Active-inActive) station distribution proposed by **NILU**
- Preprocessing:
  - **Cleaning the dataset:** removing duplicates + outside European domain observations
  - **Bilinear Interpolation** of **modelled values** at the monitoring location

Methodology

- Bias correction trained with Base Case 2015 applied to future Scenarios (2022, 2023, and 2024) for NO_2, O_3, $PM_{2.5}$ using:
  - **Random Forest** (RF)
  - **Light Gradient Boosting Machine** (LGB)
- Considering only **background stations** and Urban, Suburban and Rural
- Corrections: **with and without including** the regridded Phase 1 **topography** as a predictor
- **Validated using the training-test** (Active-inActive) station distribution proposed by **NILU**
- Preprocessing:
  - **Cleaning the dataset:** removing duplicates + outside European domain observations
  - **Bilinear Interpolation** of **modelled values** at the monitoring location

Methodology

- Bias correction trained with Base Case 2015 applied to future Scenarios (2022, 2023, and 2024) for NO_2, O_3, $PM_{2.5}$ using:
  - **Random Forest** (RF)
  - **Light Gradient Boosting Machine** (LGB)
- Considering only **background stations** and Urban, Suburban and Rural
- Corrections: **with and without including** the regridded Phase 1 **topography** as a predictor
- **Validated using the training-test** (Active-inActive) station distribution proposed by **NILU**
- Preprocessing:
  - **Cleaning the dataset:** removing duplicates + outside European domain observations
  - **Bilinear Interpolation** of **modelled values** at the monitoring location

Methodology

- Bias correction trained with Base Case 2015 applied to future Scenarios (2022, 2023, and 2024) for NO_2, O_3, $PM_{2.5}$ using:
  - **Random Forest** (RF)
  - **Light Gradient Boosting Machine** (LGB)
- Considering only **background stations** and Urban, Suburban and Rural
- Corrections: **with and without including** the regridded Phase 1 **topography** as a predictor
- **Validated using the training-test** (Active-inActive) station distribution proposed by **NILU**
- Preprocessing:
  - **Cleaning the dataset:** removing duplicates + outside European domain observations
  - **Bilinear Interpolation** of **modelled values** at the monitoring location

Methodology

- Bias correction trained with Base Case 2015 applied to future Scenarios (2022, 2023, and 2024) for NO_2, O_3, $PM_{2.5}$ using:
  - **Random Forest** (RF)
  - **Light Gradient Boosting Machine** (LGB)
- Considering only **background stations** and Urban, Suburban and Rural
- Corrections: **with and without including** the regridded Phase 1 **topography** as a predictor
- **Validated using the training-test** (Active-inActive) station distribution proposed by **NILU**
- Preprocessing:
  - **Cleaning the dataset:** removing duplicates + outside European domain observations
  - **Bilinear Interpolation** of **modelled values** at the monitoring location

Machine Learning algorithms

Target variable: logarithmic **ratio** between the observed and modelled values

$$ln\frac{\text{obs}}{\text{model}} \sim \text{model (+ topography)} + sin\frac{x}{e_1} + cos\frac{x}{e_1} + sin\frac{y}{e_1} + cos\frac{y}{e_1} + sin\frac{x}{e_2} + cos\frac{x}{e_2} + sin\frac{y}{e_2} + cos\frac{y}{e_2}$$

Predictors:

- The interpolated **modelled values** at the station location
- The **topography** at the station location (trained **with and without** it)
- Its **location**: encoded mathematically using sin/cos waves to capture geographic patterns (Mai, 2020)
  - X and Y are the **coordinates in meters** (UTM 3035)
  - $\mathbf{e_1}$ **:** regional scale behaviour (150-400 km)
  - $\mathbf{e_2}$ **:** synoptic scale (2000-3000 km)

Bias-correction future **Scenarios** (at each grid cell):

Machine Learning algorithms

Target variable: logarithmic **ratio** between the observed and modelled values

$$ln\frac{\text{obs}}{\text{model}} \sim \text{model}\ (+\ \text{topography}) + sin\frac{x}{e_1} + cos\frac{x}{e_1} + sin\frac{y}{e_1} + cos\frac{y}{e_1} + sin\frac{x}{e_2} + cos\frac{x}{e_2} + sin\frac{y}{e_2} + cos\frac{y}{e_2}$$

Predictors:

- The interpolated **modelled values** at the station location
- The **topography** at the station location (trained **with and without** it)
- Its **location**: encoded mathematically using sin/cos waves to capture geographic patterns (Mai, 2020)
- X and Y are the **coordinates in meters** (UTM 3035)
  - $\mathbf{e_1}$ **:** regional scale behaviour (150-400 km)
  - $\mathbf{e_2}$ **:** synoptic scale (2000-3000 km)

Bias-correction future **Scenarios** (at each grid cell):

Machine Learning algorithms

Target variable: logarithmic **ratio** between the observed and modelled values

$$ln\frac{\text{obs}}{\text{model}} \sim \text{model (+ topography)} + sin\frac{x}{e_1} + cos\frac{x}{e_1} + sin\frac{y}{e_1} + cos\frac{y}{e_1} + sin\frac{x}{e_2} + cos\frac{x}{e_2} + sin\frac{y}{e_2} + cos\frac{y}{e_2}$$

Predictors:

- The interpolated **modelled values** at the station location
- The **topography** at the station location (trained **with and without** it)
- Its **location**: encoded mathematically using sin/cos waves to capture geographic patterns (Mai, 2020)
  - X and Y are the **coordinates in meters** (UTM 3035)
  - $\mathbf{e_1}$: regional scale behaviour (150-400 km)
  - $\mathbf{e_2}$: synoptic scale (2000-3000 km)

Bias-correction future **Scenarios** (at each grid cell):

Mai, 2020 arXiv:2003.00824

Machine Learning algorithms

Target variable: logarithmic **ratio** between the observed and modelled values

$$ln\frac{\text{obs}}{\text{model}} \sim \text{model (+ topography)} + sin\frac{x}{e_1} + cos\frac{x}{e_1} + sin\frac{y}{e_1} + cos\frac{y}{e_1} + sin\frac{x}{e_2} + cos\frac{x}{e_2} + sin\frac{y}{e_2} + cos\frac{y}{e_2}$$

Predictors:

- The interpolated **modelled values** at the station location
- The **topography** at the station location (trained **with and without** it)
- Its **location**: encoded mathematically using sin/cos waves to capture geographic patterns (Mai, 2020)
  - X and Y are the **coordinates in meters** (UTM 3035)
  - $\mathbf{e_1}$ **:** regional scale behaviour (150-400 km)
  - $\mathbf{e_2}$ **:** synoptic scale (2000-3000 km)

Bias-correction future **Scenarios** (at each grid cell):

$$\text{bias\_corrected\_model}_i = \text{model}_i \cdot exp\left(ln\frac{\text{obs}}{\text{model}}\right)$$

Results

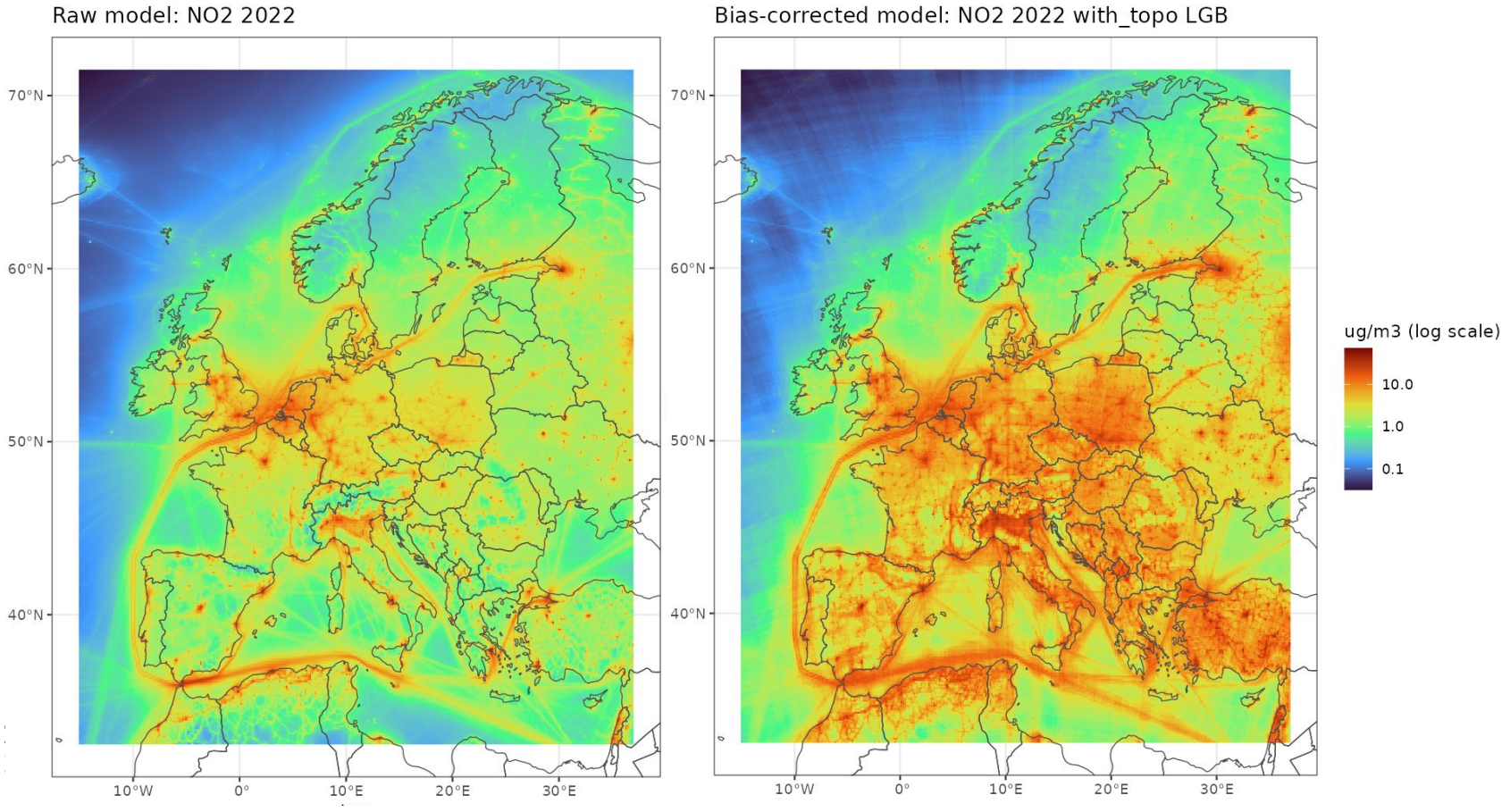

BSC
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

Statistical Performance Training 2015

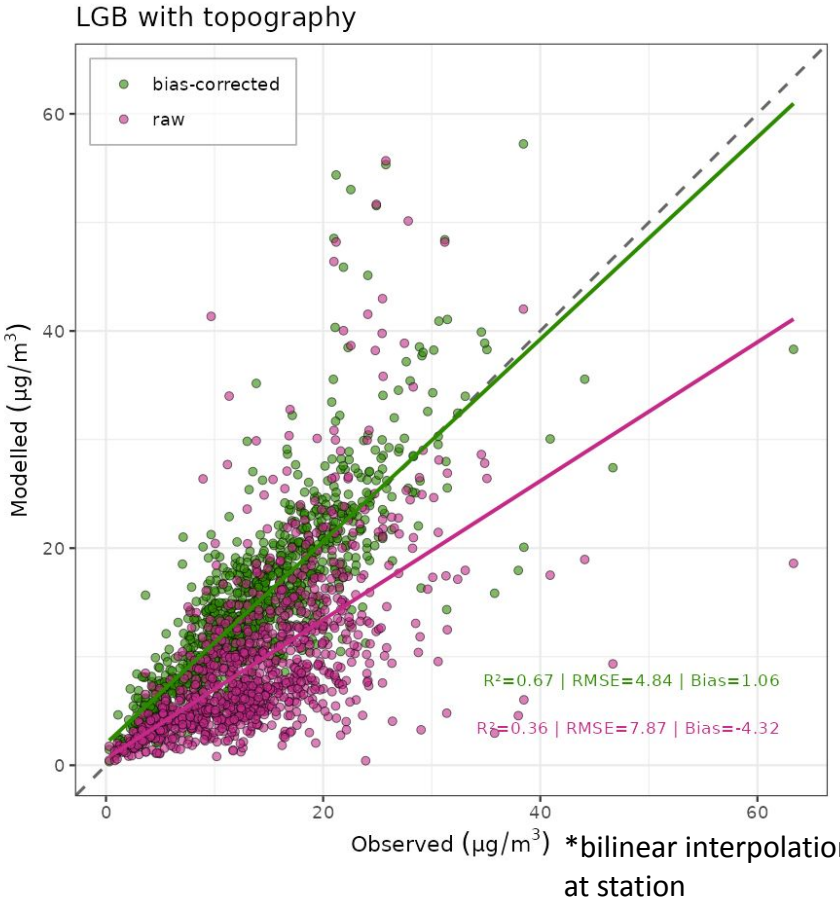
Table 1. Root mean square error (RMSE, µg m^{-3}) and Pearson correlation coefficient (R) between observed concentrations and the values estimated by the machine learning bias-correction models (Random Forest, RF, and LightGBM, LGB) at monitoring stations. Results are shown for the original model output (before bias correction) and for the corrected estimates over the training (Active stations) and test (inActive stations) datasets defined by the NILU classification.

		Original	Random Forest		light Gradient Boosting Machine	
		model vs. obs	Training	Test	Training	Test
NO_2	RMSE (µg m^{-3})	9.8	5.9	6.1	6.1	6.1
	Correlation (R)	0.7	0.8	0.8	0.8	0.8
O_3	RMSE (µg m^{-3})	14.2	8.5	9.2	8.9	9.5
	Correlation (R)	0.6	0.7	0.6	0.6	0.6
$PM_{2.5}$	RMSE (µg m^{-3})	5.3	3.5	2.8	3.5	2.9
	Correlation (R)	0.7	0.9	0.9	0.9	0.9

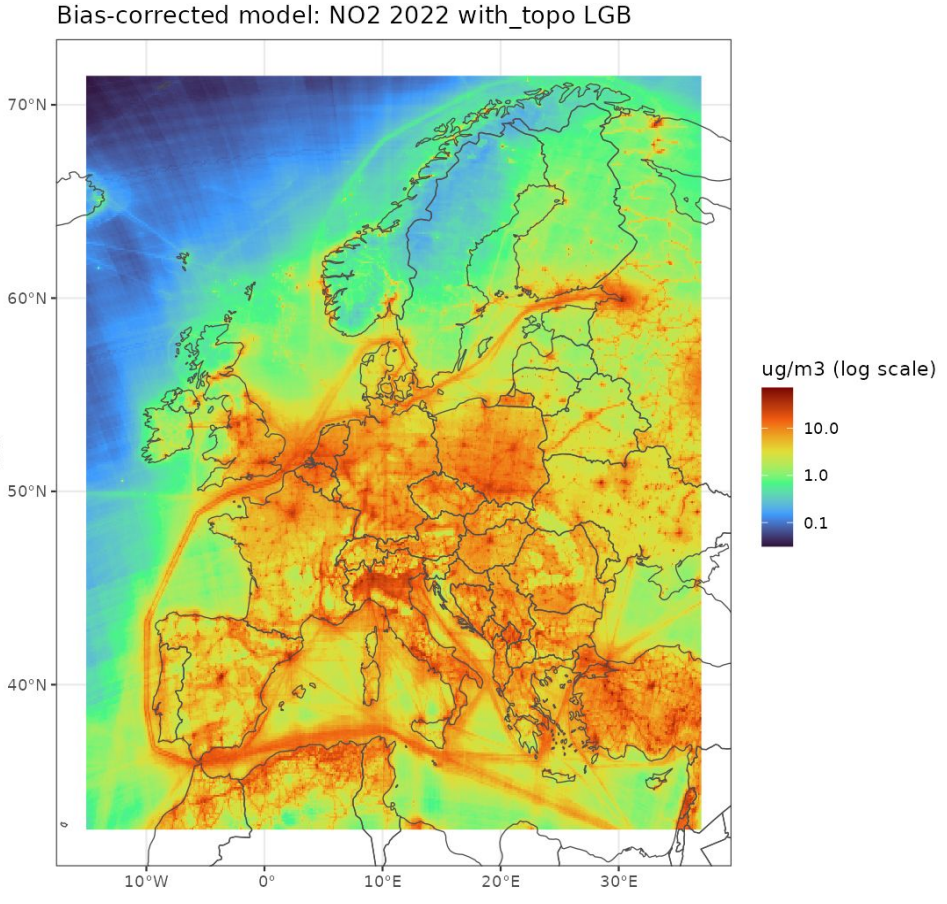
NO_2 Scenario 2022



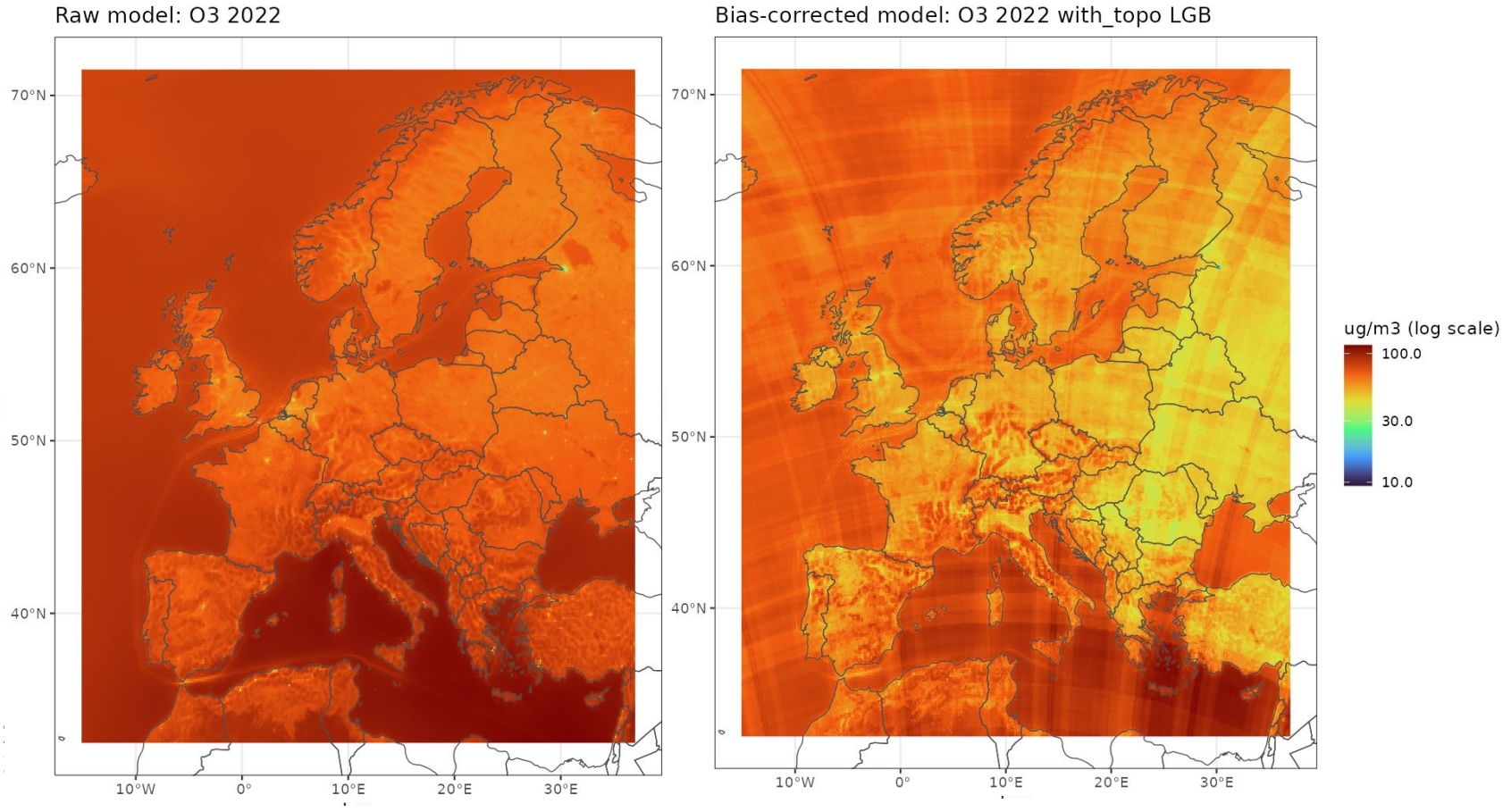
NO_2 Scenario 2022



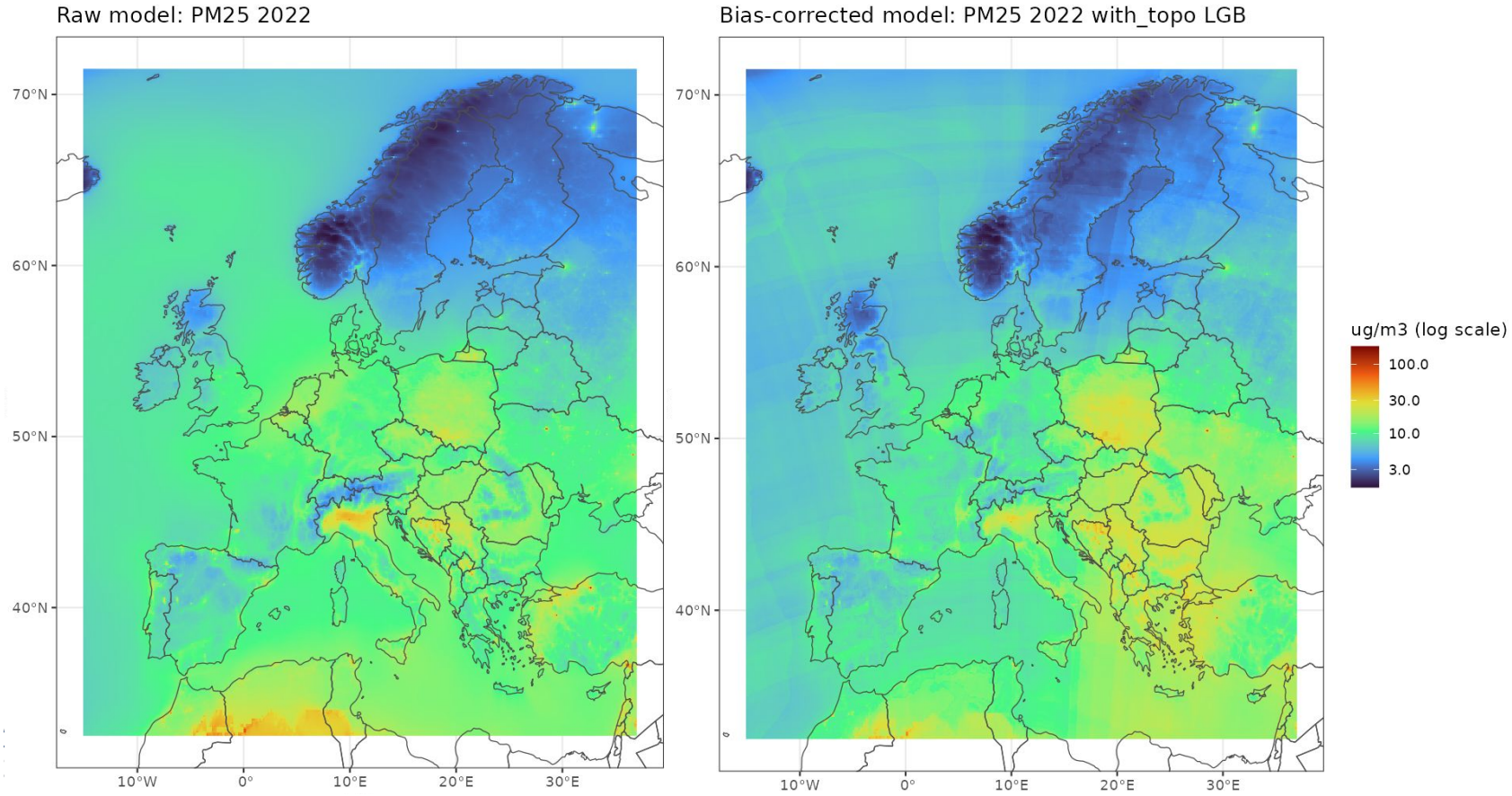
*bilinear interpolation at station



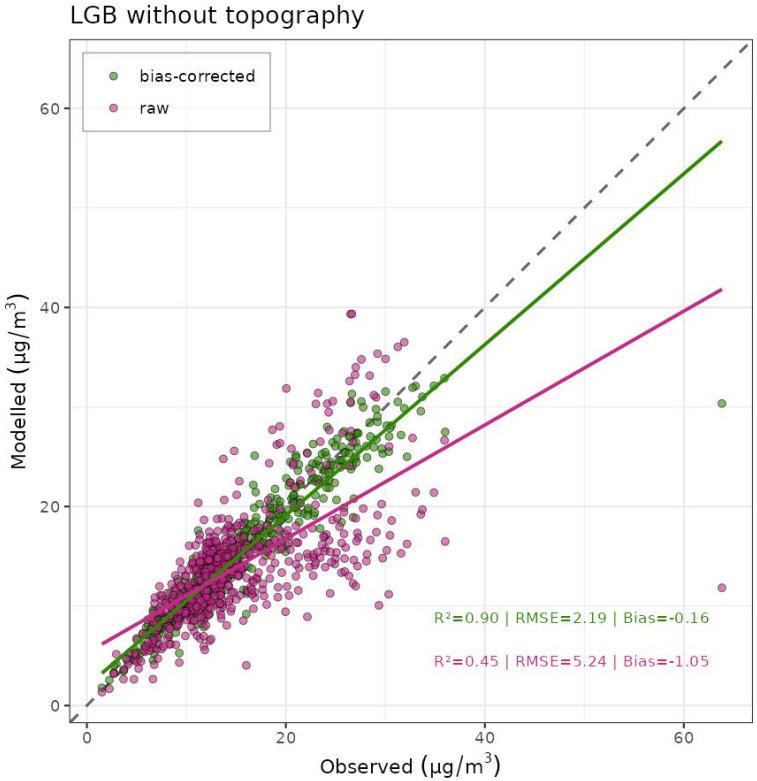
O_3 Scenario 2022

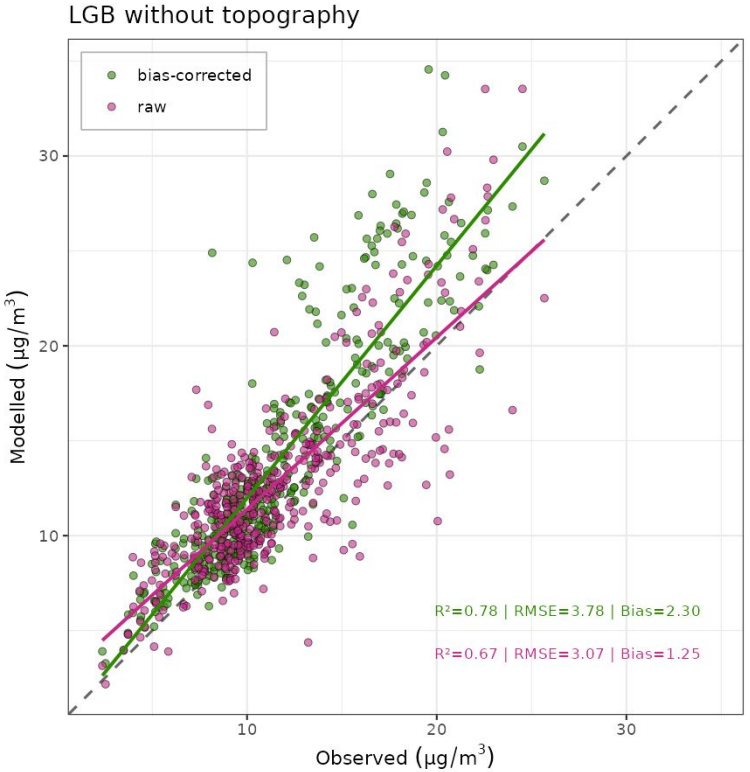


$PM_{2.5}$ Scenario 2022



$PM_{2.5}$ Base Case 2015 | Scenario 2022





*bilinear interpolation at station

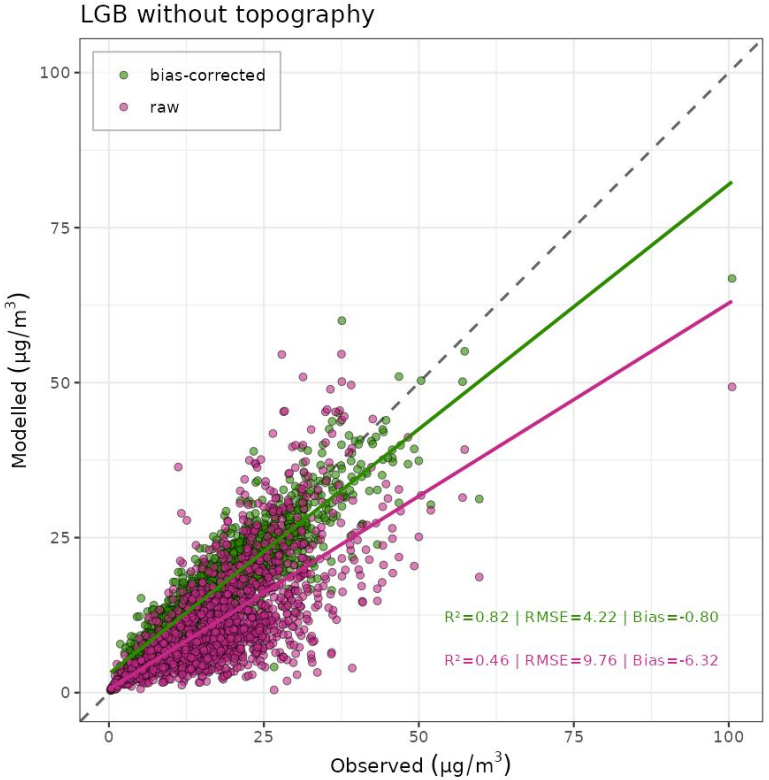
Thank you!

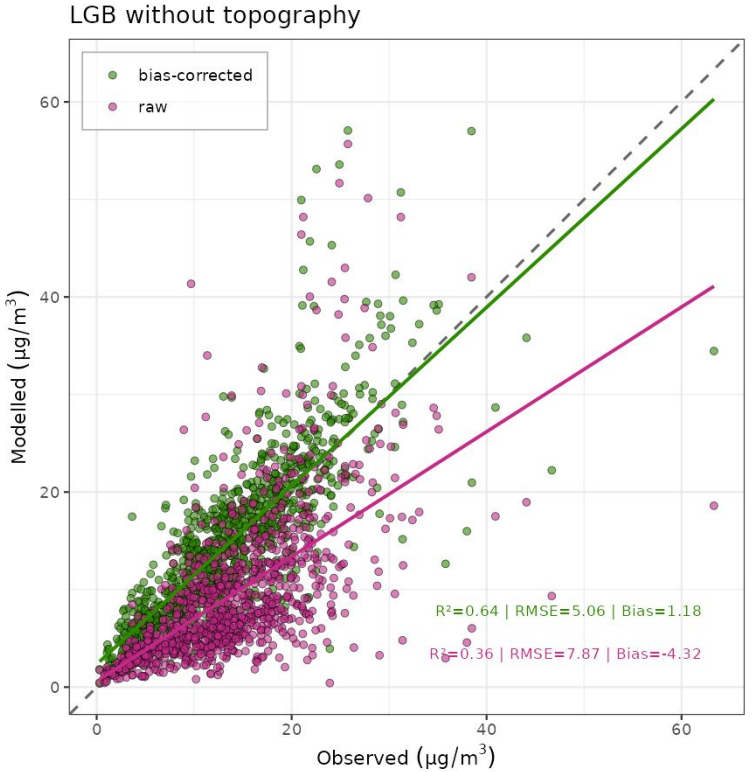

BSC
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

Extra slides

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

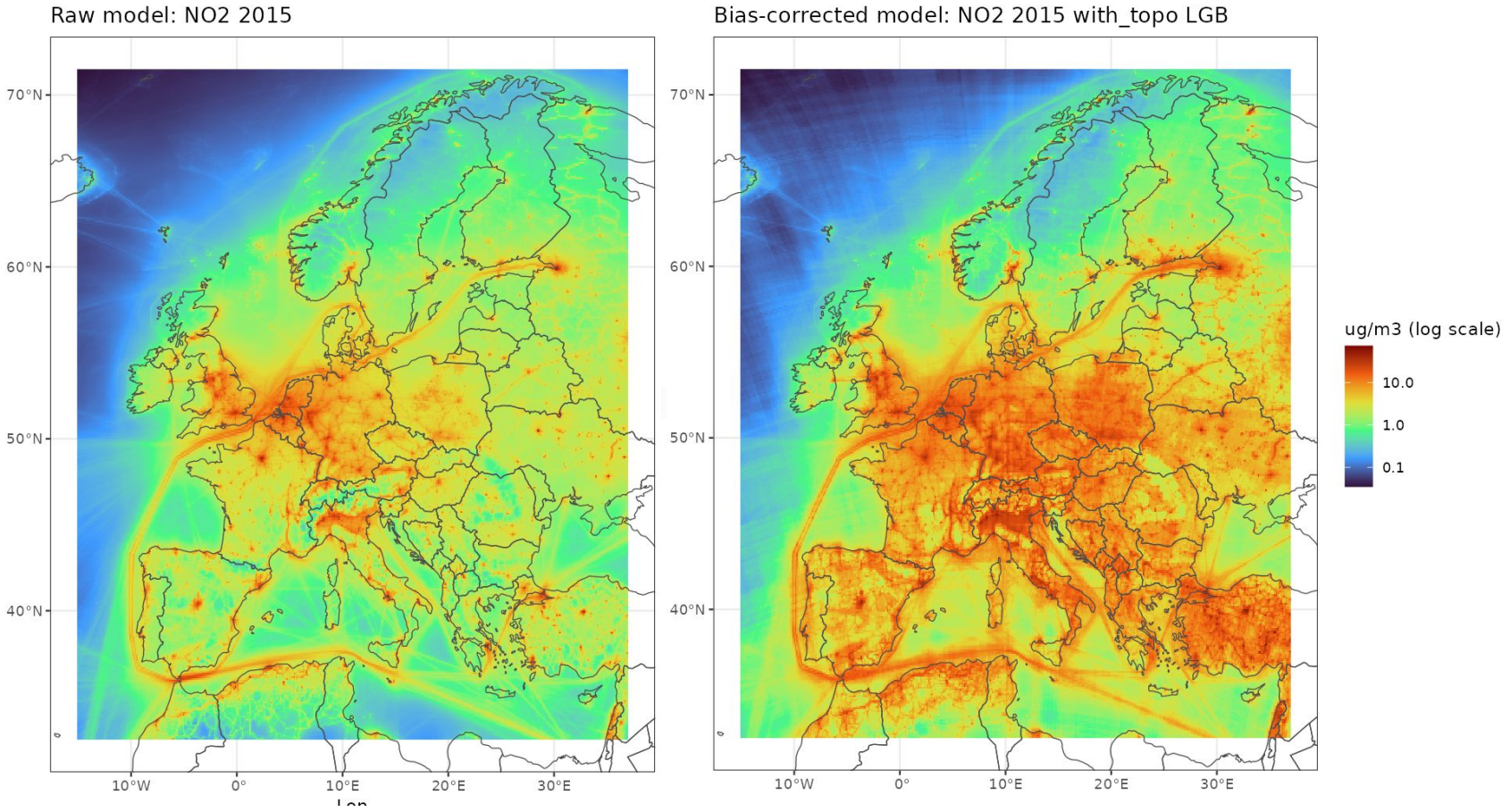
NO_2 Base Case 2015 | Scenario 2022



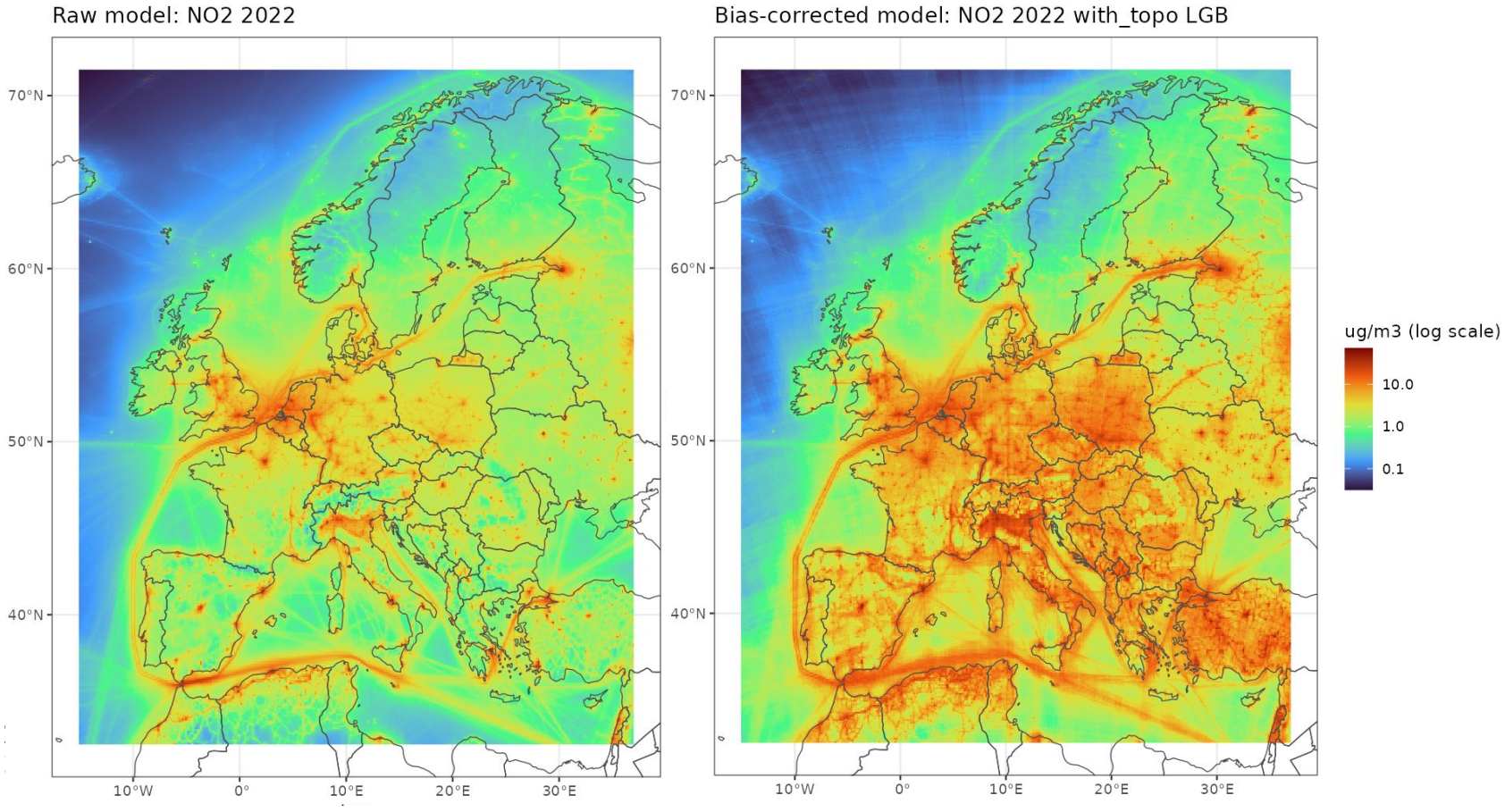


*bilinear interpolation at station

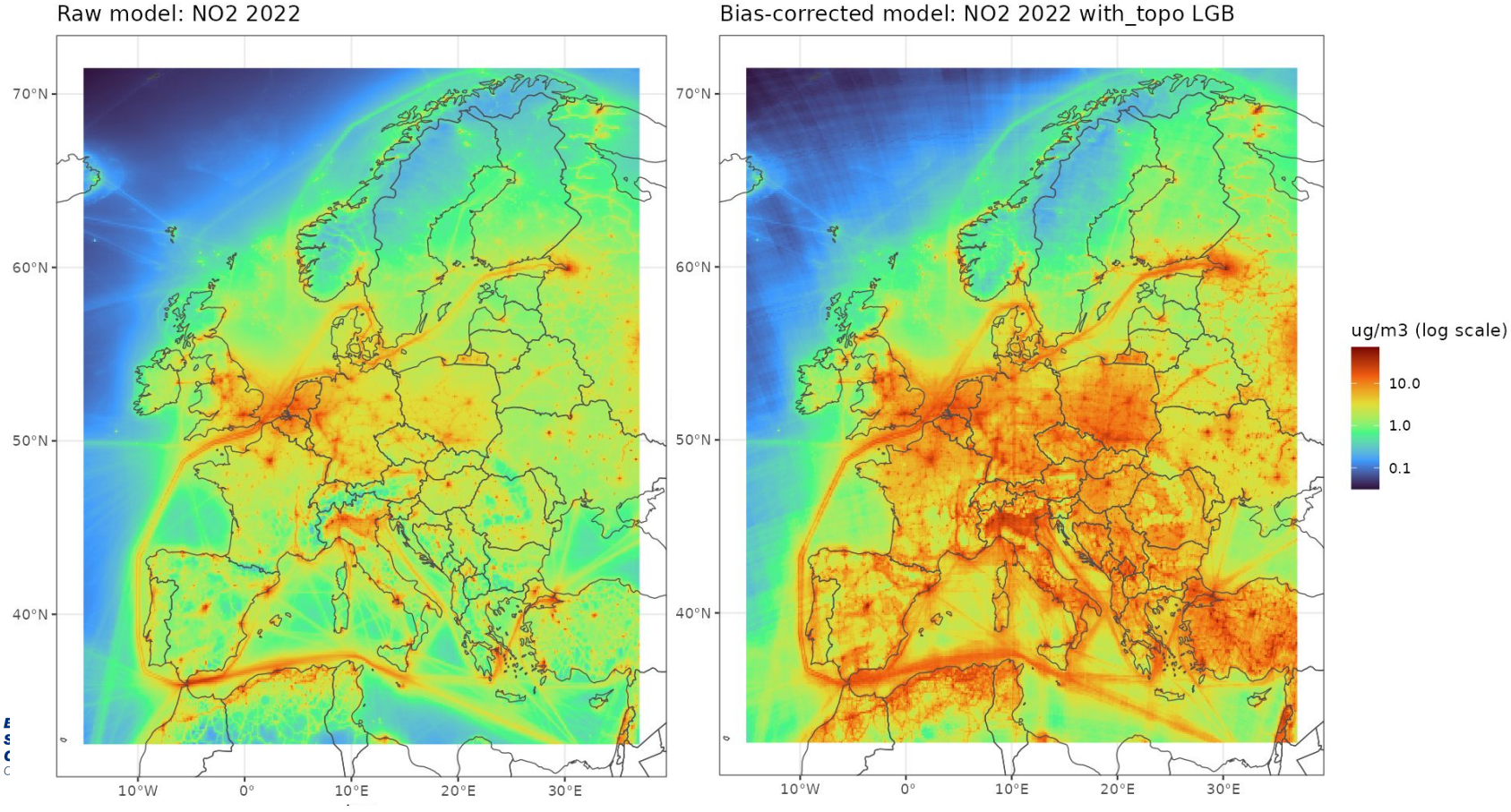
NO_2 Base Case 2015



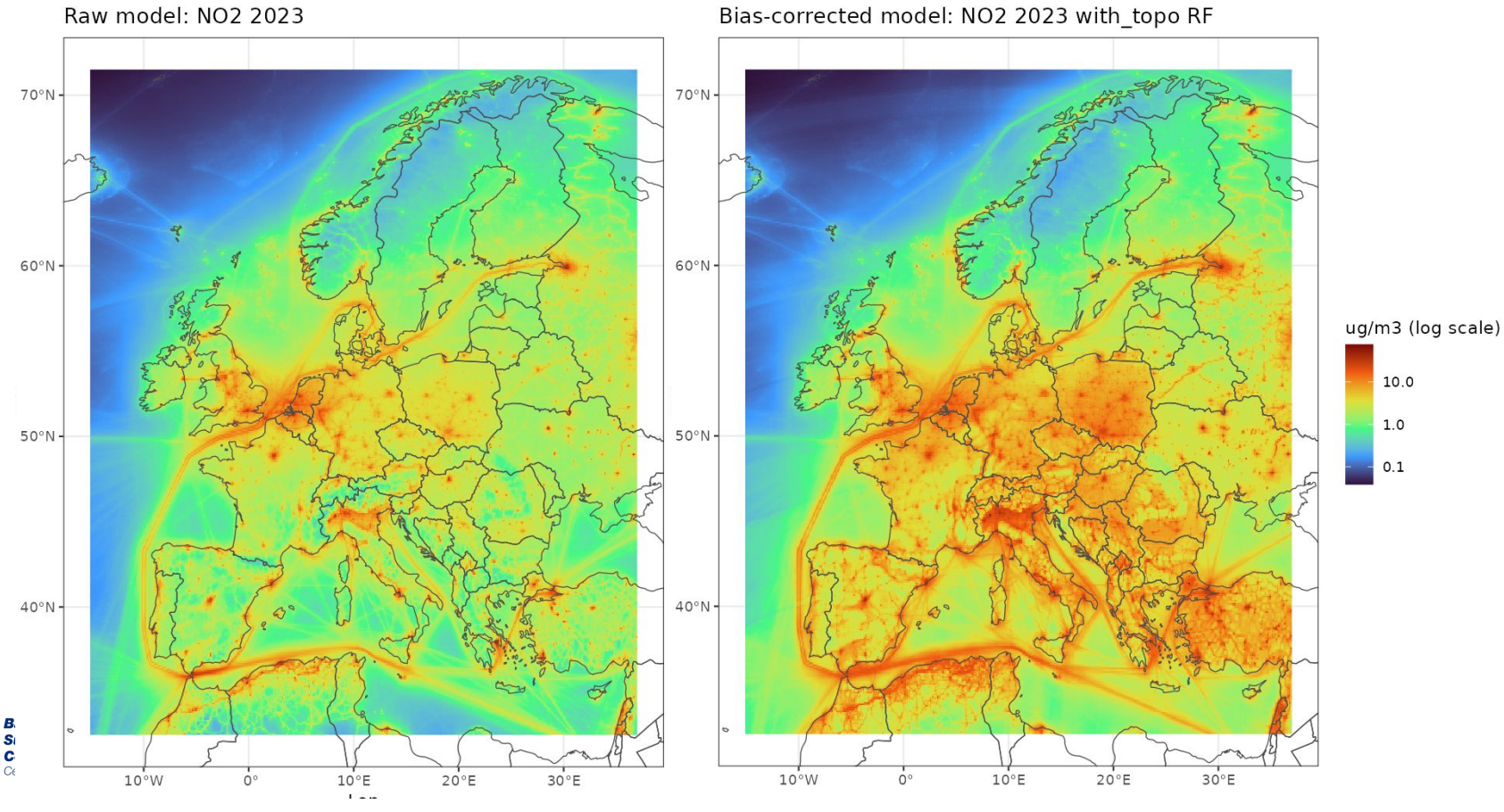
NO_2 Scenario 2022



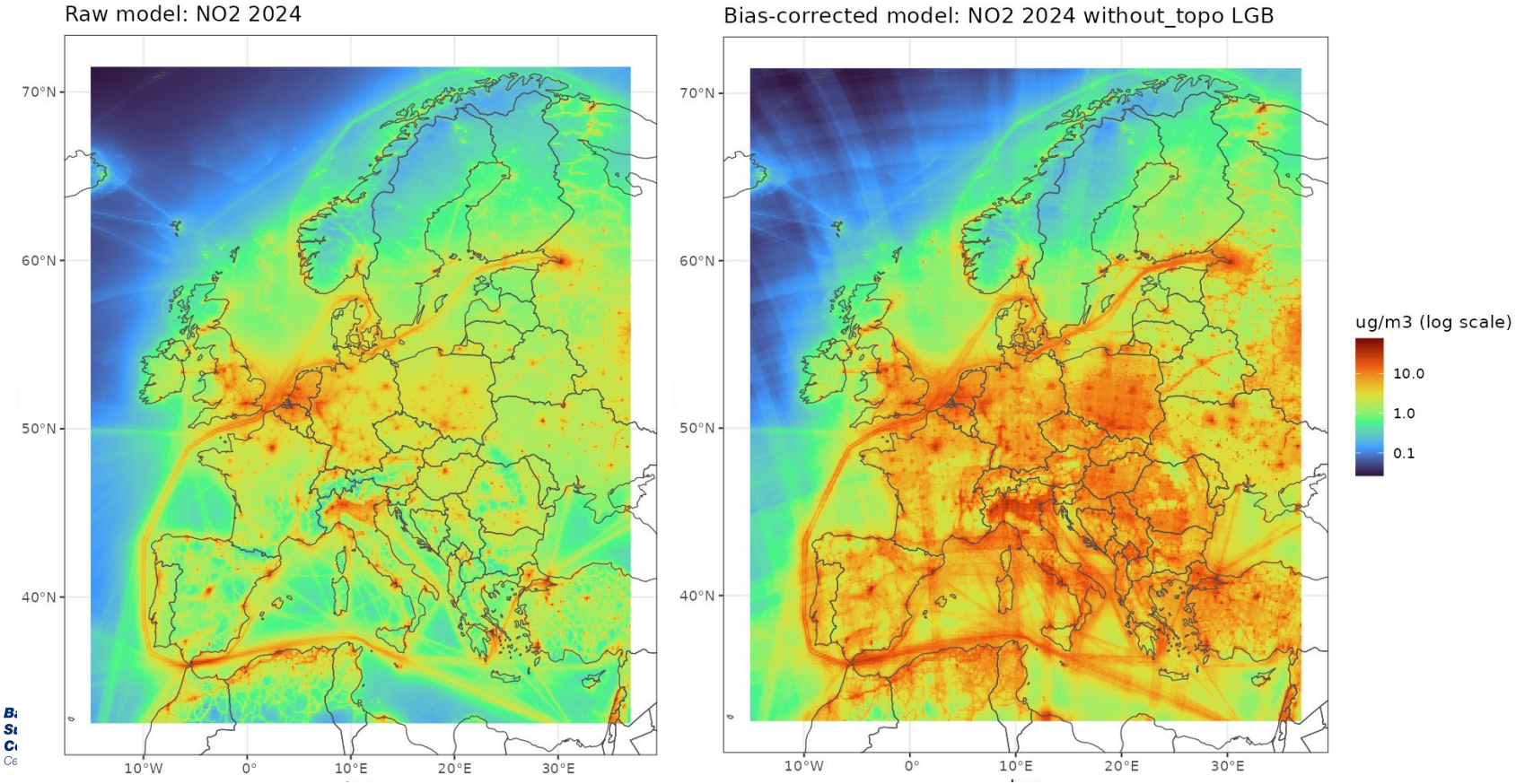
NO_2 Scenario 2022



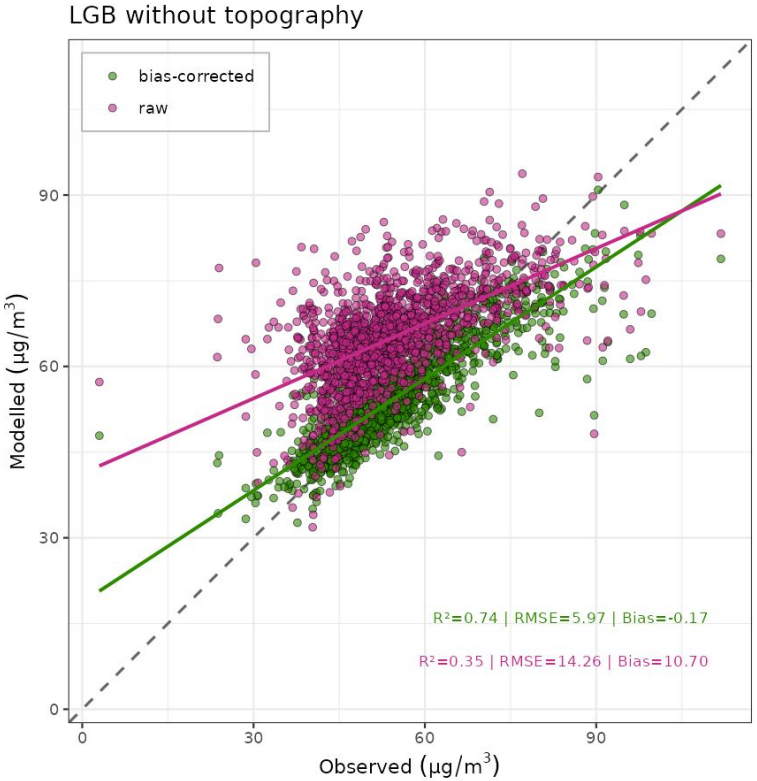
NO_2 Scenario 2023

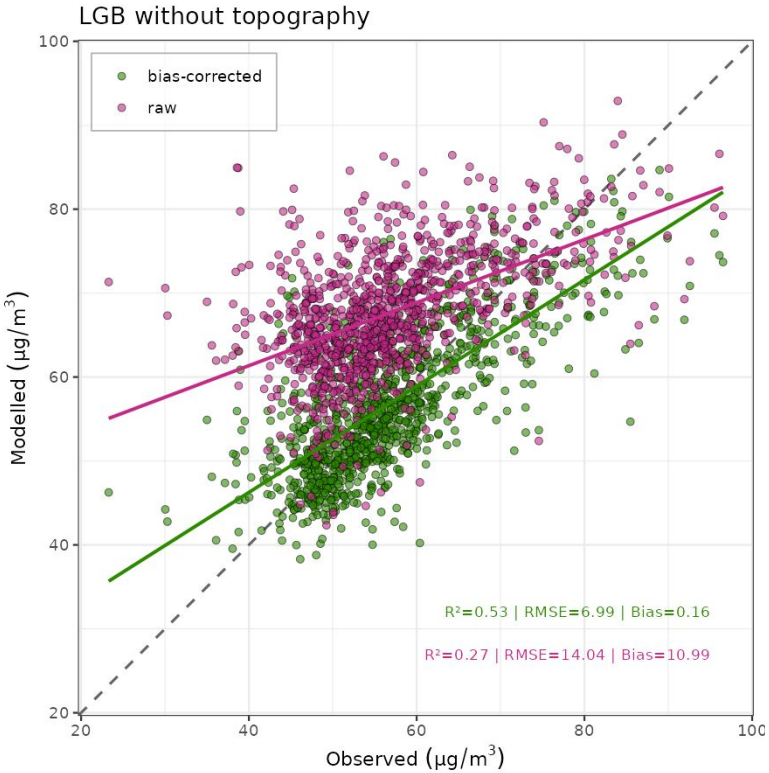


NO_2 Scenario 2024



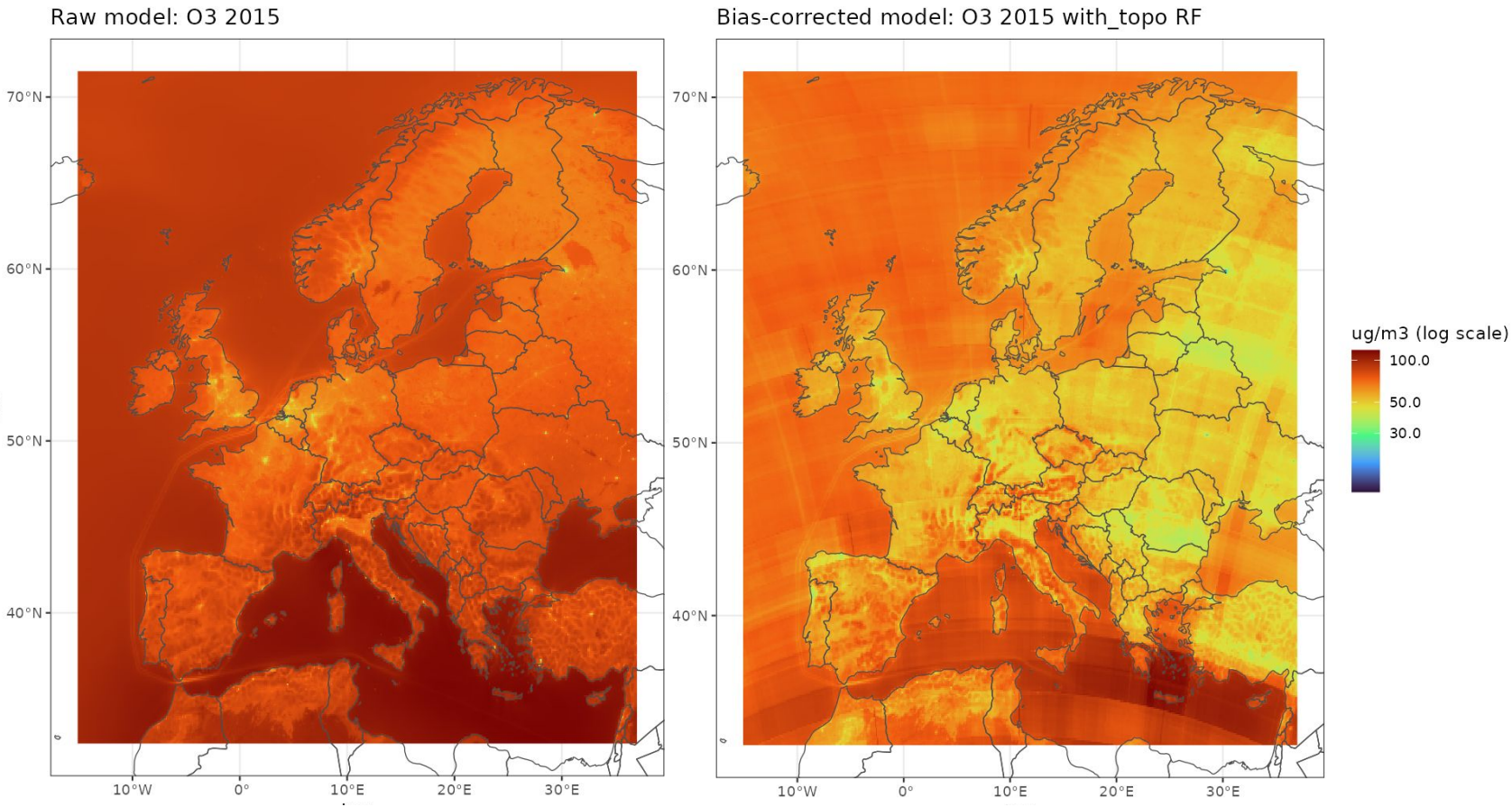
O_3 Base Case 2015 | Scenario 2022



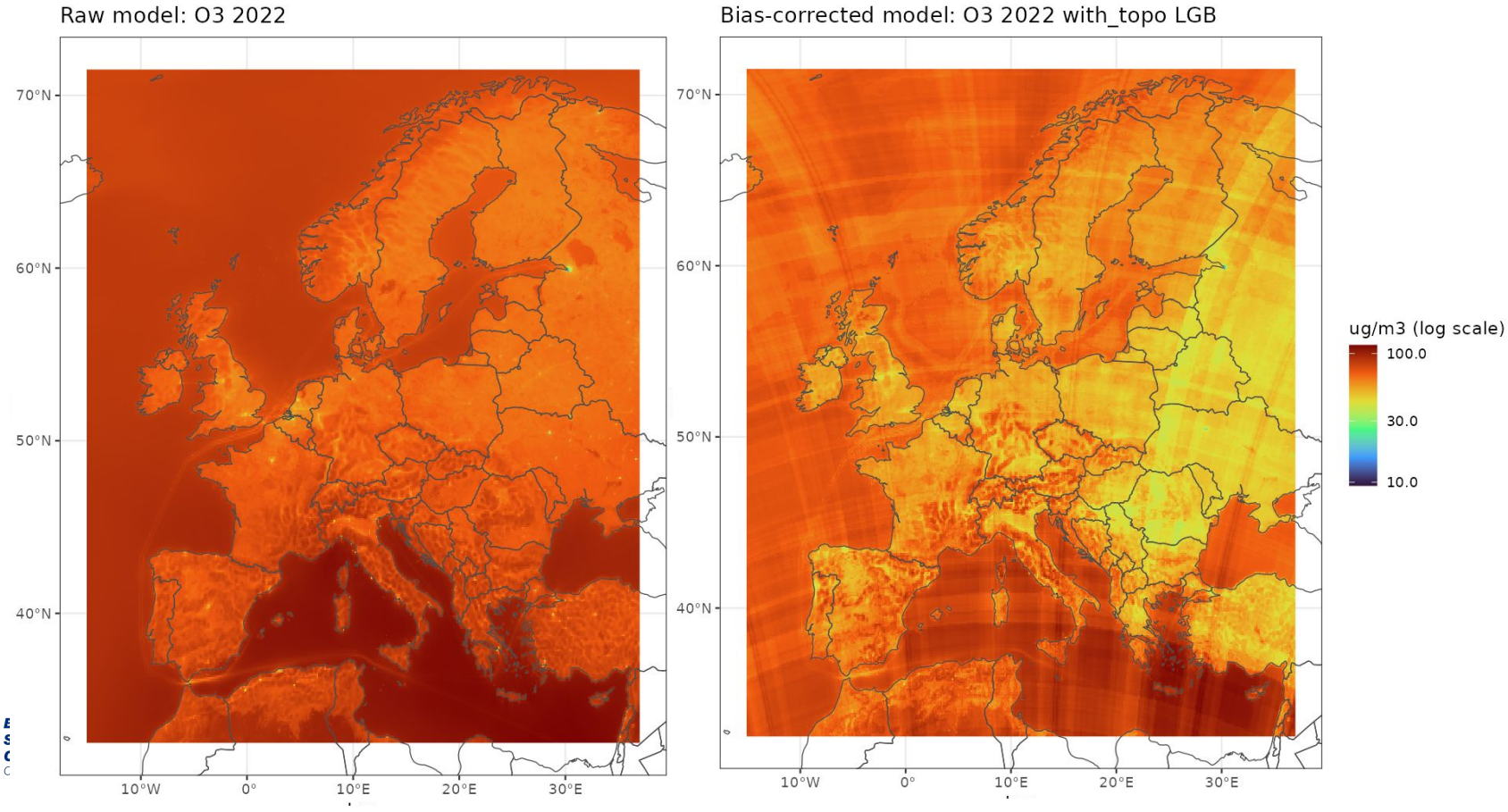


*bilinear interpolation at station

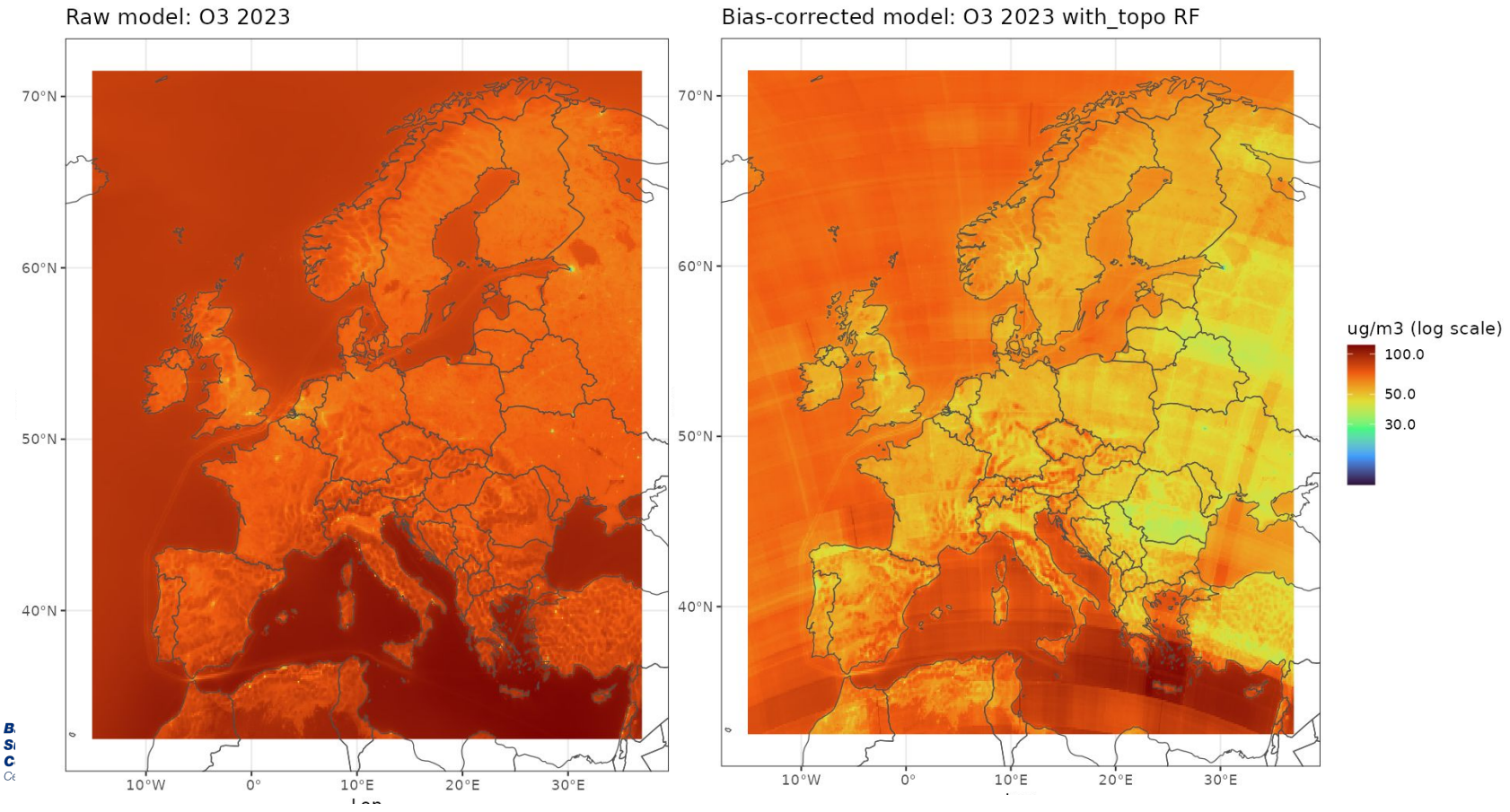
O_3 Base Case 2015



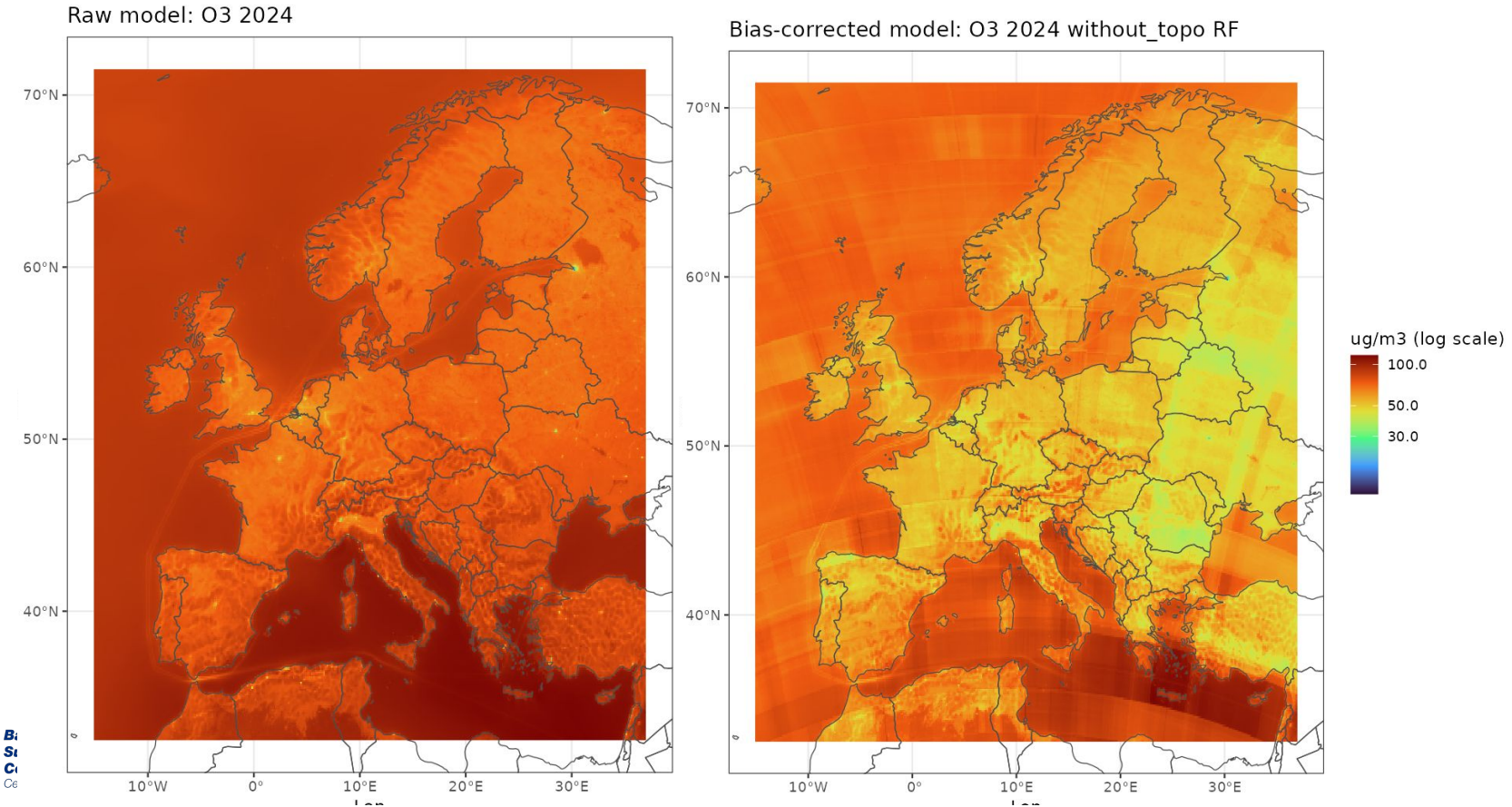
O_3 Scenario 2022



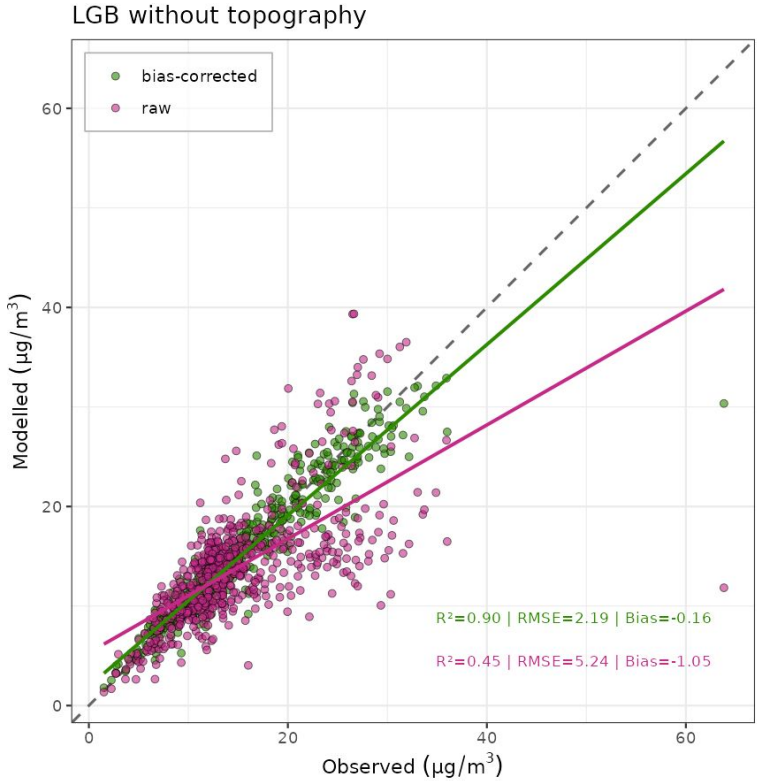
O_3 Scenario 2023

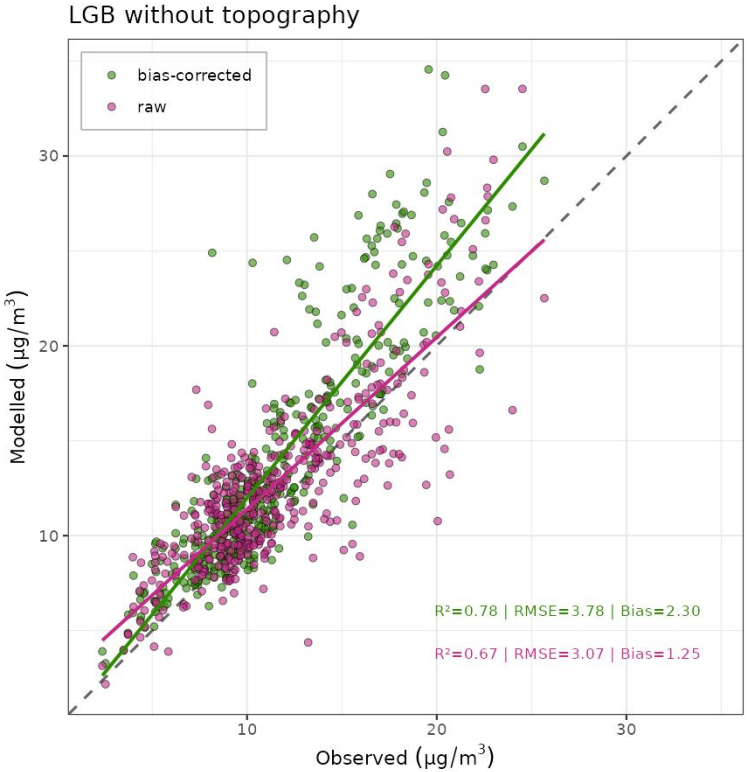


O_3 Scenario 2024



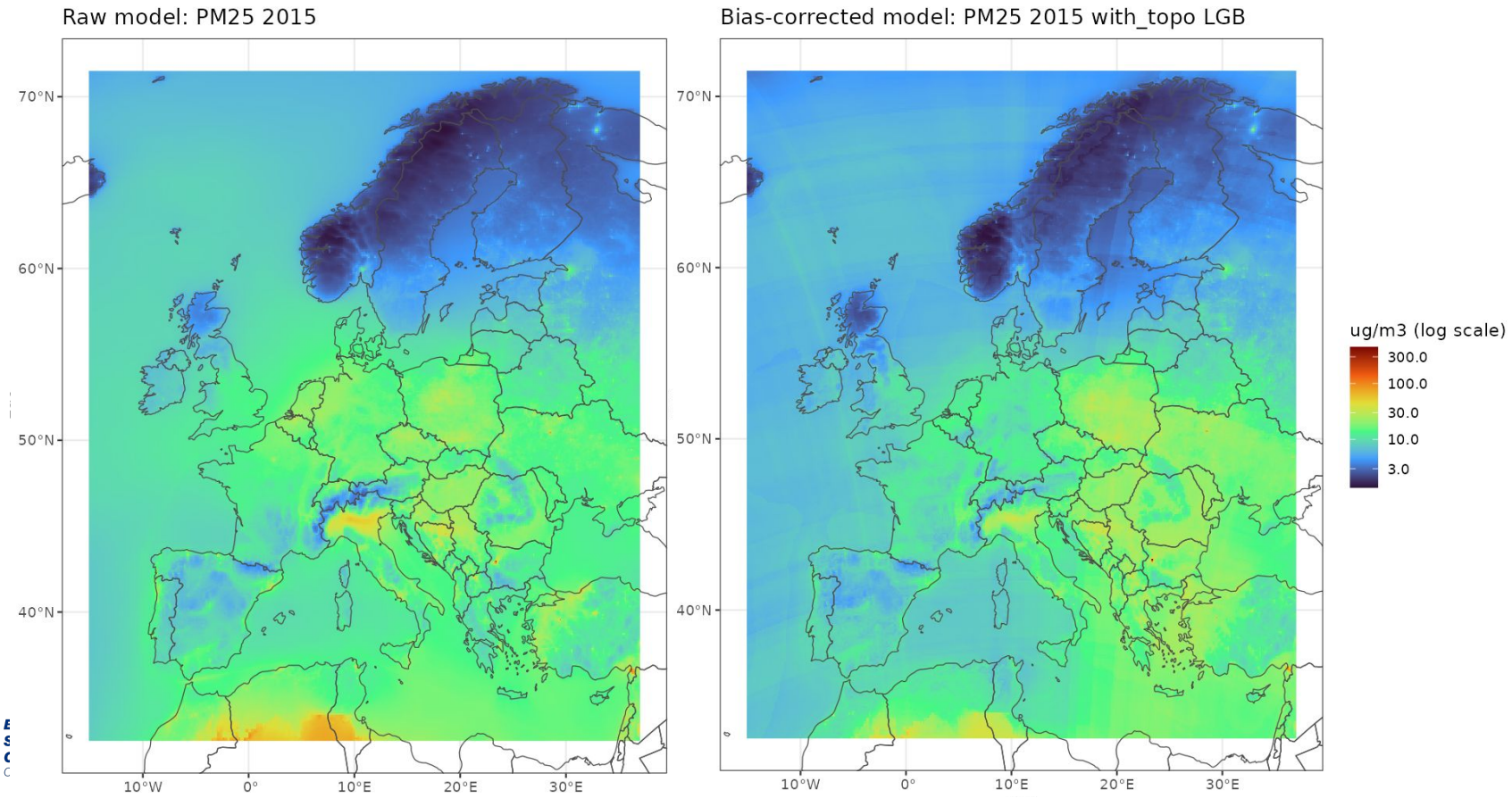
$PM_{2.5}$ Base Case 2015 | Scenario 2022



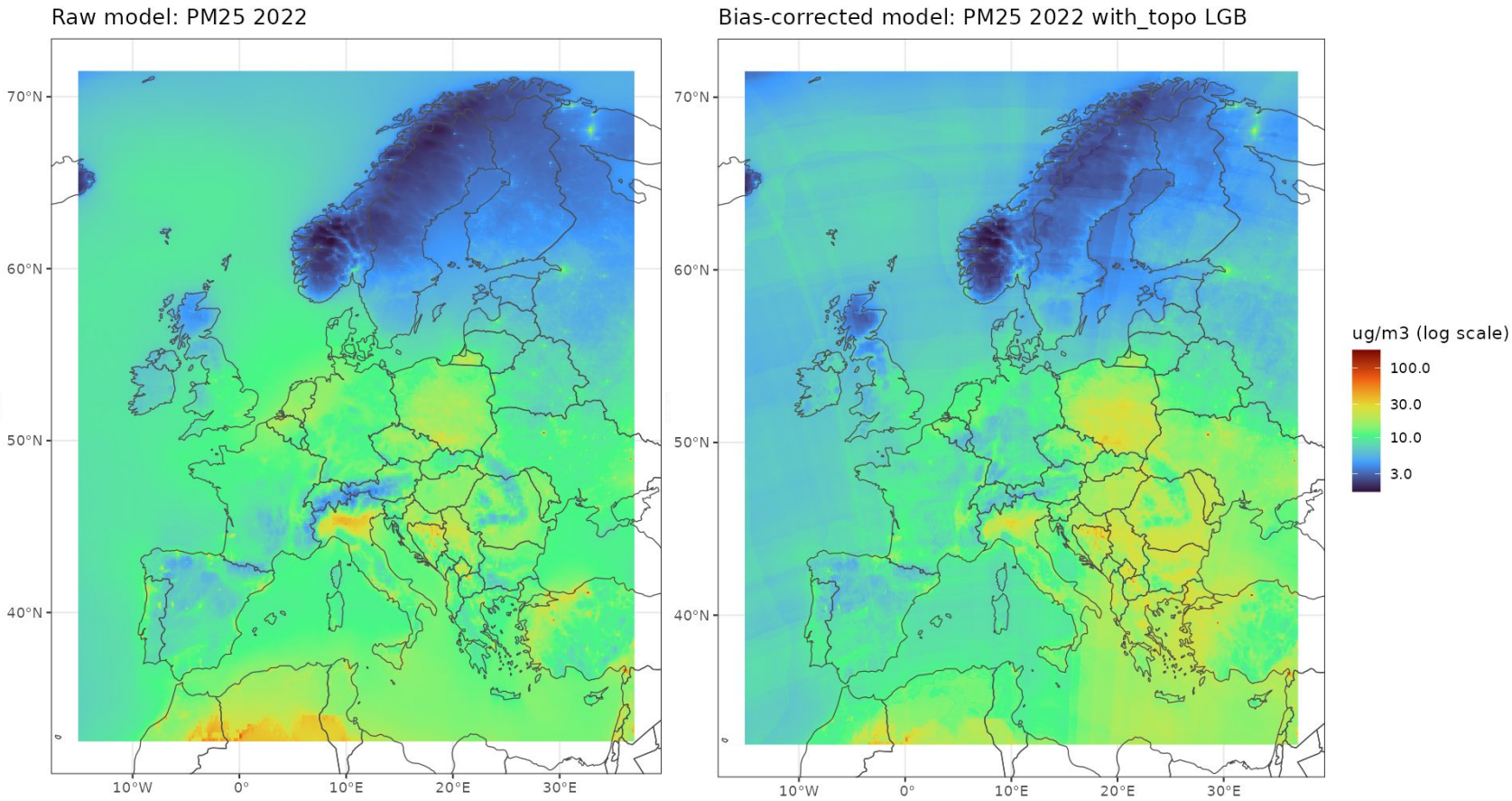


*bilinear interpolation at station

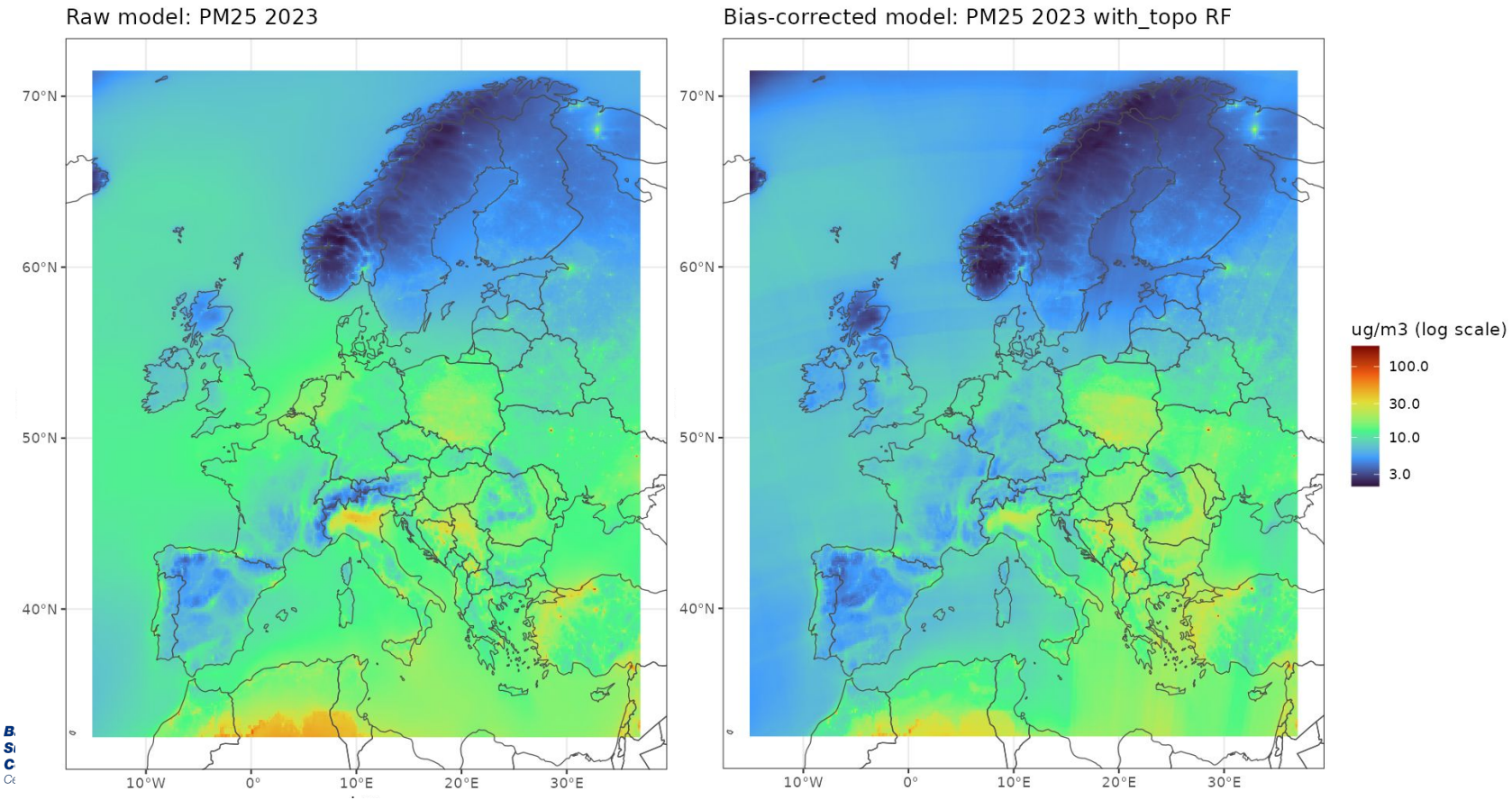
$PM_{2.5}$ Base Case 2015



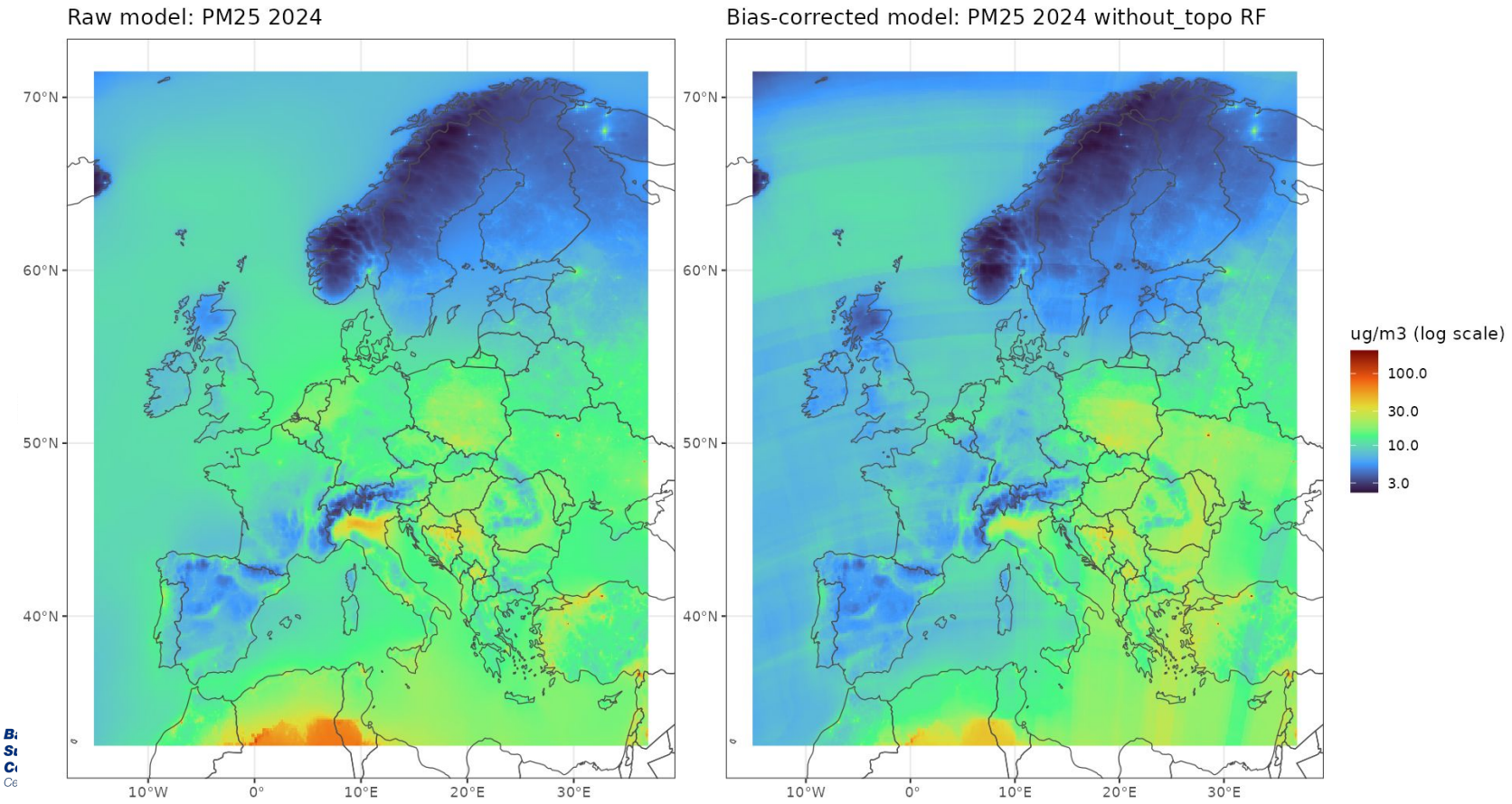
$PM_{2.5}$ Scenario 2022



$PM_{2.5}$ Scenario 2023



$PM_{2.5}$ Scenario 2024



General file description

File format: NetCDF
Projection: lon lat WGS 84 (EPSG:4326)
Spatial resolution lon: 0.1 degrees
Spatial resolution lat: 0.1 degrees

Filename format:
BSC_<phase>_<scenario>_<pollutant>_<sequence>_<calibration>_<correction>_<method>_<data_selection>_<topography>.<format>

<phase>: phase_2
<scenario>: BaseCase_2015; Scenario_2022, Scenario_2023, Scenario_2024
<pollutant>: NO2; O3; PM25
<sequence>: CA
<calibration>: Each
<correction>: Mult
<method>: RF, LGB
<data_selection>: B (only using Background stations in the calibration)
<topography>: with_Topo (used topography from phase_1); without_Topo (not using topography)
<format>: nc
<Temporal_Res>: YEARLY

Step 2: Training ML algorithms

The calibration framework was developed using two **Machine Learning (ML)** techniques:

1. **Random Forest (RF):**
   - Is an ensemble learning method that **constructs** a large number of **decision trees parallelly** during a training stage
   - Each tree is built on a **random bootstrap sample** of the data and a **random subset of predictors** at each split
   - **Predictions** are obtained by **averaging the outputs** of all trees
   - Its main **hyperparameters** are: number of trees, features considered per split, and minimum node size
   - Its parallel and independent architecture makes the method **robust to overfitting**, and **stable under noisy data**

2. **Light Gradient Boosting Machine (LGB):**
   - Is a gradient boosting framework that **builds decision trees sequentially**
   - Each new tree is trained to **correct the residual errors** of the ensemble accumulated so far
   - **Predictions** are obtained by **summing** all the incremental corrections
   - Its main **hyperparameters** are: learning rate, number of leaves and regularisation
   - It requires more careful hyperparameter tuning and early stopping to **prevent overfitting**

Both algorithms where trained using the **same formula**, in which each components is a different predictor. The **target variable** is the **logarithmic ratio** between the observed and modeled value at each monitoring station.

Workflow

Calibration:

1. **Loading data:** extracting at each station (tabular mode) the observational value, modelled value, coordinates and spatial information
2. **Split:**
    a. **Training stations (Active):** used to build the ML algorithm
    b. **Test stations (inActive):** locked away until the validation
3. **5-nested-fold cross-validation:**
    a. **Inner loop:** trying all the hyperparameter combinations (setting of the ML)
    b. **Outer loop:** confirm best hyperparameters on larger held-out folds
4. Pick **hyperparameter set** with the lowest error
5. **Overfitting evaluation:** training with all training stations and predicting the results at the holdout test set, and report RMSE and correlation
6. **Full calibration:** training overall the dataset (training and test) to apply on future scenarios

Application:

7. Prepare a tabular dataset of the future scenario grid cells with: modelled value, topography, sin/cos of the coordinates
8. The algorithm is then applied to future scenario data, under the assumption that **the model's systematic errors behave similarly in the future as they did historically**
9. **Multiply each grid cell** with the ML estimated value