Effect of Global Atmospheric Datasets in Modeling Meteorology and Air Quality in the Andean Region of Ecuador

Rene Parra*

Instituto de Simulación Computacional (ISC-USFQ), Colegio de Ciencias e Ingeniería, Universidad San Francisco de Quito, Ecuador

* Corresponding author: rrparra@usfq.edu.ec

Abstract

Several components such as initial (IC) and boundary conditions (BC), generated from global atmospheric datasets for studies at the regional and mesoscale levels, play a role in numerical modeling. The analysis datasets include operationally available observations during the time of global meteorology forecast whereas reanalysis products, though available later, include more observations. Thus, it is expected that the IC and BC generated by reanalysis will be of higher quality, thereby improving the modeling results. To generate the IC and BC of a domain with a high spatial resolution (1 km) that covers the city of Cuenca, an urban area located in the Andean region of Ecuador, we employed two analysis (GFS, FNL) and three reanalysis (NCEPR2, ERA-Interim, ERA5) products. We used the Eulerian Weather Research & Forecasting model with Chemistry (WRF-ChemV3.2) to simulate meteorological and air quality variables. FNL and GFS were best fit for modeling both meteorological and air quality variables. Likewise, it also suggests their use in generating the IC and BC for modeling purposes in the Andean region of Ecuador. Currently, a few observations from the equatorial Andean region are incorporated into global atmospheric datasets. Furthermore, the atmospheric processes in this region are particularly complex and have been less studied. These limitations, which appear to be more prevalent for reanalysis products, have an impact on the generation, and thus the quality of the information stored by the atmospheric datasets for the equatorial Andean zone. Reanalysis products are not always the best choice for modeling in this area.

Keywords: Initial conditions, Boundary conditions, WRF-Chem, Reanalysis, Cuenca
1. INTRODUCTION

Atmospheric modeling involves a variety of components. For mesoscale and regional studies, numerical models work with resolutions of a few km, requiring the generation of initial (IC) and boundary conditions (BC) from global atmospheric datasets.

The Global Forecast System (GFS), produced by the National Centers for Environmental Prediction (NCEP) (GFS, 2021), is one of these datasets. It is generated by assimilation of atmospheric observations available at the time of generating the IC for global weather forecasting using the GFS model.

Another NCEP product is the FNL Global Operational Analysis (Final) (Commerce, 2000), which is generated using similar weather forecast model and assimilation scheme as the GFS. Despite FNL having a delay of 60 to 90 min, it incorporates approximately 10% more observations than GFS records.

When compared to analysis products, retrospective analysis or reanalysis atmospheric datasets are generated later, thus ingesting an additional amount of records that were not operationally available at the time of generating the meteorological forecast. As a result, reanalysis datasets incorporate observations and cover previous time periods (Parker, 2016). The National Centers for Environmental Prediction (Kanamitsu et al., 2002) generate the atmospheric NCEP Reanalysis 2 (NCEPR2), which include more meteorological data than the GFS and FNL.
NCEPR2 is an updated version that corrects human errors in the first NCEP reanalysis version. As a result, NCEPR2 serves as a first-generation atmospheric reanalysis basis.

ERA-Interim is an atmospheric reanalysis database produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) (Berrisford et al., 2011; Dee et al., 2011), with features such as satellite radiance assimilation, the use of 4-dimensional variational analysis, improved moisture analysis, variational bias correction for satellite data, and data management improvements. ERA-Interim covers the period from January 1979 to August 2019 and has been replaced by ERA5, a fifth-generation reanalysis dataset generated by the ECMWF (ERA5, 2021). ERA5 incorporates a detailed record of the global atmosphere, land surface, and ocean waves (Hersbach et al., 2020), and it has a finer spatio-temporal resolution and a better representation of the troposphere, tropical cyclones, precipitation and evaporation, precipitation on tropical soil, soil moisture and sea surface temperature than ERA-Interim (ECMWF, 2021). NCEPR2, ERA-Interim, and ERA5, are part of the world’s most important atmospheric databases (OMM, 2018) and are used in meteorology, climate, and other related fields (e.g. Doddy Clarke et al., 2021; Tarek et al., 2020; Bhattacharya et al., 2020; Zandler et al., 2019; Moalafhi et al., 2017).

From the global atmospheric datasets, information for variables such as temperature, vectorial components of wind, relative humidity, surface pressure, sea level pressure, soil moisture, and skin temperature are taken for the generation of IC and BC.
Some studies assessed the usefulness of both analysis and reanalysis datasets. When modeling volcanic ash dispersion in South America's southern cone, GFS outperformed ERA-Interim (Mulena et al., 2016). Compared to ERA5 and ERA-Interim, GFS had a better capture of the evolution of tropical cyclones (Malakar et al., 2020). However, ERA5 showed better performance than FNL in temperature modeling in South Korea, although FNL was more accurate in modeling wind speed (Mun et al., 2020). In most of North America, ERA5 performed similarly to the direct use of observations in hydrological modeling (Tarek et al., 2020).

In Ecuador, since 2016 we have generated the forecast of ash dispersion caused by Vulcanian eruptions in Tungurahua, one of the country's active volcanoes (Parra et al., 2016; Parra, 2018a; Parra et al., 2020a). With this goal in mind, our prediction system uses GFS to obtain the IC and BC for nested domains and the Eulerian Weather Research and Forecasting (WRF) model to generate the meteorological forecast in Ecuador. Once the meteorological forecast is completed, certain meteorological variables are fed into FALL3D, an Eulerian model that simulates the dispersion of volcanic ash, based on assumptions about the start, duration, height of the emission, and volcanic ash properties.

Recent findings support the use of an online approach for atmospheric simulation, including meteorology and air quality feedbacks (Baklanov et al., 2014; Baklanov and Zhang, 2020). With this in mind, we began to assess the impact of the components that play a role in the simulation of
both meteorology and air quality in Ecuadorian cities. We use FNL to generate the IC and BC to evaluate the influence of the planetary boundary layer schemes in Cuenca (Parra, 2018b), an Andean city located in Southern Ecuador (2500 masl, Fig. 1).

FNL was also used to simulate the impact on Cuenca's air quality caused by the replacement of diesel buses with electric buses (Parra and Espinoza, 2020b). This scenario is considered in the current national regulation.

1.1 Emission Inventory of Cuenca

Cuenca has a population of approximately 640,000 people. Its terrain is complex, with altitudes ranging from 1000 to 4000 masl (Fig. 1). The most recent emission inventory was conducted in 2014 (EMOV EP, 2016), and the results showed that on-road traffic was the primary source of most atmospheric pollutants (94.9% carbon monoxide (CO), 71.2% nitrogen oxides (NOx), 42.4% fine particulate matter (PM$_{2.5}$), 39.6% of non-methane volatile organic compounds (NMVOC)). Other relevant sources of NMVOC were the use of solvents (29.7%), and the vegetation (19.5%). The industrial sector was the most significant source (60.1%) of sulfur dioxide (SO$_2$). A thermoelectric plant located NE of the urban area generated 35.1% of SO$_2$ and 18.5% of NO$_x$ emissions. Furthermore, it was reported that several artisan brick producers located NW of the urban area generated 38.5% of total PM$_{2.5}$ emissions. The spatial distribution of annual PM$_{2.5}$ emissions is depicted in Figure 1.
Emission inventories are typically the components with the highest level of uncertainty among the elements that participate in air quality simulation. This feature is especially important in all of Ecuador's emissions inventories prepared to date (MAE, 2014a; MAE, 2014b). The high levels of uncertainty are primarily due to a lack of national emission factors, which have necessitated the use of emission factors from international literature. In the case of Cuenca, the quality of the corresponding emission maps can be tested in an atmospheric model to reproduce both the observed meteorology and air quality in order to develop a new inventory with less uncertainty.

To accomplish this, the best atmospheric dataset for generating the IC and BC must be identified.

1.2 Air Quality Network from Cuenca

To monitor meteorology and air quality in 2014, the city had an automatic station in the historic center (MUN station, Fig. 1). In addition, the monitoring network has approximately 20 passive stations that measure the monthly mean concentrations of NO₂ and O₃. The availability of a unique automatic station in 2014 for monitoring the air quality in the urban area of Cuenca is consistent with the minimum requirement established by the European Directive about air quality and cleaner air for Europe (Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe, 2008), which indicate to assess compliance the PM₂.₅ exposure, it is required at least one sampling point per million inhabitants. After 2014, the air quality network improved, and currently, it has six automatic stations for measuring PM₂.₅ and PM₁ (Parra et al., 2022). The municipality of Cuenca operates
the air quality network, which has been accredited by the competent national authority, in accordance with the methods established in Ecuadorian regulations.

Between 2008 and 2019, mean annual NO$_2$ concentrations higher than the World Health Organization (WHO) 2006 guideline (40 µg m$^{-3}$) (WHO, 2006) were measured in BCB and VEG (Fig. 1), two passive microscale stations located in street canyons with high traffic of gasoline and diesel vehicles (mainly buses) (Parra and Espinoza, 2020b).

PM$_{2.5}$ concentrations (mean value of 24 h) were higher than the WHO guideline of 2006 (25 µg m$^{-3}$) for fourteen days between 2012 and 2019. In four years of this period, the PM$_{2.5}$ annual mean concentration was higher than the WHO guideline value (10 µg m$^{-3}$) proposed in 2006.

O$_3$ concentrations may be higher than the WHO guideline value (100 g m$^{-3}$, maximum average of 8 h) on specific days in September, due in part to high levels of solar radiation arriving perpendicular to the Equator during this month (Parra and Espinoza, 2020b), which promotes the photochemical production of O$_3$.

The WHO recently updated its guidelines for air quality, recommending stricter concentrations of particulate matter, O$_3$, NO$_2$, SO$_2$, and CO (WHO, 2021). In 2020 in Cuenca, the annual mean concentrations of NO$_2$ (19.7 µg m$^{-3}$) and PM$_{2.5}$ (8.5 µg m$^{-3}$) were higher than the new recommended values (10.0 µg m$^{-3}$ for NO$_2$, and 5.0 µg m$^{-3}$ for PM$_{2.5}$). Furthermore, higher daily mean concentrations of NO$_2$ (25.0 µg m$^{-3}$) and PM$_{2.5}$ (15.0 µg m$^{-3}$) were recorded over 91 and 49
days, respectively, at the new guide values. The records highlight the importance of better understanding the quality problems, improving the emissions inventory, and defining a suitable approach for modeling meteorology and air quality.

1.3 Objectives

Although the results of the simulations performed thus far have been promising, it is necessary to assess the influence of other components in order to select those that contribute to improving the modeling of meteorology and air quality in Cuenca. Given that atmospheric reanalysis datasets contain more observations than analysis products, the IC and BC generated by reanalysis are expected to improve modeled results. Thus, the primary goals of the contribution are meant to address the following questions:

• Are reanalysis datasets better to model meteorology and air quality in the Andean region of Ecuador?

• What atmospheric dataset provides the best IC and BC to model meteorology and air quality in this region?

• What is the recommended atmospheric dataset for atmospheric modeling as a quality indicator activity for a new emissions inventory?

2 METHODS

2.1 Modeling approach
Using the Eulerian Weather Research & Forecasting with Chemistry model (WRF-ChemV3.2), we generated the IC and BC for modeling meteorological and air quality variables in Cuenca using GFS, FNL, NCEPR2, ERA-Interim, and ERA5. WRF is a non-hydrostatic 3-D model widely used for meteorological research and forecasting (WRF, 2021), which allows modeling of chemical transport of air pollutants to be included at the same time. A master domain and three nested subdomains (Fig. 1) were used to run the meteorological simulations. The nested grid ratio between domains was 3 to 1. The internal subdomain is a grid of 100 x 82 cells, 1 km on each side, that covers the territory of Cuenca (Fig. 1c). We used 35 vertical levels, up to the top pressure of 50 hPa (about 20 km altitude), being comparable with the vertical resolution of ERA-Interim and ERA5 datasets (38 vertical levels) and more detailed than GFS (27), FNL (27), and NCEP2 (18) vertical resolution. The option of chemical transport of pollutants was activated for the internal subdomain, allowing the reading of hourly emissions previously elaborated from the 2014 emission inventory. Based on previous modeling studies in Cuenca (EMOV EP, 2016; Parra, 2018b), BC for chemical species were assumed, using 85% of the default concentrations from WRF-Chem V3.2. Due to the O₃ levels, the study period run from 4 to 27 September 2014. Furthermore, the activity of on-road traffic and other emission sources is typical of other months. On-road traffic emissions for September were estimated from the total yearly emissions through the fuel consumption ratio during this month to the total annual amount. The daily traffic
emissions were estimated based on the traffic intensities during weekdays and weekends. Hourly on-road traffic emissions were defined from typical daily profiles provided by the Municipality of Cuenca. NMVOC emissions from vegetation were estimated from modeled maps of temperature and solar radiation for each day of the study period. The Carbon Bond Mechanism - Z (CBMZ) (Zaveri and Peters, 1999) and the Model for Simulation Aerosol Chemistry and Interaction (MOSAIC) were used to speciate hourly emissions (Zaveri et al., 2008). The main characteristics of the atmospheric bases used in this study are shown in Table S1. Table S2 shows the main parameters and options chosen for the WRF-Chem model. The option to use the online approach, which considers the direct effects of aerosols on meteorology, was activated.

We made the simulations for three-day periods, using six h as spin-up time for each one, and activating the option for using the default WRF-ChemV3.2’s idealized profile to initialize chemistry.

The study covers 24 days, with a proportion of 80% of the monthly period, which is consistent with the required proportion (at least 75%) when aggregating data and calculating statistical parameters, according to the European Directive about ambient air quality and cleaner air for Europe (Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe, 2008).

2.2 Modeling performance
The meteorology model's performance was defined using the MUN station's hourly records of temperature, wind speed, and wind direction. For temperature and wind speed, we use the following metrics:

\[ GE = \frac{1}{N} \sum_{i=1}^{N} |P_i - O_i| \]  
\[ MB = \frac{1}{N} \sum_{i=1}^{N} (P_i - O_i) \]  
\[ IOA = \frac{\frac{1}{N} \sum_{i=1}^{N} (P_i - O_i)^2}{\frac{1}{N} \sum_{i=1}^{N} (P_i - P_m + O_i - O_m)^2} \]  
\[ RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (P_i - O_i)^2} \]  

where GE is the gross error, MB is the mean bias, IOA is index of agreement, RMSE is the root mean square error, \( N \) is the number of values, \( P_m \) is the mean modeled value, \( O_m \) is the mean observed value, \( P_i \) is the modeled value, and \( O_i \) is the record. Wind direction was assessed through GE and MB. Wind direction also was evaluated by comparing the modeled wind roses to the one obtained from the records.

To evaluate the performance of modeling short-term air quality, the CO (maximum mean in 1 h and 8 h per day), PM\(_{2.5}\) (mean in 24 h), and O\(_3\) (maximum mean in 8 h per day) records were used. These periods are consistent with national air quality legislation and the WHO guidelines (WHO, 2006). Based on the recommendations of Simon (Simon et al., 2012), we present as metrics the MB, RMSE, the fractional bias (FB), the mean normalized bias (MNB), and the correlation coefficient (r). The following equations define the last three indicators:
Furthermore, we calculated the percentages of meteorological and air quality records captured by modeling. We consider that a record was “positive modeled or captured” if the maximum deviation between the observed and modeled value agreed with the accuracy proposed by the European modeling quality Directive (EEA, 2011) (Table S3). A percentage of 100% means that all records were captured. This approach was applied in previous studies to assess the modeling performance in Cuenca (Parra, 2021, 2018b).

For long-term air quality records (monthly means of NO2 and O3), modeling performance was defined as the percentage of passive stations with a maximum difference of 30%. The percentage of 100% indicates that the monthly mean concentrations in all passive stations that measure each pollutant were properly modeled.

3 RESULTS

3.1 Meteorology

The metrics indicate that GFS, FNL, and NCEPR2 provided the best fit to temperature measurements. As shown in Table 1, the GE indicator values were in the reference range for all
datasets (reference: <2 °C, minimum: 1.3 °C for FNL, maximum: 1.8 °C for ERA5), whereas the MB indicator was within the ± 0.5 °C boundaries for all databases, except for ERA5 and ERA-Interim. Although, on average, ERA-Interim overestimated (MB = 1.0 °C) and ERA5 underestimated (MB = -0.7 °C) the temperature, a strong relationship between measured and modeled values (IOA = 0.9) was established for all datasets.

When FNL, GFS, ERA-Interim, NCEPR2, and ERA5 were applied, 77.3%, 75.7%, 74.1%, 72.4%, and 64.3% of the hourly temperature records were captured, respectively (Table 3). In addition, the FNL produced the highest proportion of modeled values within the range of variation established for temperature (Fig. 2b), as well as matched the recorded daily temperature profile most closely (Fig. 2f). Although a higher MB was obtained for GFS (0.4 °C) relative to FNL (0.2 °C), the former model produced output that was most consistent with records. On the other hand, NCEPR2 tended to underestimate and overestimate temperatures during the early morning hours and at noon, respectively. As can be seen from Figure 3, for 25 September 2014 (13:00 LT), the GFS and FNL modeled temperatures between 18 °C and 21 °C (shown in violet color) throughout the urban area, whereas values between 21 °C and 24 °C (red color) were computed using the reanalysis datasets.

According to the obtained metrics, all atmospheric datasets yielded acceptable modeled wind speed values. Even though all RMSE, MB, and IOA values are within their respective reference
ranges, as can be seen from Table 1, the FNL produced the best fit to the data (RMSE = 0.9°C, 
MB = 0.2°C, IOA = 0.9), as it captured 77.6% of the recorded values (Table 3). In contrast, 
ERA5 captured 63.8% of the records. For all databases, RMSE (reference: <2 m s⁻¹, minimum: 
0.9 m s⁻¹ for FNL, maximum: 1.3 m s⁻¹ for ERA5) and MB (reference: ± 0.5 m s⁻¹, minimum: 
0.2 m s⁻¹ for FNL, maximum: 0.3 m s⁻¹ for the remaining datasets) were within the reference 
range. A strong linear relationship between the records and modeled wind speed values was 
obtained in all cases (IOA of 0.9 and 0.8 calculated for FNL and other models, respectively).

According to the records, winds were primarily coming from the east-northeast (ENE) 
direction (Fig. 4a). However, GFS, FNL, NCEPR2, and ERA5 indicated east (E), while only the 
output produced by ERA-Interim unmatched the recorded pattern (Fig. 4e). Although FNL (64.0 
°) and GFS (65.0 °) were better in terms of GE, none of the datasets reached values for wind 
direction into the benchmark range (< 30°). NCEPR2 (-0.8 °) and ERA-Interim (-4.4 °) got MB 
values into the benchmark ranges. However, the ERA-Interim MB value resulted from significant 
negative and positive differences between wind speed records and computed values, as shown by 
GE (90.1 °).

The metrics indicated that GFS, FNL, and NCEPR2 provided the best fit to model temperature, 
wind speed, and wind direction.

3.2 Air quality
The GFS, FNL and ERA-Interim produced the best fit to model the maximum CO daily 1-h mean concentration. It is important to note that GFS and ERA-Interim did not overestimate or underestimate (MB = 0.0 mg m$^{-3}$, Table 2) whereas NCEPR2 (MB = 1.0 mg m$^{-3}$) and ERA5 (MB = 1.1 mg m$^{-3}$) overestimated these concentrations. The magnitude of difference between the modeled and observed values were smaller for GFS, FNL, and ERA-Interim (RMSE = 0.6 mg m$^{-3}$) compared to NCEPR2 (1.2 mg m$^{-3}$) and ERA5 (1.5 mg m$^{-3}$). However, there was no strong linear relationship between the recorded and modeled values (r = 0.4 for GFS, FNL, and ERA5).

Similarly, GFS, FNL, and ERA-Interim produced the best fit for modeling the maximum CO 8-h mean concentration (Fig. 5). On the other hand, as seen from Table 2, the datasets showed a slight underestimation (MB = -0.1 mg m$^{-3}$) in the concentrations. The magnitude of difference between the modeled and observed values were lower for GFS and FNL (RMSE = 0.2 mg m$^{-3}$). It is also worthy of note that there was no strong linear relationship between the recorded and modeled values (r = 0.4) for FNL. For these concentrations, FNL (100.0%) and GFS (95.8%) had the highest capture percentages (Table 3).

The best options for modeling daily mean PM$_{2.5}$ concentrations were GFS and ERA-Interim (Table 3, Fig. 6). On the other hand, ERA5 had the lowest performance. However, there was a
weak linear correlation between the recorded and modeled values for the GFS ($r = 0.2$) and ERA-Interim ($r = 0.3$). The GFS and ERA-Interim produced the most captured records (75.0%) to model PM$_{2.5}$ concentrations. ERA5 produced the least capture percentage (45.8%).

The majority of the databases had a capture percentage above 80.0% of the maximum O$_3$ daily 8-h mean concentrations (Table 3, Fig. 7). ERA5 had the best metrics to model these concentrations. All of the datasets provided mean O$_3$ profiles that matched the behavior of the records during the early hours of the morning, though the modeled values were higher after noon and in the afternoon (Fig. 7f). Figure S1 depicts the modeled maps of 25 September 2014 for all the atmospheric datasets. They were consistent, despite minor differences. The area adjacent to the NW of the urban area presented concentrations between 75.0 µg m$^{-3}$ and 80.0 µg m$^{-3}$ (purple color) for GFS and FNL. However, ERA-Interim and ERA5 generated values between 80.0 µg m$^{-3}$ and 85.0 µg m$^{-3}$ (red color). Despite the high percentages of records captured, the linear relationship between recorded and modeled values was low ($r$ between -0.1 and 0.2, Table 3, Fig. 7).

In terms of NO$_2$ monthly mean concentrations, GFS, FNL, and ERA-Interim captured 100.0% records from all passive stations that measure the pollutant while 73.3% was captured by ERA5 (Table 3, Fig. S2a). However, ERA5 captured the O$_3$ monthly mean concentrations in 87.5% of the stations (Table 3, Fig. S2b).
FNL and GFS were the datasets best fit for modeling meteorology and air quality, thus suggesting their ability to generate IC and BC for atmospheric modeling purposes in the study region.

4 DISCUSSION AND CONCLUSIONS

We used a state-of-the-art 3-D numerical model to simulate meteorology and air quality, using atmospheric analysis and reanalysis datasets to define the IC and BC of a nested domain with a high spatial resolution (1 km) that covers the territory of an Andean equatorial city. The modeling approach examined the direct effects of aerosols on meteorology, based on the benefits of using an online approach reported in the literature. In this context, we begin with the premise that the best database should accurately model both meteorology and air quality.

The metrics indicated that the model's performance was better when the analysis datasets (FNL and GFS) were used. The lack of improvement, and more specifically, the decrease in the model's performance when using reanalysis products (NCEPR2, ERA-Interim, and ERA5), was unexpected due to the high quality of these datasets since they were made with more observations. FNL and GFS performed best in terms of modeling surface temperature, which is an important meteorological parameter that is strongly related to the height of the planetary boundary layer (PBL) and atmospheric stability. These databases provided minimum mean temperatures during the early morning hours close to the observations (Fig. 2f), thus suggesting that the corresponding
heights of the PBL and atmospheric stability were also adequately modeled. The peak of vehicular traffic emissions in Cuenca typically occurs between 07:00 and 08:00 LT, due to stable atmospheric conditions that prevent dispersion and produce higher CO concentrations around these hours. This relationship is consistent with the best results and metric values for modeling CO concentrations in the analysis datasets. With lower PBL modeled heights (not indicated) and higher CO concentrations, NCEPR2 and ERA5 produced lower minimum temperature values than the records. This interaction explains why these datasets overestimated CO concentrations.

In the case of wind speed, despite that all datasets obtained metrics within the reference ranges, FNL performed the best and, on average, its results were more in line with the daily behavior of this parameter (not indicated).

All datasets overestimated PM$_{2.5}$ concentrations, suggesting that other components should be evaluated, such as the quality of the emissions inventory and the influence on the modeling of direct and indirect effects between aerosols and meteorology.

Although ERA5 captured the majority of the short-term and long-term O$_3$ records, the model performance with respect to the meteorological variables and other air quality variables was the lowest. On average, all datasets modeled higher hourly solar radiation levels than the records (not indicated), with ERA5 producing the most accurate results. Because solar radiation influences the photochemical production of O$_3$, the more approximate levels of solar radiation of ERA5 explains,
at least in part, why this database presented a higher performance in O$_3$ modeling. The overestimation of solar radiation highlights the importance of evaluating other schemes, such as cumulus parameterization, O$_3$ concentrations in IC and BC, and the emission inventory of O$_3$ precursors.

The results suggest that instead of reanalysis datasets, FNL or GFS should be used to generate the IC and BC for modeling meteorology and air quality in the equatorial Andean region. This finding is consistent with the preliminary decrease reported when using NCEPR2 instead of GFS or FNL to model four volcanic ash emission and dispersion events in Ecuador, which occurred in December 2012, July 2013, February 2014, and August 2015 in the Tungurahua volcano and the Cotopaxi volcano, respectively (Parra, 2018c). We emphasize that these events occurred over a period of months and years, and that the modeling done on a national scope. The current contribution and the volcanic ash dispersion study results were consistent, and it was suggested that the IC and BC be generated using GFS or FNL for other regions of Ecuador and at the national level.

Some studies found similar results in other parts of the world. Table S4 summarizes the main conclusions of selected studies on the impact of global atmospheric datasets. These conclusions, as well as the results of this contribution, suggest that reanalysis products are not always the best default option. Global atmospheric datasets require dedicated evaluations to determine their use.
Although NCEPR2 includes a higher number of observations compared to GFS and FNL, this reanalysis product is made with lower horizontal (2.5°) and vertical (18 vertical layers) resolution (Table S1). The lower spatial resolution of NCEPR2 could partly explain why this reanalysis obtained a lower performance than GFS (0.5°, 27 vertical layers) or FNL (1.0°, 27 vertical layers) (Wagner et al., 2018). However, when using ERA5, a fifth-generation atmospheric dataset and one of the most advanced reanalysis products with high horizontal (0.25°) and vertical (38 vertical layers) resolution, overall modeling performance decreased.

Although the study of the causes of the decline in the performance of the reanalysis datasets is beyond the scope of this manuscript, we argue that one reason corresponds to the complex conditions imposed by the Andes Mountains. The limitations of the reanalysis datasets are evident in areas with complex orography, particularly in mountainous regions and in areas where assimilation and processing schemes cannot reproduce actual atmospheric processes (OMM, 2018).

Although the national meteorological and hydrological services of the Andean countries provide a large amount of surface data, some of these observations do not meet the World Meteorological Organization's requirements (WMO, 2020). In addition, national databases are partially shared with international entities in charge of generating global atmospheric databases. On the other hand, high altitude areas are poorly instrumented (Condom et al., 2020; Kull et al.,
2021), and the equatorial Andean region has poor coverage of radiosonde observations (Durre et al., 2018; Cazorla et al., 2021). These limitations imply an incomplete description of the atmospheric conditions in this region, which has an impact on the assimilation processes used to prepare global atmospheric datasets.

Aside from the few observations from the equatorial Andean region, the reduction in modeling quality when using reanalysis datasets, particularly ERA5, was unexpected. Increased records from this region are required to prepare global atmospheric datasets and improve understanding of atmospheric processes in the equatorial Andean region. The atmosphere in this region exhibits complex behavior, with unique characteristics that combine a complex topography, the influence of the Intertropical Convergence Zone, the presence of breezes from the Pacific Ocean and the Amazon region, which promote vertical convective movements, and consequently, the persistence of clouds. The atmospheric models include parameterizations and schemes that were developed and tested primarily for the northern hemisphere. These constraints are currently affecting the development of atmospheric databases. These limitations, which appear to be more prevalent for reanalysis products, have an impact on the generation, and therefore the quality of the information stored by the atmospheric datasets for the equatorial Andean zone. Reanalysis products are not always the best choice for modeling in this area.
Initial and boundary conditions are essential components for mesoscale and regional atmospheric modeling. This contribution provided information for selecting the best atmospheric datasets to model meteorology and air quality in the Andean region of Ecuador. Other parameters and components, such as the parameterization of the earth's surface and cumulus clouds, the effect of assimilation of meteorological and air quality data, the influence of spin-up time, and the indirect effects between meteorology and aerosols, must be evaluated. Complementary studies based on updated and improved emission inventories, with coverage of other months and the use of other chemical mechanisms and aerosol models, will be required in the future.

Within the surface layer, the bottom 20 to 200 m of the planetary boundary layer, the frictional drag, heat conduction, and evaporation from the surface produce significant changes in wind speed and temperature with height, defining the land-atmosphere interaction. The surface layer and components as the planetary boundary layer and land surface are typically parameterized in atmospheric modeling (Table S2). Although this study focused on the influence of IC and BC, we highlight the need to assess the effect of parameterized variables in the future.

ACKNOWLEDGMENTS
This research is part of the project “Emisiones atmosféricas y Calidad del Aire en el Ecuador 2022–2023”. Simulations were done at the High-Performance Computing system at the Universidad San Francisco de Quito.

REFERENCES


Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U.,
Balmaseda, M.A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A.C.M., Berg, L. van de,
Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Ger, A.J., Haimberger, L.,
et al. (2011). The ERA-Interim reanalysis: configuration and performance of the data

air quality and cleaner air for Europe (2008), OJ L.

Reanalysis Dataset Should We Use for Renewable Energy Analysis in Ireland?

Atmosphere 12, 624. https://doi.org/10.3390/atmos12050624

Durre, I., Yin, X., Vose, R.S., Applequist, S., Arnfield, J. (2018). Enhancing the Data Coverage
1770. https://doi.org/10.1175/JTECH-D-17-0223.1

ECMWF. (2021). What are the changes from ERA-Interim to ERA5? - Copernicus Knowledge
Base - ECMWF. URL
https://confluence.ecmwf.int/pages/viewpage.action?pageId=74764925 (accessed 12 July
2021).


ERA5. (2021). Hourly data on pressure levels from 1979 to present. URL
https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-pressure-
levels?tab=overview (accessed 12 July 2021).

EEA. (2011). The application of models under the European Union’s Air Quality Directive : a
technical reference guide. Publications Office, LU.


Mon. Weather Rev. 121, 764–787. https://doi.org/10.1175/1520-
0493(1993)121<0764:PEOAUB>2.0.CO;2

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J.,
Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X.,

https://doi.org/10.1002/qj.3803

for the Bulk Parameterization of Clouds and Precipitation. Mon. Weather Rev. 132, 103–


MAE. (2014b). Inventario de emisiones en las ciudades de Loja, Azogues, Babahoyo y Quevedo. Ministerio del Ambiente, Quito-Ecuador


27
Table 1. Metrics for meteorological modeling. Green background color indicates metrics with values into the benchmark ranges.

<table>
<thead>
<tr>
<th>Global Meteorological Dataset:</th>
<th>1 GFS</th>
<th>2 FNL</th>
<th>3 NCEPR2</th>
<th>4 ERA-Interim</th>
<th>5 ERA5</th>
<th>Benchmark</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Surface temperature:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GE (°C)</td>
<td>1.4</td>
<td>1.3</td>
<td>1.4</td>
<td>1.5</td>
<td>1.8</td>
<td>&lt; 2 °C</td>
</tr>
<tr>
<td>MB (°)</td>
<td>0.4</td>
<td>0.2</td>
<td>-0.1</td>
<td>1.0</td>
<td>-0.7</td>
<td>&lt; ±0.5 °C</td>
</tr>
<tr>
<td>IOA</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>≥ 0.8</td>
</tr>
<tr>
<td><strong>Wind speed at 10 m:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMSE (m s⁻¹)</td>
<td>1.1</td>
<td>0.9</td>
<td>1.1</td>
<td>1.0</td>
<td>1.3</td>
<td>&lt; 2 m s⁻¹</td>
</tr>
<tr>
<td>MB (m s⁻¹)</td>
<td>0.3</td>
<td>0.2</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>&lt; ± 0.5 m s⁻¹</td>
</tr>
<tr>
<td>IOA</td>
<td>0.8</td>
<td>0.9</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>≥ 0.6</td>
</tr>
<tr>
<td><strong>Wind direction at 10 m:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GE (°)</td>
<td>65.0</td>
<td>64.0</td>
<td>71.1</td>
<td>90.1</td>
<td>80.4</td>
<td>&lt; 30 °</td>
</tr>
<tr>
<td>MB (°)</td>
<td>-24.1</td>
<td>-22.4</td>
<td>-0.8</td>
<td>-4.4</td>
<td>16.5</td>
<td>&lt; ± 10 °</td>
</tr>
</tbody>
</table>
Table 2. Metrics for air quality modeling. Green background color indicates metrics with the best performance values.

<table>
<thead>
<tr>
<th>Global Meteorological Dataset:</th>
<th>1 GFS</th>
<th>2 FNL</th>
<th>3 NCEPR2</th>
<th>4 ERA-Interim</th>
<th>5 ERA5</th>
<th>Ideal value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CO, daily maximum 1-h mean:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MB (mg m(^{-3}))</td>
<td>0.0</td>
<td>0.1</td>
<td>1.0</td>
<td>0.0</td>
<td>1.1</td>
<td>0.0</td>
</tr>
<tr>
<td>RMSE (mg m(^{-3}))</td>
<td>0.6</td>
<td>0.6</td>
<td>1.2</td>
<td>0.6</td>
<td>1.5</td>
<td>0.0</td>
</tr>
<tr>
<td>FB (%)</td>
<td>1.8</td>
<td>3.5</td>
<td>46.5</td>
<td>-1.3</td>
<td>50.6</td>
<td>0.0</td>
</tr>
<tr>
<td>MNB (%)</td>
<td>4.2</td>
<td>5.9</td>
<td>66.9</td>
<td>0.8</td>
<td>78.6</td>
<td>0.0</td>
</tr>
<tr>
<td>r</td>
<td>0.4</td>
<td>0.4</td>
<td>0.3</td>
<td>0.4</td>
<td>0.2</td>
<td>1.0</td>
</tr>
<tr>
<td>CO, daily maximum 8-h mean:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MB (mg m(^{-3}))</td>
<td>-0.1</td>
<td>-0.1</td>
<td>0.3</td>
<td>-0.1</td>
<td>0.4</td>
<td>0.0</td>
</tr>
<tr>
<td>RMSE (mg m(^{-3}))</td>
<td>0.2</td>
<td>0.2</td>
<td>0.4</td>
<td>0.3</td>
<td>0.6</td>
<td>0.0</td>
</tr>
<tr>
<td>FB (%)</td>
<td>-15.3</td>
<td>-12.6</td>
<td>29.8</td>
<td>-14.9</td>
<td>35.6</td>
<td>0.0</td>
</tr>
<tr>
<td>MNB (%)</td>
<td>-13.3</td>
<td>-11.3</td>
<td>37.4</td>
<td>-13.0</td>
<td>45.3</td>
<td>0.0</td>
</tr>
<tr>
<td>r</td>
<td>0.3</td>
<td>0.4</td>
<td>0.2</td>
<td>0.3</td>
<td>0.2</td>
<td>1.0</td>
</tr>
<tr>
<td>PM(_{2.5}), daily 24-h mean:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MB (µg m(^{-3}))</td>
<td>0.7</td>
<td>1.1</td>
<td>2.2</td>
<td>0.4</td>
<td>3.2</td>
<td>0.0</td>
</tr>
<tr>
<td>RMSE (µg m(^{-3}))</td>
<td>3.0</td>
<td>3.4</td>
<td>4.1</td>
<td>2.8</td>
<td>5.5</td>
<td>0.0</td>
</tr>
<tr>
<td>FB (%)</td>
<td>10.9</td>
<td>15.1</td>
<td>29.3</td>
<td>5.7</td>
<td>39.6</td>
<td>0.0</td>
</tr>
<tr>
<td>MNB (%)</td>
<td>32.8</td>
<td>46.1</td>
<td>67.6</td>
<td>25.3</td>
<td>96.7</td>
<td>0.0</td>
</tr>
<tr>
<td>r</td>
<td>0.2</td>
<td>-0.1</td>
<td>-0.1</td>
<td>0.3</td>
<td>-0.4</td>
<td>1.0</td>
</tr>
<tr>
<td>O(_3), daily maximum 8-h mean:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MB (µg m(^{-3}))</td>
<td>16.3</td>
<td>14.5</td>
<td>11.5</td>
<td>14.5</td>
<td>7.9</td>
<td>0.0</td>
</tr>
<tr>
<td>RMSE (µg m(^{-3}))</td>
<td>19.6</td>
<td>18.1</td>
<td>16.1</td>
<td>18.9</td>
<td>16.0</td>
<td>0.0</td>
</tr>
<tr>
<td>FB (%)</td>
<td>24.4</td>
<td>22.0</td>
<td>17.9</td>
<td>22.0</td>
<td>12.6</td>
<td>0.0</td>
</tr>
<tr>
<td>MNB (%)</td>
<td>32.6</td>
<td>29.3</td>
<td>24.3</td>
<td>29.8</td>
<td>-17.1</td>
<td>0.0</td>
</tr>
<tr>
<td>r</td>
<td>0.1</td>
<td>0.2</td>
<td>0.0</td>
<td>-0.1</td>
<td>0.2</td>
<td>1.0</td>
</tr>
</tbody>
</table>
Table 3. Percentages of captured records by modeling.

<table>
<thead>
<tr>
<th>Global Meteorological Dataset:</th>
<th>1 GFS</th>
<th>2 FNL</th>
<th>3 NCEPR2</th>
<th>4 ERA-Interim</th>
<th>5 ERA5</th>
<th>Number of records</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meteorology:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Surface temperature</td>
<td>75.7</td>
<td>77.3</td>
<td>72.4</td>
<td>74.1</td>
<td>64.3</td>
<td>572</td>
</tr>
<tr>
<td>Wind speed at 10 m</td>
<td>71.5</td>
<td>77.6</td>
<td>70.6</td>
<td>74.0</td>
<td>63.8</td>
<td>572</td>
</tr>
<tr>
<td>Wind direction at 10 m</td>
<td>35.8</td>
<td>35.7</td>
<td>28.1</td>
<td>18.0</td>
<td>28.1</td>
<td>572</td>
</tr>
<tr>
<td>Short-term air quality:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CO, daily maximum 1-h mean</td>
<td>83.3</td>
<td>91.7</td>
<td>45.8</td>
<td>79.2</td>
<td>33.3</td>
<td>24</td>
</tr>
<tr>
<td>CO, daily maximum 8-h mean</td>
<td>95.8</td>
<td>100.0</td>
<td>66.7</td>
<td>83.3</td>
<td>58.3</td>
<td>24</td>
</tr>
<tr>
<td>PM$_{2.5}$, daily 24-h mean</td>
<td>75.0</td>
<td>62.5</td>
<td>58.3</td>
<td>75.0</td>
<td>45.8</td>
<td>24</td>
</tr>
<tr>
<td>O$_3$, daily maximum 8-h mean</td>
<td>79.2</td>
<td>83.3</td>
<td>83.3</td>
<td>83.3</td>
<td>87.5</td>
<td>24</td>
</tr>
<tr>
<td>Long-term air quality:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NO$_2$, monthly mean</td>
<td>100.0</td>
<td>100.0</td>
<td>86.7</td>
<td>100.0</td>
<td>73.3</td>
<td>15</td>
</tr>
<tr>
<td>O$_3$, monthly mean</td>
<td>56.3</td>
<td>56.3</td>
<td>62.5</td>
<td>56.3</td>
<td>87.5</td>
<td>16</td>
</tr>
</tbody>
</table>
Fig. 1. Location of: (a) Ecuador and (b) Cuenca. (c) PM$_{2.5}$ emissions (t y$^{-1}$). (d) Urban area of Cuenca and the air quality stations (red dots). At the Historic Center MUN station (2500 masl) provides short-term air quality levels and meteorology. The others are passive stations for monitoring monthly-mean O$_3$ and NO$_2$ values. Nomenclature of stations: MUN, Municipio; EHS, Escuela Héctor Sempérgui; ECC, Escuela Carlos Crespi; CCA, Colegio Carlos Arizaga; MAN, Machángara; CEB, Cebollar; BAL, Balzay; VEG, Vega Muñoz; TET, Terminal Terrestre; EIA, Escuela Ignacio Andrade; MEA, Mercado El Arenal; BCB, Bomberos; CHT, Colegio Herlinda Toral; LAR, Calle Larga; MIS, Misticata; CRB, Colegio Rafael Borja; EIE, Escuela Ignacio Escandón; EVI, Escuela Velasco Ibarra; ODO; Facultad de Odontología; ICT, Ictocruz. Red rectangles indicated the domains for modeling: Master domain (70 x 70 cells, 27 km), first subdomain (52 x 52 cells, 9 km), second subdomain (61 x 43 cells, 3 km), internal subdomain (100 x 82 cells, 1 km).
Fig. 2. Observed versus modeled temperature (MUN station): (a) 1 GFS, (b) 2 FNL, (c) 3 NCEPR2, (d) 4 ERA-Interim, (e) 5 ERA5. (f) Mean daily profiles. Parallel lines indicate the range ±2 °C which depicts the zone of required accuracy for modeling temperature.
Fig. 3. Modeled surface wind and temperature on 25 September 2014 (13:00 LT): (a) 1 GFS, (b) 2 FNL, (c) 3 NCEPR2, (d) 4 ERA-Interim, (e) 5 ERA5.
Fig. 4. Wind roses (MUN station): (a) Observed, (b) 1 GFS, (c) 2 FNL, (d) 3 NCEPR2, (e) 4 ERA-Interim, (f) 5 ERA5.
Fig. 5. Observed versus modeled daily CO 8-h maximum mean (MUN station): (a) 1 GFS, (b) 2 FNL, (c) 3 NCEPR2, (d) 4 ERA-Interim, (e) 5 ERA5. (f) Mean daily profiles.
Fig. 6. Observed versus modeled daily PM$_{2.5}$ 24-h mean (MUN station): (a) 1 GFS, (b) 2 FNL, (c) 3 NCEPR2, (d) 4 ERA-Interim, (e) 5 ERA5. (f) Mean daily profiles.
Fig. 7. Observed versus modeled daily $\text{O}_3$ 8-h maximum mean (MUN station): (a) 1 GFS, (b) 2 FNL, (c) 3 NCEPR2, (d) 4 ERA-Interim, (e) 5 ERA5. (f) Mean daily profiles.