Data Assimilation Schemes in Colombian Geodynamics - Cooperative Research Plan for 2017 - 2020 Between Universidad EAFIT and TUDelft, with the Help of Universidad de Antioquia and Universidad Nacional de Colombia Sede Medellin

Start date: 1 January 2017
End date: 30 December 2020

Statistical error analysis between LOTOS-EUROS and MACC model

Medellin Air qUality Initiative MAUI

MAUI-RP006

Executing Entity	Universidad EAFIT Cra 49 No 7sur - 50 Medellín, Colombia
	Grupo de investigación en modelado matemático – GRIMMAT Grupo reconocido por COLCIENCIAS Categoría A Grupo de investigación en Biodiversidad, Evolución y Conservación - BEC
Responsible	Prof. Olga Lucia Quintero Montoya Prof. Nicolás Pinel Peláez. Researchers
Cooperating Entities	Department of Applied Mathematics - Tu Delft. Delft, The Netherlands TNO Universidad de Antioquia
Responsible	Arnold Heemink Martin Verlaan Arjo Segers Martjin Schaap Angela Rendón

EDITION AND DISTRIBUTION CONTROL

Edition	Action*	Name	Signature	Entity	Date (DD/MM/YYYY)
1	Creation	Andrés Yarce		Universidad EAFIT	16/06/2017
2	Modification	Santiago López		Universidad EAFIT	16/06/2017
3	Modification	Martín Rodríguez		Universidad EAFIT	16/06/2017
4	Modification	Nicolás Pinel		Universidad EAFIT	18/06/2017
4	Revision	O. Lucía Quintero		Universidad EAFIT	19/06/2017
5	Modification	Nicolás Pinel		Universidad EAFIT	22/06/2017

* Specify type of action: Creation - Revision - Modification - Distribution.

Content

Introduction 4

Methodology 4

1.1 Mean Square Error 5

1.2 Fractional Bias (FB) 6

1.3 Normalized Mean Square Error (NMSE) 6

1.4 Correlation Coefficient (R) 7

Results 8

1.1 Carbon monoxide (CO) 0.14º LE grid 8

1.2 Carbon monoxide (CO) 0.25º LE grid 9

1.3 Nitrogen dioxide ($NO2$) 0.14º LE grid 10

1.4 Nitrogen dioxide ($NO2$) 0.25º LE grid 11

1.5 Ozone ($O3$) 0.14º LE grid 12

1.6 Ozone ($O3$) 0.25º LE grid 13

1.7 Sulfur dioxide ($SO2$) 0.14º LE grid 14

1.8 Sulfur dioxide ($SO2$) 0.25º LE grid 15

1.9 Summarized Statistics 16

DISCUSSION 18

CONCLUSION 18

References 19

INTRODUCTION

The present report describes the procedures implemented with the objective of assessing the initial performance of the model LOTOS-EUROS in the Tropical Andes Domain, evaluated against the Copernicus MACC data. This evaluation aimed to

1.) establish the feasibility of running the model within the domain;

2.) initiate the development of an evaluation framework at the domain level;

3.) identify areas of the domain that may present systematic trends on the evaluated statistics in order to recognize potential difficulties for the model.

METHODOLOGY

The model set up was:

Start date: 2015-04-24 06:00:00

End date: 2015-12-24 24:00:00

The Southwest corner of the domain was located at: -79.9° lon / -3.8° lat

Simulation resolutions:

1.) 0.14° lat x 0.14° lon
2.) 0.25° lat x 0.25° lon

Spacing: nx = 102; ny = 121

The outputs of the model on the variables listed below were evaluated against Copernicus MACC data for the final date. The performance statistics were calculated for every grid within the domain. The evaluated variables were:

- Carbon Monoxide (CO)
- Nitrogen Dioxide (NO_2)
- Ozone (O_3)
- Sulphure Dioxide (SO_2)

The following statistics were obtained as metrics of the performance of the model when compared to the Copernicus MACC data. It is proposed that similar statistics be adopted once the meteorological data from the IDEAM become available.

The following conventions are used (Chang & Hanna, 2004; Chen *et al*., 2014; Cox & Tikvart, 1990; Hanha, 1988; Thunis, Pederzoli, & Pernigotti, 2012):

Cp: Model output,

Co: Observations,

$\acute{C}$: Average time series,

σ_c: Standard deviation of the time series.

1.1 Mean Square Error

The Mean Square Error (MSE) is an estimator that measures the error square between the model and the observations. The MSE may be calculated as (Poli & Cirillo, 1993; Solazzo & Galmarini, 2016):

$$\mathrm{MSE} = E(\text{mod-obs})^2 = \frac{\sum_{i=1}^{nt}(\text{mod}_i - \text{obs}_i)^2}{n_t},$$

Where, mod_1 is the model output at the time i, obs_i is the observation at the time i and n_t is the maximum observation time. The estimator can be decomposed as the sum of the variance and the bias squared between the model and the observations.

$$\mathrm{MSE} = \mathrm{var}(\text{mod-obs}) + \mathrm{bias}^2,$$

This expression may be calculated as:

$$\mathrm{MSE} = (\overline{\text{mod}} - \overline{\text{obs}})^2 + (\sigma_{\text{mod}} - \sigma_{\text{obs}})^2 + 2(1-r)\sigma_{\text{mod}}\sigma_{\text{obs}}$$

Where, are the $\acute{mod}\ y\ \acute{obs}$ means of the model output and the observations respectively, $\sigma_{mod}\, y\, \sigma_{obs}$ are the variances of the model output and the observations respectively, and *r* is the correlation coefficient between both series of time. The bias between the model and the observations measures the systematic error of the model.

The bias is commonly used to measure the proximity between the mean value of the time series of the model and the reality (expressed in the observations). The variance between the model and the observation shows whether the variability of the model is compatible with the variability of the observations. Finally, the covariance term represents the unexplained proportion of the MSE due to the remaining non-systematic errors, that is to say, it represents the remaining error after deviations of the mean values have been considered. This last term is a measure of the lack correlation of the model with comparable observations and is considered the least disturbing part of the error.

1.2 Fractional Bias (FB)

The fractional Bias (FB) measures the systematic error of the model outputs against the observations. With the FB value it is possible to conclude if the model tends to underestimate or overestimate the variable values.

The FB is based on a linear scale and the systematic bias refers to the arithmetic difference between Cp and Co . The FB values are between -2 and 2. Positive values denote a model underestimation, while negative values denote a model overestimation. An FB value of 0 shows a perfect model estimate. The fractional bias was chosen because of two desirable characteristics. First, the fractional bias is symmetric and limited. The values for the fractional bias oscillate between -2.0 (extreme over prediction) and +2.0 (extreme under prediction). Second, the fractional bias is a dimensionless number, which makes it convenient to combine the results of the data categories that have significantly different concentration levels.

$$\mathrm{FB} = \frac{(\overline{\mathrm{C_o}} - \overline{\mathrm{C_p}})}{0.5(\overline{\mathrm{C_o}} + \overline{\mathrm{C_p}})},$$

1.3 Normalized Mean Square Error (NMSE)

Contrary to bias, in the NMSE the deviations are summed (absolute values) instead of the differences. If the model has a very low NMSE, it is performing well in space and time. On the other hand, high NMSE values do not necessary mean that a model is completely wrong. That case could be due to the change of time and/or space. Moreover, it should be noted that differences in peaks have a higher weight in NMSE than differences in other values.

$$\mathrm{NMSE} = \frac{\overline{(\mathrm{C_o} - \mathrm{C_p})^2}}{\overline{\mathrm{C_o}}\,\overline{\mathrm{C_p}}},$$

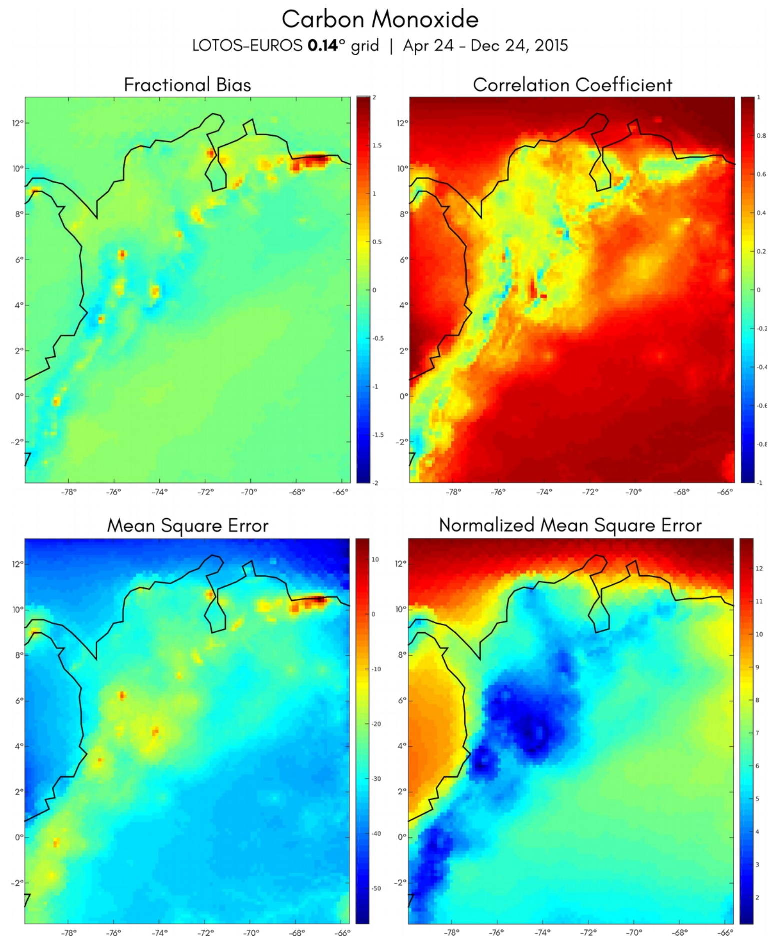
1.4 Correlation Coefficient (R)

The correlation coefficient is a measure that determines the degree to which two dynamic variables are associated in time (for this case). The range of values for the correlation coefficient is from -1.0 to 1.0. A correlation of -1.0 indicates a perfect negative correlation, while a correlation of 1.0 indicates a perfect positive correlation.

$$R = \frac{\overline{(C_o - \overline{C_o})(C_p - \overline{C_p})}}{\sigma_{C_p}\sigma_{C_o}}$$

RESULTS

1.1 Carbon monoxide (CO) 0.14º LE grid

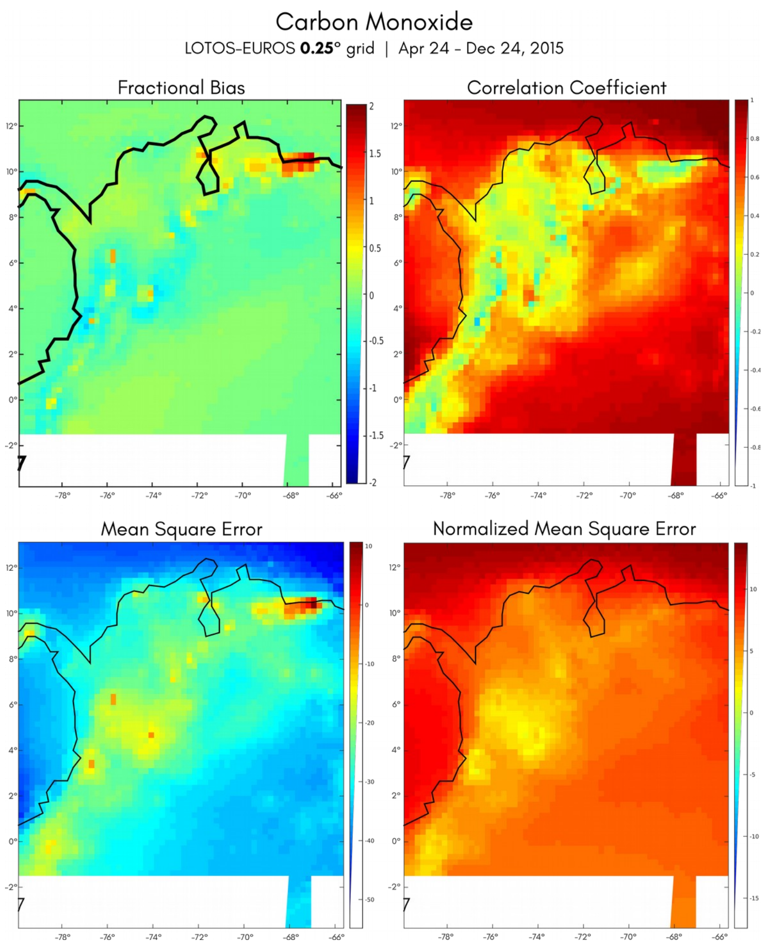


Aggregate Statistics:

	Fractional Bias	Correlation	Mean Square Error	Normalized Mean Square Error
Minimum	-0.7277	-0.4353	2.24e-06	1.1849

Maximum	1.8820	0.9960	2.38e+01	12.9092
Median	-0.0609	0.6070	8.80e-03	6.8665
Sum			1.09e+02	8.4746e+04

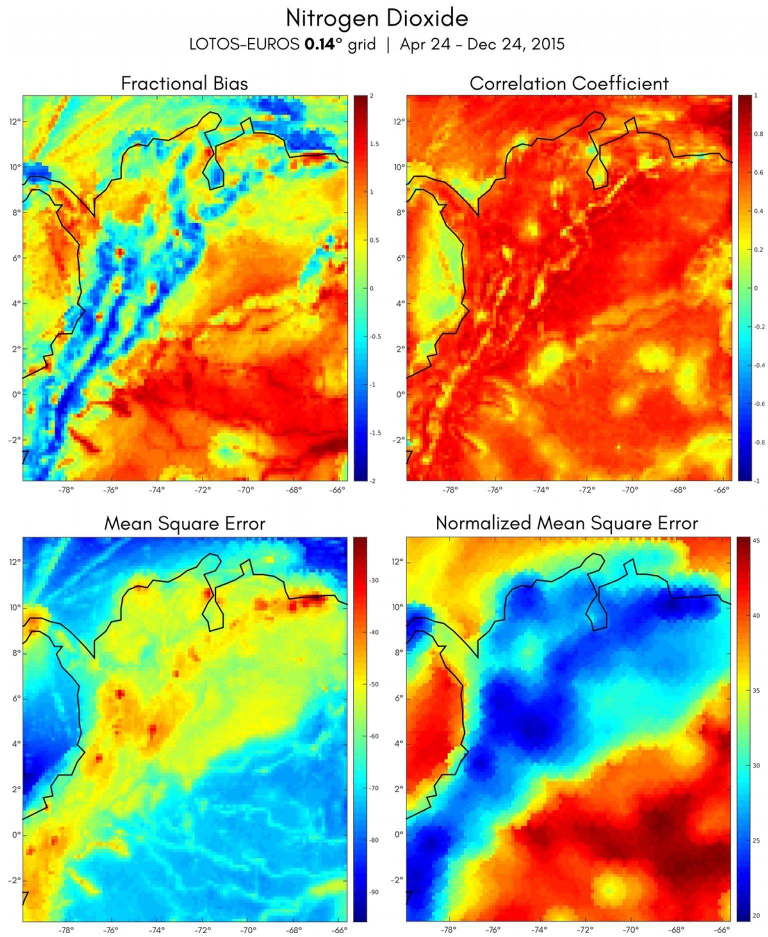
1.2 Carbon monoxide (CO) 0.25º LE grid



Aggregate statistics:

	Fractional Bias	Correlation	Mean Square Error	Normalized Mean Square Error
Minimum	-0.8104	-0.3488	3.25e-06	-17.4365
Maximum	1.8109	0.9943	1.16e+01	13.7285
Median	-0.0697	0.7017	7.00e-03	8.2759
Sum			5.91e+01	6.9517e+04

1.5 Nitrogen dioxide (NO_2) 0.14º LE grid

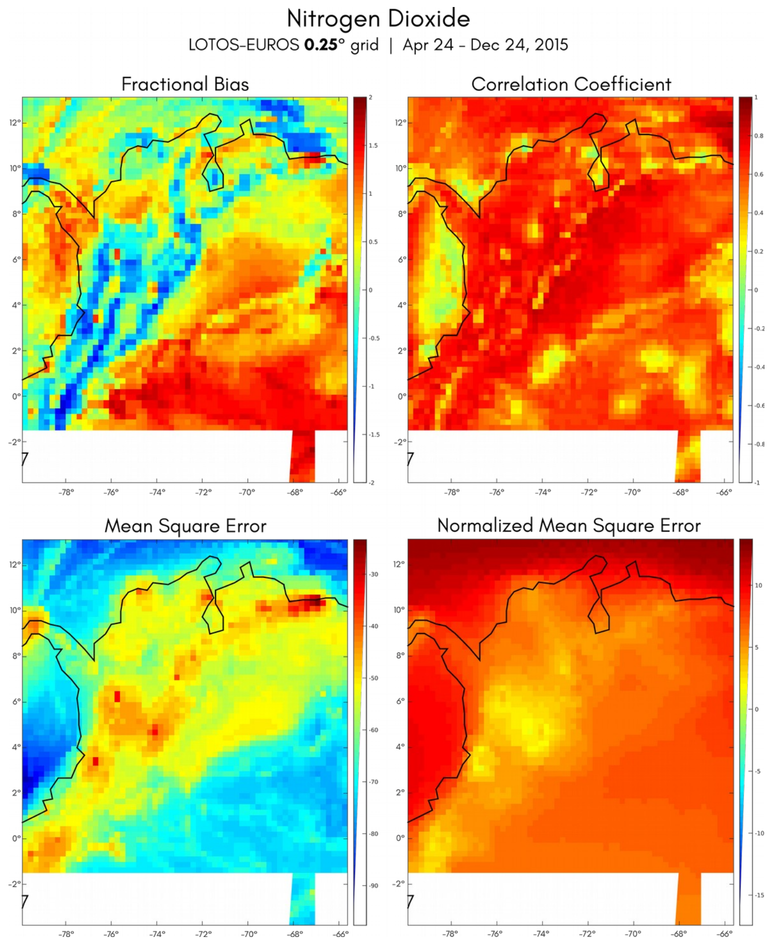


Aggregate statistics:

	Fractional Bias	Correlation	Mean Square Error	Normalized Mean Square Error
Minimum	-1.7572	-0.2798	2.66e-10	91.1883
Maximum	1.8893	0.9630	6.70e-03	3.3388e+04
Median	0.4164	0.5561	7.91e-06	5.7793e+03

Sum			9.76e-02	7.1328e+07

1.6 Nitrogen dioxide (NO_2) 0.25º LE grid

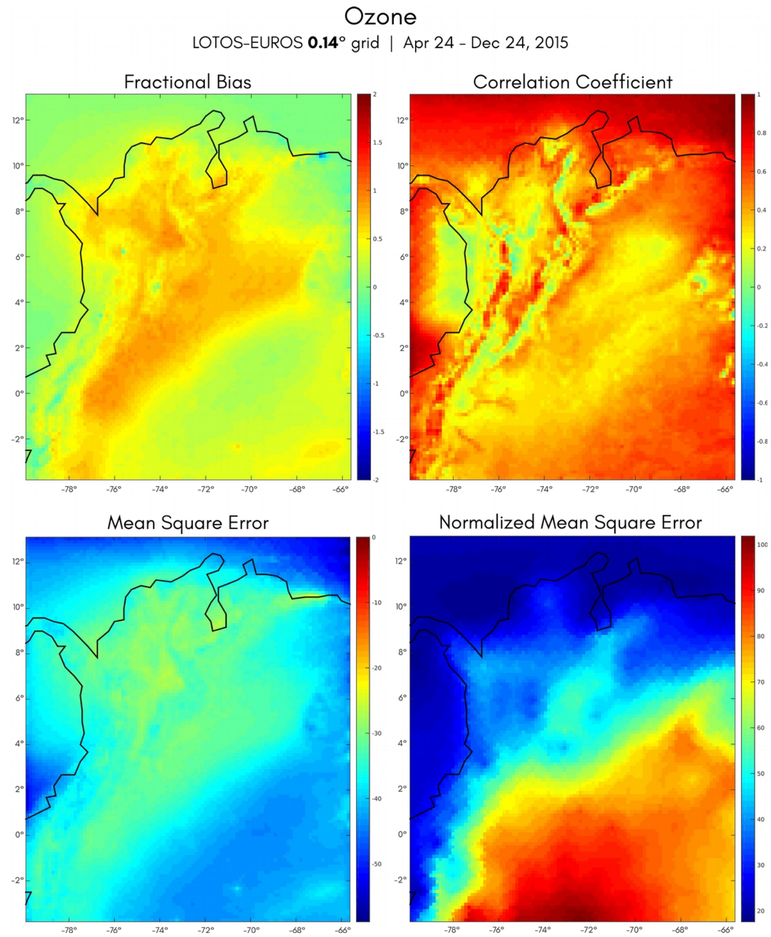


Aggregate statistics:

	Fractional Bias	Correlation	Mean Square Error	Normalized Mean Square Error

Minimum	-1.6892	-0.3191	1.72e-10	102.2123
Maximum	1.8885	0.9657	4.60e-03	3.3580e+04
Median	0.3393	0.5509	4.52e-06	5.7891e+03
Sum			3.80e-02	4.8629e+07

1.7 Ozone (O_3) 0.14º LE grid

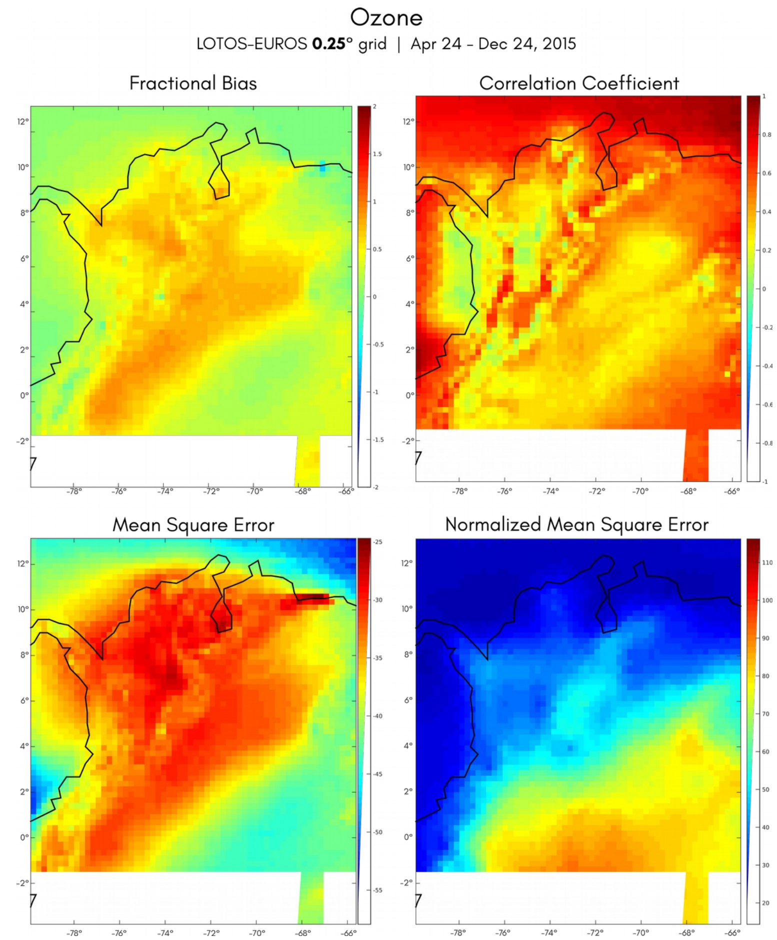


Aggregate statistics:

	Fractional Bias	Correlation	Mean Square Error	Normalized Mean Square Error
Minimum	-0.9412	-0.1719	1.30e-06	17.6125
Maximum	0.9555	0.9846	4.90e-03	101.8686
Median	0.3526	0.4564	3.20e-04	49.8383

Sum			3.95e+00	6.1510e+05

1.8 Ozone (O_3) 0.25º LE grid

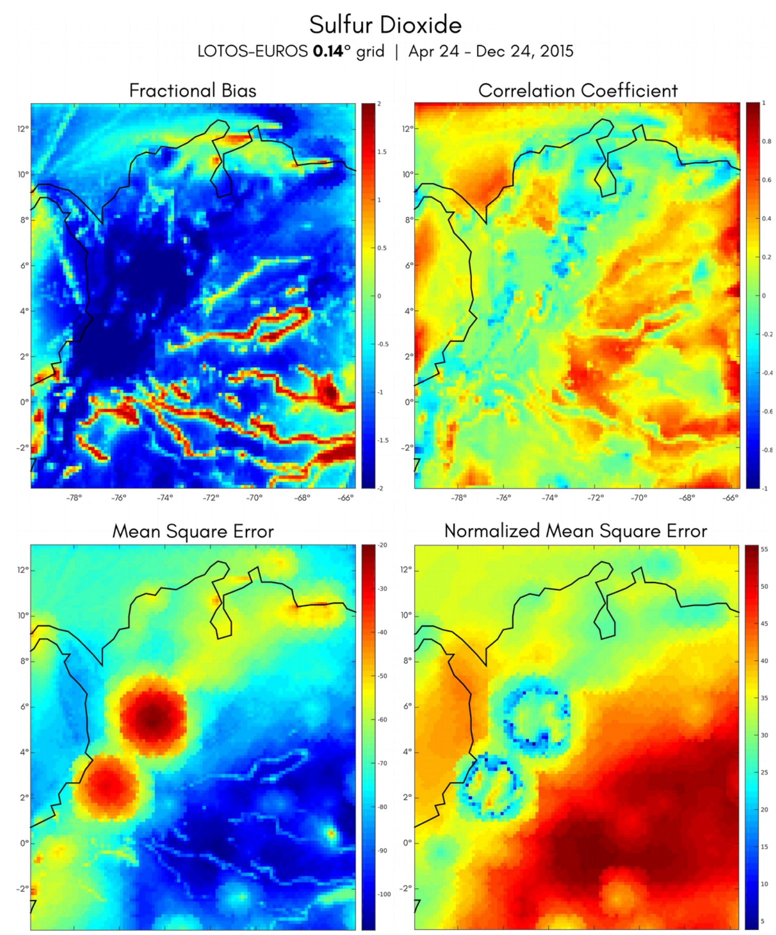


Aggregate statistics:

	Fractional Bias	Correlation	Mean Square Error	Normalized Mean Square Error
Minimum	-0.7087	-0.0305	1.61e-06	14.3536

Maximum	0.9304	0.9815	3.40e-03	116.2194
Median	0.2310	0.5972	2.02e-04	44.9223
Sum			1.70e+00	3.7735e+05

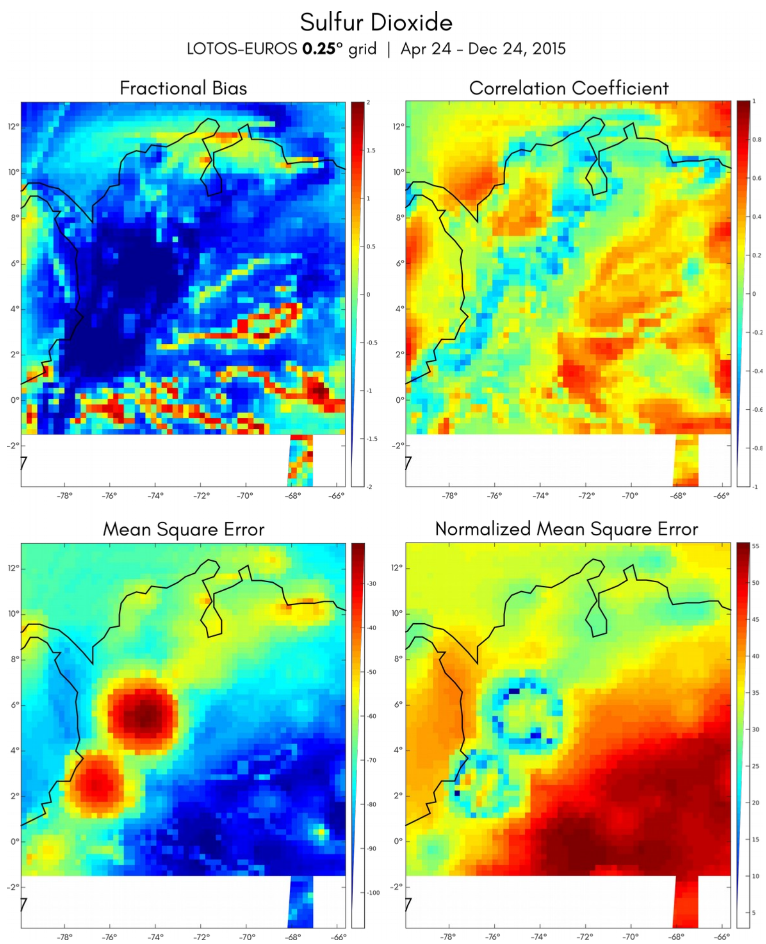
1.9 Sulfur dioxide (SO_2) 0.14º LE grid



Aggregate statistics:

	Fractional Bias	Correlation	Mean Square Error	Normalized Mean Square Error
Minimum	-1.9986	-0.5936	1.49e-11	-6.3021e+03
Maximum	1.9936	0.9355	1.00e-02	3.5622e+05
Median	-0.9003	0.1719	4.75e-05	4.4377e+04
Sum			5.86e-01	5.4771e+08

1.10 Sulfur dioxide (SO_2) 0.25º LE grid



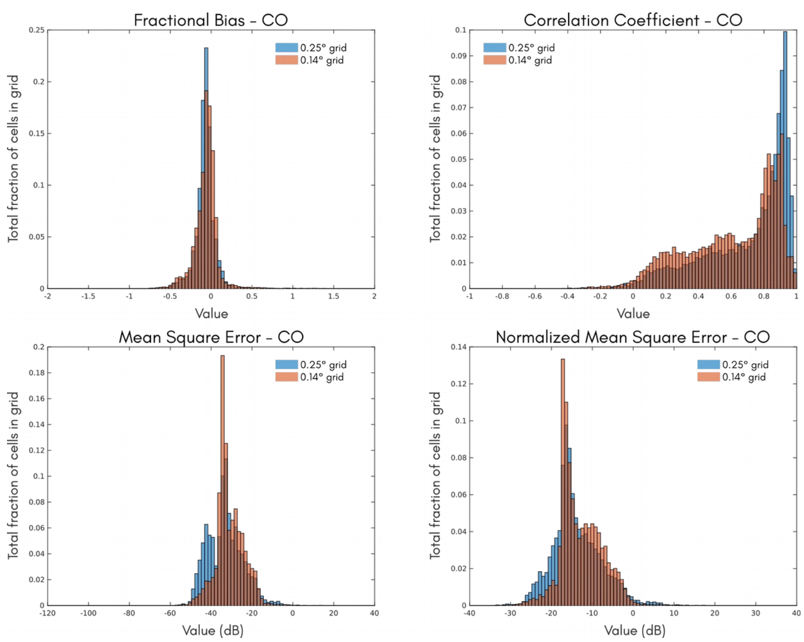
Aggregate statistics:

	Fractional Bias	Correlation	Mean Square Error	Normalized Mean Square Error
Minimum	-1.9989	-0.6621	1.56e-11	-6.0989e+03
Maximum	1.9853	0.8940	8.20e-03	3.5069e+05
Median	-0.8503	0.2417	2.42e-05	2.6770e+04

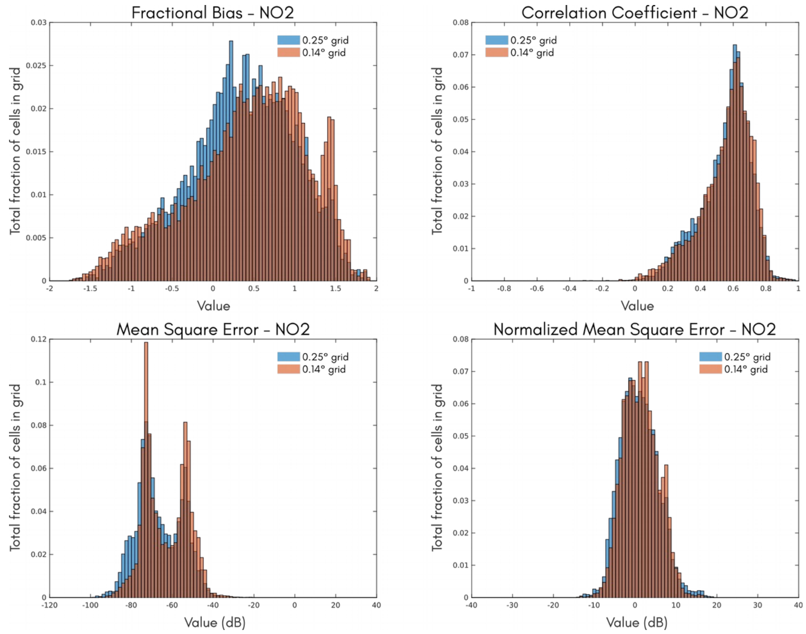
Sum			2.03e-01	2.2487e+08

1.11 Summarized Statistics

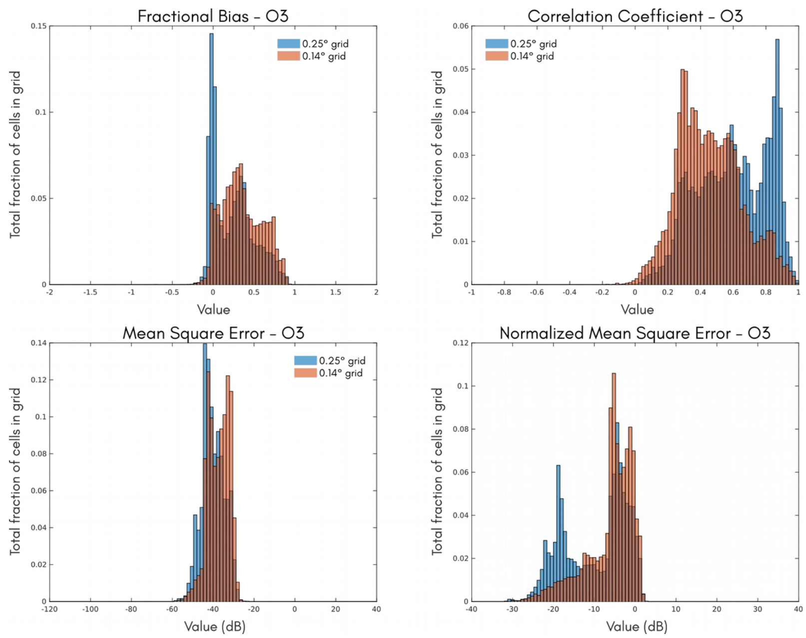
Carbon monoxide (CO)



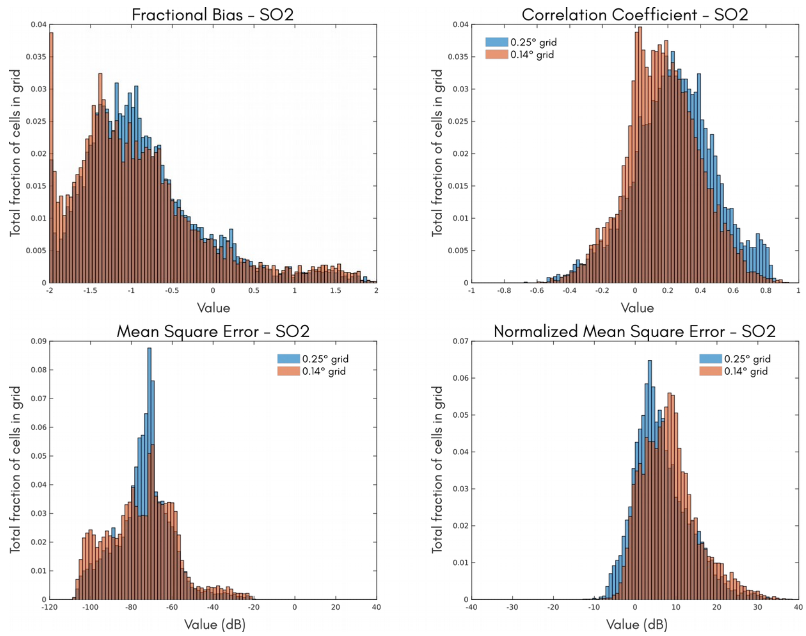
Nitrogen dioxide (NO_2)



Ozone (O_3)



Sulfur dioxide (SO_2)


Fractional Bias - SO2
Correlation Coefficient - SO2
Mean Square Error - SO2
Normalized Mean Square Error - SO2
Total fraction of cells in grid
Value
Value (dB)
0.25° grid
0.14° grid

DISCUSSION

From the values of Fractional Bias over the running domain, it can be seen that in general, the model appears to be underestimating the ouput variables CO, SO_2 and NO_2 over the mountainous terrain of the country, especifically the three branches of the Andes that represent the Western, Central and Eastern Coordilleras. For O_3, the trend is the opposite, with the model overestimating the values of this gas over the mountains. For the variable NO2, the model appears to overestimate the values over the Amazon basin, while for Ozone the model has a FB close to zero over the Amazon, but it appears to overestimate the values of the Altillanura biogeographic region in the Eastern part of Colombia. The model underestimates systemically the values for SO_2.

Two areas are notorious for their large MSE values in the SO_2 estimates. They are seen as large circles over the Central Coordillera. Curiously enough, these to large areas of error appear to be located over the two major volcanic foci of the country. At the same time, included within these areas of error, given the current spacial resolution of the model, are the major Colombian cities of Bogota, Medellin and Cali. It will be important to account for the different sources of emissions not yet incorporated into the model.

The apparent discrepancies observed when comparing the four statistics for a common geographical location and chemical species indicate the need for a more nuanced analysis of the dynamics. It will be necessary to evaluate the behavior of the dynamics at selected points through time to understand the weight that extreme values may be excerting over the cummulative statistics.

CONCLUSION

The preliminary results for the LE model over the TAD are encouraging. A more elaborate, time-aware evaluation scheme is needed to assess the current strengths and weaknesses of the model for the TAD domain.

REFERENCES

Chang, J. C., & Hanna, S. R. (2004). Air quality model performance evaluation. *Meteorology and Atmospheric Physics*, *87*(1–3), 167–196. https://doi.org/10.1007/s00703-003-0070-7

Chen, G., Li, J., Ying, Q., Sherman, S., Perkins, N., Rajeshwari, S., & Mendola, P. (2014). Evaluation of observation-fused regional air quality model results for population air pollution exposure estimation. *The Science of the Total Environment*, *485–486*(979), 563–74. https://doi.org/10.1016/j.scitotenv.2014.03.107

Cox, W. M., & Tikvart, J. A. (1990). A statistical procedure for determining the best performing air quality simulation model. *Atmospheric Environment Part A, General Topics*, *24*(9), 2387–2395. https://doi.org/10.1016/0960-1686(90)90331-G

Fontes, T., & Barros, N. (2010). Interpolation of Air Quality Monitoring Data in an Urban Sensitive Area : the Oporto / Asprela Case. *Revista da Faculdade de Ciência et Tecnologia, 18*(7), 6–18.

Hanha, S. R. (1988). Air Quality Model Evaluation and Uncertainty. *Japca*, *38*(4), 406–412. https://doi.org/10.1080/08940630.1988.10466390

Kumar, J. D., Sabesan, M., Das, A., Vinithkumar, N. V., & Kirubagarn, R. (2011). Evaluation of [missing information]