Distribution Agreement

Signature:

____

Approval Sheet

A Bayesian Approach to Modeling Particulate Matter Over an Italian Domain Using MAIA Ancillary Geographic Product Data.

By

Harrison Goodall
MPH

Department of Epidemiology

____
Yang Liu, PhD
Committee Chair

____
Mike Goodman, MD, MPH
Committee Member

A Bayesian Approach to Modeling Particulate Matter Over an Italian Domain Using MAIA Ancillary Geographic Product Data.

By

Harrison Goodall

Thesis Committee Chair: Yang Liu, PhD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Epidemiology
2023

Abstract

A Bayesian Approach to Modeling Particulate Matter Over an Italian Domain Using MAIA Ancillary Geographic Product Data.

By Harrison Goodall

Particulate Matter (PM) is a major cause of morbidity and mortality worldwide. This study examines the use of five models using a Bayesian Hierarchical Downscaling model structure to predict $PM_{2.5}$ (PM < 2.5 µm) across a region in central Italy in 2015. We build upon previous modeling work done in this region of Italy and provide an alternative way to create models to predict $PM_{2.5}$ using fewer spatiotemporal and spatial predictors, smaller training data sets as well as the ability to calculate uncertainty measurements. The Bayesian models used in this paper predicted $PM_{2.5}$ concentrations with a mean overall cross validation R^2 of .72. Using extinction as our main predictor (aerosol optic density (AOD) divided by planetary boundary layer (PBL)) and data from NASA's Multiple Angle Imager for Aerosol Ancillary Geographic Product (MAIA AGP) and Italian collaborators, we demonstrated that the MAIA AGP variables can be used to reliably predict $PM_{2.5}$ and generate R^2 values equivalent to those generated from models run with parameters processed by our Italian collaborators. The ability of our Bayesian model to integrate MAIA AGP variables and predict annual and daily $PM_{2.5}$ concentrations with reasonable accuracy and uncertainty measurements provides future exposure studies with important data about model uncertainty, and the ability to predict $PM_{2.5}$ across resource limited domains.

A Bayesian Approach to Modeling Particulate Matter Over an Italian Domain Using MAIA Ancillary Geographic Product Data.

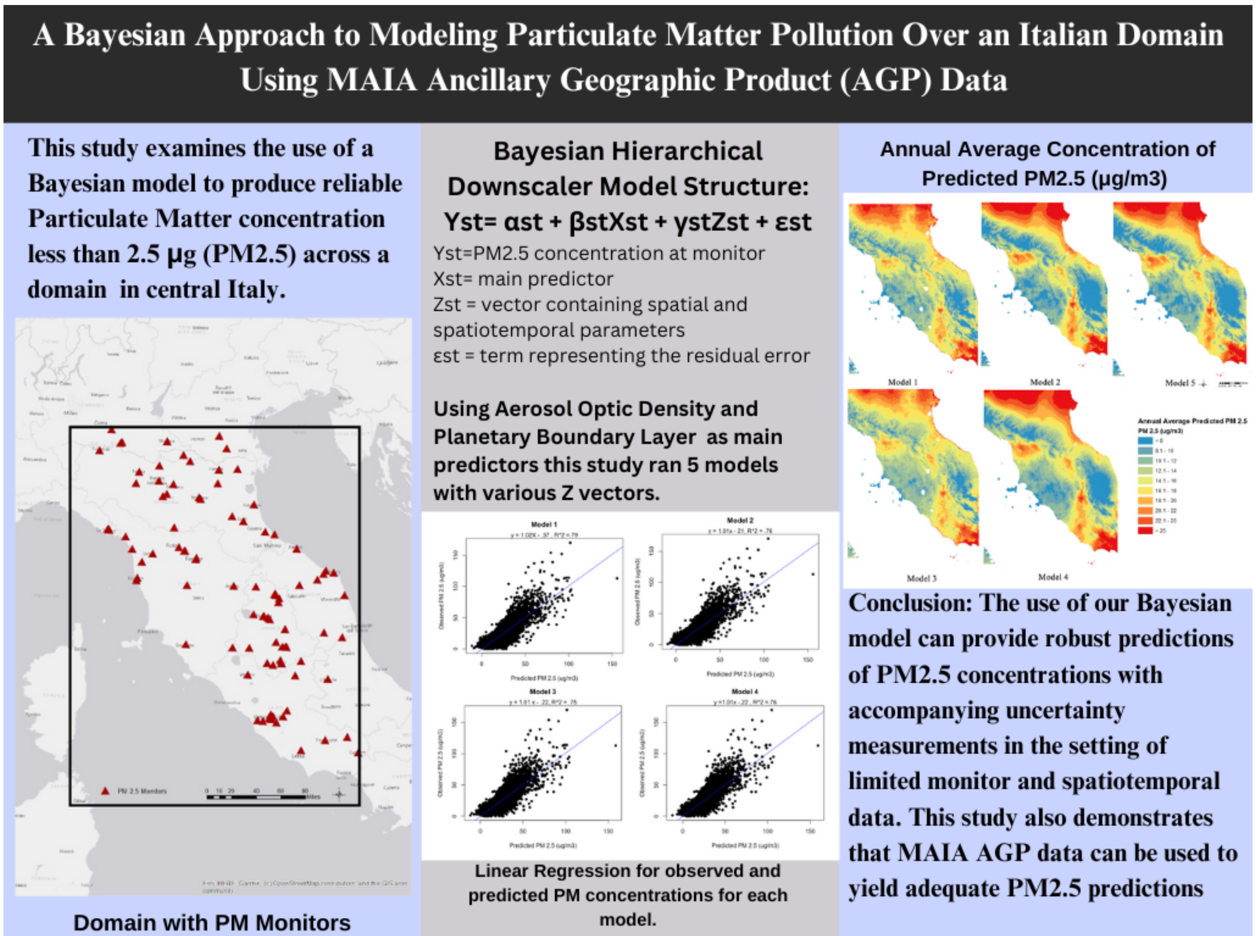
By

Harrison Goodall

Thesis Committee Chair: Yang Liu, PhD

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Epidemiology
2023

Visual abstract:



Key Words: Multiple Angle Imager for Aerosol (MAIA) , Aerosol Optic Density (AOD), Statistical Downscaling, Rome

Highlights:

1. To quantify health effects of $PM_{2.5}$ across the globe, fine scale spatial and temporal $PM_{2.5}$ predictions are needed, especially for regions with limited meteorological and air quality data.
2. We built upon data provided by Italian collaborators to run numerous models over an Italian domain. The purpose was to examine ways of predicting $PM_{2.5}$ with less air quality data and ancillary predictors than traditional machine learning approaches. A large portion of this project was also aimed at integrating the Multiple Angle Imager for Aerosol Ancillary Geographic Product (MAIA AGP) into our models to determine how models utilizing this data compare to models utilizing predictor variables collected at a local scale.
3. We used a Bayesian Hierarchical Downscaler model structure with a variety of predictors to determine the effects of inputs on $PM_{2.5}$ prediction capability.
4. Use of our Bayesian statistical model with MAIA AGP data generated reliable $PM_{2.5}$ concentration predictions and uncertainty measurements.
5. The ability to of our Bayesian statistical model to generate reliable $PM_{2.5}$ measurements in the context of limited ground monitoring data, and few spatiotemporal and spatial variables with MAIA AGP data suggests these methods and data sources could be used for various domains around the globe with limited PM and meteorological monitoring ability.

1. Introduction

Air pollution is a significant cause of morbidity and mortality worldwide (Fuller et al., 2022). In 2019, it is estimated that air pollution led to 6.7 millions deaths (Fuller et al., 2022). Air pollution can be divided into two main categories, ambient and indoor air pollution. While indoor air pollution has improved over the past two decades, deaths due to ambient air pollution have continued to rise from 2.9 million estimated deaths in 2000 to 4.5 million estimated deaths in 2019 (Fuller et al., 2022). As studies across the past several decades have demonstrated, a major component health related outcomes are due to small inhalable particles called particulate matter air pollution (PM): specifically particles 2.5 microns or less in width, referred to as $PM_{2.5}$. (Brook et al., 2010; Dockery et al., 1993; Franklin et al., 2015). $PM_{2.5}$ exposure has been linked with a wide array of health outcomes including increased cardiovascular-disease related mortality(Brook et al., 2010; Franklin et al., 2015) and increased incidence of lung cancer and respiratory morbidity and mortality (Faustini et al., 2011; Raaschou-Nielsen et al., 2013; Turner et al., 2011).

To continue the study of $PM_{2.5}$ exposure on health, it is necessary to have accurate methods of predicting $PM_{2.5}$ concentrations across large domains. While it is possible to measure ambient air pollution, including $PM_{2.5}$, with ground monitors, there are rarely enough monitors to use ground based data alone to accurately assess $PM_{2.5}$ exposure across a geographic domain, especially if the domain contains rural areas, which typically have fewer ground monitors (Stafoggia et al., 2019). Over the past decades, Aerosol optic density (AOD), a satellite based parameter, has been widely used to predict particulate matter over large areas (Engel-Cox et al., 2004; Lee et al., 2016). AOD is a measure of extinction that quantifies the ability of aerosols suspended in a vertical column of air to refract, absorb and scatter light (Acharya et al., 2021).

AOD has been widely used to predicted $PM_{2.5}$, PM_{10} and other aerosol concentrations in the atmosphere around the globe (Acharya et al., 2021; de Hoogh, Héritier, et al., 2018; Engel-Cox et al., 2004; Lee et al., 2016; Liu et al., 2007; Stafoggia et al., 2019). The quality of AOD data has increased in recent years with the release of NASA's Multi-Angle Implementation Correction (MAIAC), which provides high quality AOD data at 1-km^2 resolution(Lyapustin et al., 2011, 2018). MAIAC data has been used across various domains, most notably in the United States and Europe(Liu et al., 2007; Stafoggia et al., 2019).

Various statistical and machine learning models have been used to predict $PM_{2.5}$ based on AOD measurements, and overcome the challenges (Hoff & Christopher, 2009) to accurately utilizing remote sensing data (Chang et al., 2014; Geng, Murray, Chang, et al., 2018; Shtein et al., 2020; Stafoggia et al., 2019). In this paper we use a statistical model to predict $PM_{2.5}$ values. Our model builds off previous research that demonstrated the ability of statistical models using MAIAC data to predict $PM_{2.5}$ at a fine spatial scale(Hu et al., 2014; Kloog et al., 2014; Liu et al., 2004).

To overcome the challenge of using aerial gridded AOD data to predict spatial point-measurements from PM ground monitors, we utilized a unified hierarchical Bayesian downscaling model introduced by Chang et al.(Chang et al., 2014). By treating the relationship between gridded AOD measurements and the spatial points of PM ground monitors as temporally and spatially correlated random effects, statistical downscaling overcomes issues of spatial misalignment, allowing for the prediction of $PM_{2.5}$ at any spatial point within a given grid cell (Chang et al., 2014). The unified Bayesian hierarchical framework of this model structure allows for quantification of uncertainty of $PM_{2.5}$ predictions as Bayesian interference allows for uncertainty propagation via prediction intervals and prediction standard deviations(Chang et al., 2014). As in other PM modeling studies, we use meteorological data to increase the predictive capacity of our model, as meteorological and land use parameters can influence the relationship between PM and AOD (Beloconi et al., 2018; Just et al., 2015; Kloog et al., 2012).

In this paper we apply this statistical model to central Italy with the domain centering around Rome, a large city in Italy, that has been previously shown to have high levels of particulate matter air pollution (Fattorini & Regoli, 2020; Stafoggia et al., 2019). This study seeks to build upon previous high quality analyses that utilized machine learning (de Hoogh, Héritier, et al., 2018; Shtein et al., 2020; Stafoggia et al., 2017, 2019) and land-use regression models to predict PM concentrations over Europe (de Hoogh, Chen, et al., 2018; de Hoogh et al., 2016; Eeftens et al., 2012). While these two forms of modeling, used in conjunction with air quality and chemical transport models, appear to be the most popular techniques employed by Europe centered analyses, Beloconi et al. used a Bayesian Geostatistical model to predict annual PM concentrations across Europe, showing Bayesian statistical models can be used effectively over the European domain (Beloconi et al., 2018; Beloconi & Vounatsou, 2021). This paper is the first to use a unified Bayesian hierarchical downscaling model to predict daily $PM_{2.5}$ concentrations at high resolution across a European domain. We utilize statistical models to provide daily as opposed to yearly predictions and use AOD as a central predictor instead of relying solely on statistical performance in model selection, differentiating our methods from previous Bayesian and statistical models utilized over Europe.

This paper seeks to build upon a 2019 analysis by Stafoggia et al. that estimated daily $PM_{2.5}$ values for all of Italy at a 1-km^2 resolution using random-forest and ensemble learning models (Stafoggia et al., 2019). While Stafoggia's paper provides high quality PM predictions across Italy, the hierarchical structure of the model, where outputs from one step are used as inputs for the next, prevent prediction of uncertainty measurements to accompany prediction values. Using Stafoggia's data for the year 2015, we ran our Bayesian hierarchical downscaling model to compare the predictive quality of PM measurements generated by machine learning and statistical approaches, as well as generate uncertainty measurements to address limitations of previous studies. Additionally, we ran several variations of the same model using spatial and spatiotemporal predictors from our Italian collaborators and the Multi-angle Imager for Aerosols Ancillary Geographic Product (MAIA AGP) to examine model fit across various sources of data, from local data to widely available data from the MAIA AGP product. Rome has been the site of many epidemiologic investigations of the health effects of $PM_{2.5}$ and likely will continue to be a commonly studied domain in the future, thus uncertainly predictions are crucial to adequately quantifying exposure and assessing corresponding health effects(Amoatey et al., 2020; De Marco et al., 2018).

2 Materials and Methods:

2.1.1 Domain

The target domain of this study a 93,143 -km^2 region of central Italy that includes the city of Rome, the largest city in Italy (Figure 1). The coordinates of the domain were contributed by the NASA MAIA team. The Apennine Mountain range extends through the center of the domain, separating the east and west coasts of the country. The eastern coast of the domain borders the Adriatic Sea, while the western coast borders the Tyrrhenian and Ligurian Seas. The domain extends to the south of Rome terminating before Naples and extends to the north up to Parma and Bologna stopping just shy of the Po Valley region. Data was collected across this domain for the year 2015.

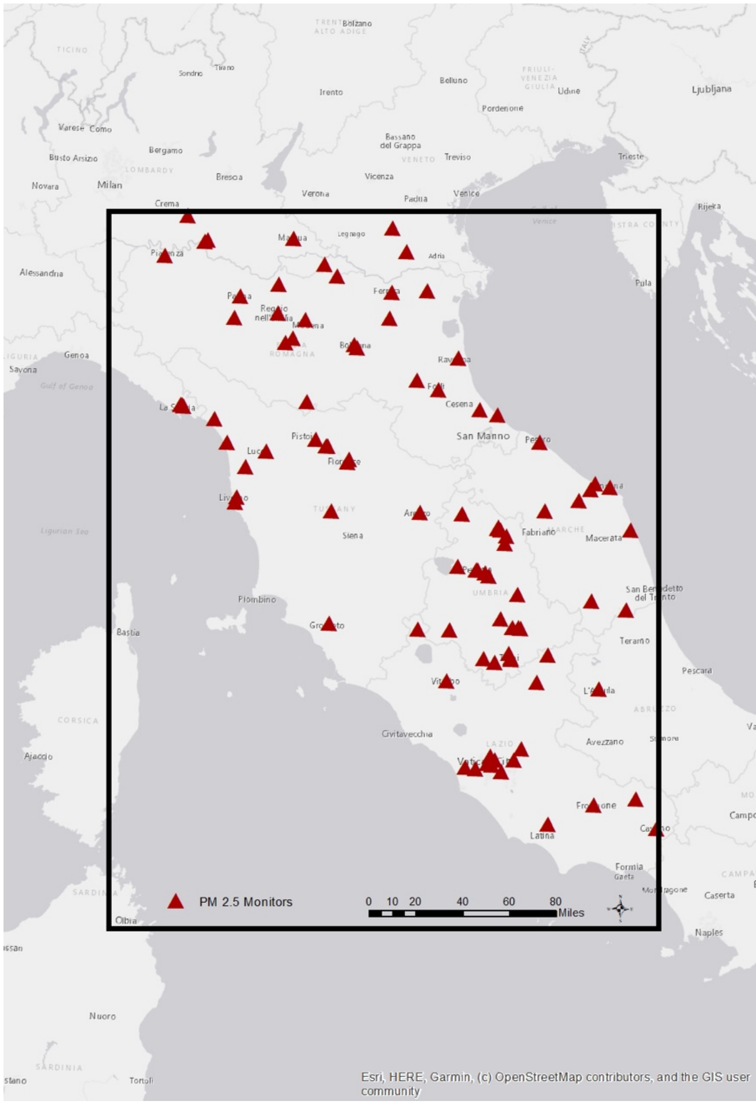


Figure 1: Study domain (indicated by the black box) with $PM_{2.5}$ monitors (red triangles) and Italian Regions (grey outlines).

2.1.2 Data Sets

As mentioned previously, data used in this analysis comes from two sources. The first and main source of data was data prepared by Stafoggia et. al (Stafoggia et al., 2019). The second source was the April 2020 MAIA AGP. The AGP is the product of geostatistical regression models that calibrate data from MAIA satellite image data and earth surface features to aid in the prediction of PM concentration. We chose to focus on one year, 2015, for this analysis. All variables used in the analysis can be seen in Table 1.

2.1.3 PM 2.5

Particulate Matter measurements were provided by the Italian Institute for Environmental Protection and Research. The domain included 109 air quality monitoring stations that collected daily PM 2.5 measurements. The daily PM 2.5 values used reflect the 24 hour mean from each monitor.

2.1.4 AOD

The AOD values used in this analysis were measurements of AOD at the 550nm (AOD55) wavelength based on the MAIAC AOD, which utilized 6 Modis Aqua L1B data for 2015.
In addition to MAIAC AOD values at 550nm, we utilized a gap filled AOD value to account for missing AOD values. This gap filled AOD at 550nm (AOD55.GF) was the result of an imputation process completed by Stafoggia et. al (Stafoggia et al., 2019). Stafoggia used co-located Copernicus Atmospheric Monitoring Service (CAMS) AOD measurements as input for their random forest with daily MAIAC AOD as their output variable.

2.1.5 Meteorologic Parameters

Meteorologic parameters were downloaded from the European Center for Medium-Range Weather forecasts, specifically from the ERA (ECMWF Re-Analysis) Interim Reanalysis at resolution of $0.125° \times 0.125°$. These parameters include temperature, wind speed and direction, and planetary boundary layer height. Wind speed and direction are composed of two vectors, u10, which represents wind moving parallel to the x axis, and v10, which represents wind moving parallel to the y axis.

2.1.6 Spatial Parameters

Elevation data was provided at a 30-meter spatial resolution by the Copernicus Land Monitoring Service- European Digital Elevation Model.

2.1.7 Other Spatiotemporal Parameters

We utilized monthly estimates of Normalized Difference Vegetative Index (NDVI) taken from the MODIS NDVI product (MOD13A3) at 1km resolution.

2.1.8 MAIA AGP Variables

Land Use/Land Cover from the AGP product was sourced from the 2009 European Space Agency global land cover data ("GlobCover") at a resolution of 30m/pixel. The data was reduced to 1km/pixel resolution in the conversion to the AGP target area.

Population Density data was taken from the Gridded Population of the World Version 4 (GPWv4) based on United Nations data from 2015 and processed at a $1km^2$ resolution.

Urban Density data was taken from the Urban Settlement Density Product of the AGP, which utilized data from Open Street Map (OSM), Global Man-made Impervious Surface (GMIS), and Global Human Built-up And Settlement Extent (HBASE) at a 30m/pixel resolution. Weighted 30m pixels were tabulated to provide urban density scores at a 1km resolution.

The Land/Water Identifiers product utilized data from a variety of sources, however most of the data came from the European Commission's Joint Research Center "Global Surface Water Transitions" product, which provides data at a 27m/pixel scale which was converted to a 1km scale using a nearest neighbor algorithm.

Elevation Slope Aspect was calculated using a 3x3 kernel projection of image elevation at a 333.3m/pixel resolution within 1km of the Albers Equal Area projection.

Average Elevation was calculated using the mean elevation of 100m/pixel elevation values within 1km of the Albers Equal Area projection. Elevation values used originated from the MAIA IAGP DEM elevation data set.

Table 1: Description, source, and spatial resolution of variables

Variable	**Description**	**Source**	**Spatial Resolution**
Domain	93,143 x 1km^2 grid cells	MAIA AGP Target Area	1 km
$PM_{2.5}$ Concentration (μg/m^3)	Average daily measurements from 109 ground monitors	Italian Institute for Environmental Protection and Research	-
AOD55	Aerosol Optic Density at 550nm	MAIAC AOD	0.125° × 0.125°
AOD55.GF	Gap filled AOD at 550nm	Stafoggia et al. 2019	1 km
Elevation (meters)	Average Elevation Value per 1 km grid	Copernicus Land Monitoring Service- European Digital Elevation Model	30 m
Normalized Difference Vegetative Index	Spatiotemporal green space predictor	MODIS NDVI product (MOD13A3)	1 km
Wind Speed (u10, v10)	Wind direction and speed parallel to x and y axes	ERA (ECMWF Re-Analysis) Interim Reanalysis	0.125° × 0.125°
Temperature (degrees Celsius)	Averaged temperature per grid cell	ERA (ECMWF Re-Analysis) Interim Reanalysis	0.125° × 0.125°
Planetary Boundary Layer	Layer of atmosphere with 1km of earth's surface	ERA (ECMWF Re-Analysis) Interim Reanalysis	0.125° × 0.125°
Average Elevation	Average Elevation Value per 1km grid	MAIA AGP 2020	1 km
Elevation Slope Aspect	Representation of the average surface-normal azimuth angle	MAIA AGP 2020	1 km
Urban Density	Aggregate of roads, impervious substances, and human settlement data	MAIA AGP 2020	1 km
Population Density	Number of persons per square km	MAIA AGP 2020	1 km
Land/Water Identifiers	Signifies land, ocean, inland water, and ephemeral water classes	MAIA AGP 2020	1 km
Land Cover/Use	Provides 23 descriptions of land use	MAIA AGP 2020	1 km

2.3 Model Structure

We utilized a unified Bayesian Hierarchical downscaling model to predict $PM_{2.5}$ based on AOD measurements and planetary boundary layer height at noon (PBL12). The details of this model are briefly enumerated below, if desired, more detail can be found in Chang et. al's 2014 paper (Chang et al., 2014). We applied the downscaling model to calibrate the relationship between daily average AOD/PBL12 values and $PM_{2.5}$ across our domain for all of 2015. As provided by Chang et al. (2014) and Geng et al. (2018) the relationship can be written as the following:

$$Y_{st} = \alpha_{st} + \beta_{st}X_{st} + \gamma_{st}Z_{st} + \varepsilon_{st}$$

Y_{st} represents the $PM_{2.5}$ concentration at monitor location s on day t. X_{st} is the main predictor at location s on day t. In this analysis AOD/PBL12 was the value used for X_{st} in the model. Similarly Z_{st} is a vector containing land use and meteorological parameters for location s on day t. α_{st} and β_{st} represent spatial covariate random effects which serve to correct the additive and multiplicative bias associated with AOD (Geng, Murray, Tong, et al., 2018). γ is a fixed-effects regression coefficient associated with Z_{st}. ε_{st} is a term representing the residual error which is assumed to have an independent and normal distribution with a mean of 0.

Using this model structure, we evaluated five models with different Z vectors to determine how robust this model is to variation of spatiotemporal inputs. Additionally, we examined model performance using data gathered from our Italian collaborators compared to data from the MAIA AGP. Each model used the same model structure and $PM_{2.5}$ ground monitor data. The X and Z values across the model variations can be seen in Table 2. The training data sets were smaller for Models 1 and 3, as there were many missing AOD55 values, which decreased the training data size compared to models utilizing AOD.GF which had no missing.

Table 2: X and Z values used in each model variation.

	Main Predictor (X)	Training Data Set Size	Spatial and Spatiotemporal Variables (Z)
Model 1	AOD55/PBL12	12,271	Elevation, Temperature, u10, v10, NDVI
Model 2	AOD55.GF/PBL12	36,389	Elevation, Temperature, u10, v10, NDVI
Model 3	AOD55/PBL12	12,271	Land Water Identifiers, Urban Density, Population Density, Land Cover, Elevation Slope Aspect, Average Elevation
Model 4	AOD55.GF/PBL12	36,389	Land Water Identifiers, Urban Density, Population Density, Land Cover, Elevation Slope Aspect, Average Elevation
Model 5	AOD55.GF/PBL12	36,389	Elevation, Temperature, u10, v10, NDVI, Urban Density, Population Density, Land Cover

2.4 Model Evaluation

We used a 10-fold regression calibration on the spatial and temporal components of $PM_{2.5}$ predictions separately, as well as overall calibration.

All statistical analysis was performed on R version 4.1.1. All maps were created in ArcGIS Desktop version 10.8.2 (ESRI, Redlands, CA).

3. Results:

3.1 Descriptive Statistics of Training Data Set

The descriptive statistics of the training data set can be seen in Table 3. The training data sets for models 1 and 3 were a subset of the main data set created by omitting rows with missing AOD55 values. This resulted in a training data set of n=12,271 for models using AOD55 (models 1 and 3), and a training data set of n=36,389 for models using AOD55.GF (Models 2,4 and 5). Both AOD55 and gap filled AOD (AOD.GF) share a mean of .15 and standard deviation of .08.

To visualize the pattern of missing AOD55 values we created a plot of the percentage of AOD55 values missing over the domain (Figure 2). This plot reveals a pattern of longitudinal lines where 100% of the AOD55 values are missing during out study period. As AOD55 was used in the X predictor position of our model, days with no AOD55 value cannot yield a $PM_{2.5}$ concentration prediction, which resulted in patterns of missing prediction values in our models using AOD55.

Table 3: Descriptive statistics of dependent and independent variables in the training dataset. Training data set size differed based on use of AOD55 or AOD.GF, as there were only 12,271 observations with complete AOD55 data. $PM_{2.5}$ concentrations from each size of data set are provided. Descriptive statistics of other variables are given for the full training data when utilizing AOD.GF, and have an n = 36,389.

Variable	**Mean**	**SD**	**Min**	**Max**
$PM_{2.5}$ ($\mu g/m^3$, n=36,389)	19.07	15.54	0	170
$PM_{2.5}$ ($\mu g/m^3$, n=12,271)	18.29	14.21	0	170
PBL12 (meters, n=36,389)	1102.28	631.24	12.19	3366.63
AOD55 (n=12,271)	0.15	0.08	0.01	0.82
AOD55.GF (n=36,389)	0.15	0.08	0.01	0.83
Elevation (meters, n=36,389)	160.49	212.58	-4.81	1000.95
Temperature (degrees Celsius, n=36,389)	10.38	6.92	-7.2	27.45
u10 (n=36,389)	-0.26	1.76	-11.42	8.76
v10 (n=36,389)	-0.28	1.86	-13.18	6.48
NDVI (n=36,389)	0.44	0.13	-0.02	0.88
Urban Density (n=36,389)	10.86	7.99	0	29
Population Density (population/km^2, n=36,389)	2983	3800.2	0	17569
Land Cover/Use (n=36,389)	108.1	82.88	11	210
Land Water Identifiers (n=36,389)	1.01	0.22	0	2

Elevation Slope Aspect (n=36,389)	76	86..49	0	253
Average Elevation (meters, n=36,389)	166.7	220	0	977

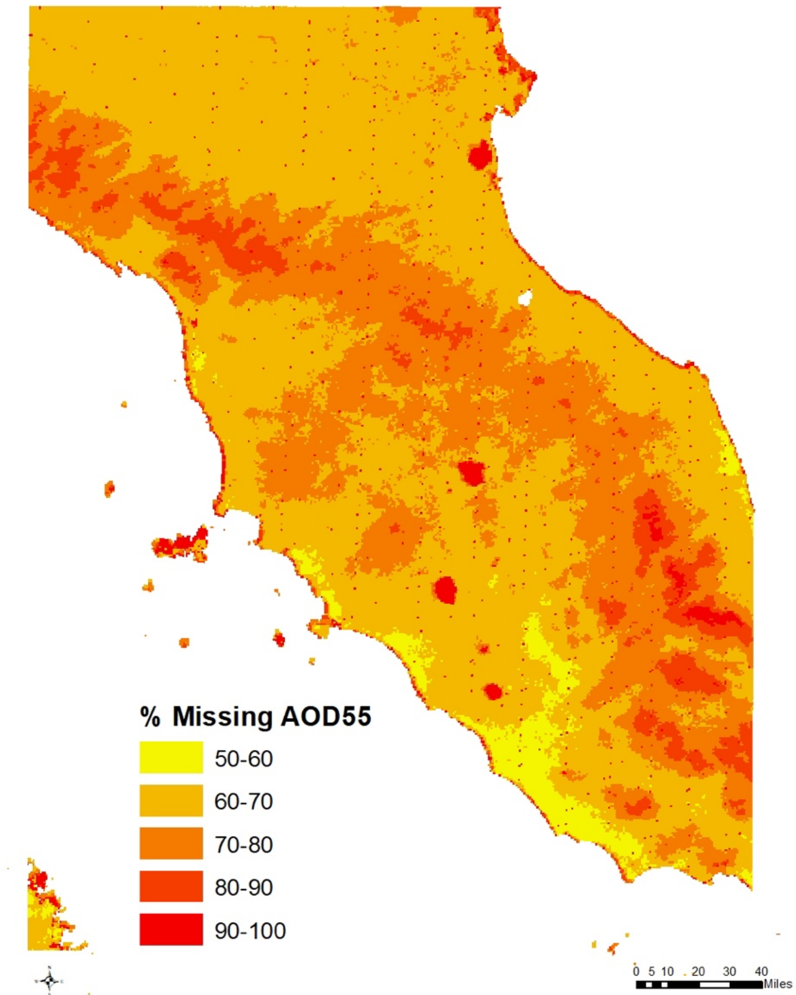


Figure 2: Percentage of Missing AOD55 values

3.2 Model Performance

Regression plots of observed and predicted $PM_{2.5}$ concentrations for each model can be seen in Figure 3. Overall, the intercepts are close to 0 (average -.22) and slopes are close to 1 (average 1.01).

Table 4 shows the result of the 10-fold cross validation. The highest overall R^2 values in the cross validation were generated by Models 2, 4 and 5 (R^2 values of .726, .721, .725 respectively). Spatial cross validation results yielded higher R^2 values (.02 higher on average), lower RMSE (.12 lower on average) and standard deviation values (2.05 lower on average) and slopes closer to unity than the temporal cross validation results. The overall R^2 values were higher than either spatial or temporal results, with a range from .712 to .725.

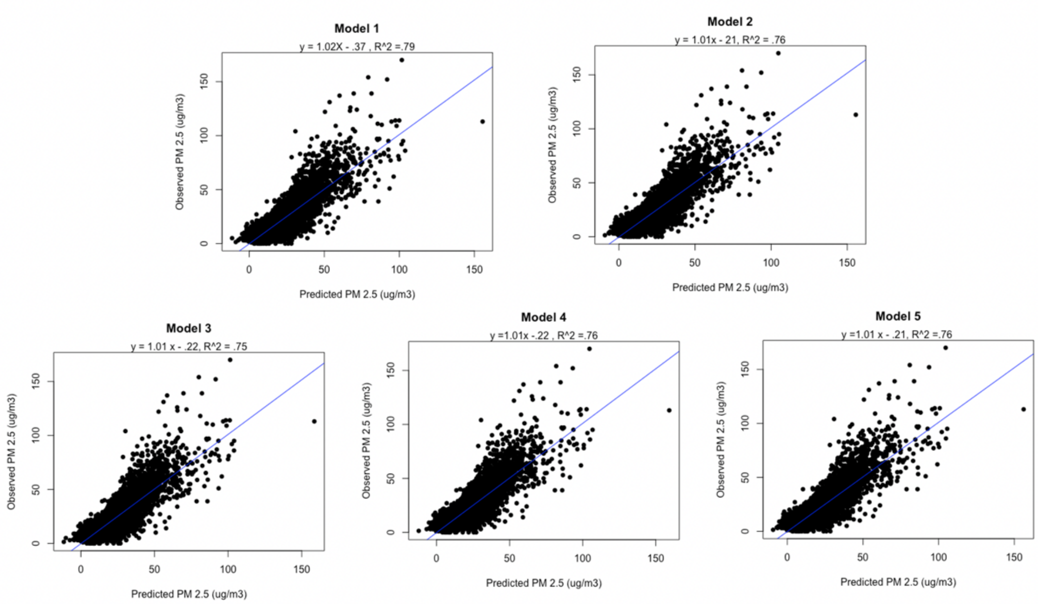


Figure 3: Linear Regression for observed and predicted PM concentrations for each model. Regression line represented by the blue line.

Table 4: Results of 10-fold cross validation

		R^2	RMSE	Rate	Slope	SD
Model 1						
	Overall	0.716	7.6129	0.958	0.958	7.594
	Spatial	0.664	8.378	0.943	0.960	7.595
	Temporal	0.665	8.200	0.960	0.934	8.566
Model 2						
	Overall	0.726	8.140	0.951	0.992	8.213
	Spatial	0.659	9.055	0.936	0.978	8.276
	Temporal	0.641	9.3623	0.969	0.913	10.894
Model 3						
	Overall	0.712	7.681	0.957	0.962	7.688
	Spatial	0.652	8.583	0.942	0.944	7.645
	Temporal	0.639	8.619	0.955	0.916	8.708
Model 4						
	Overall	0.721	8.205	0.952	0.991	8.323
	Spatial	0.655	9.125	0.936	0.957	8.360
	Temporal	0.639	9.386	0.971	0.920	11.265
Model 5						
	Overall	0.725	8.146	0.9521	0.993	8.247
	Spatial	0.653	9.113	0.935	0.956	8.254
	Temporal	0.632	9.300	0.970	0.908	10.932

3.3 Pm2.5 predictions

Figure 4 shows the spatial distribution of annual averages of predicted $PM_{2.5}$ values across the target domain. There are lower predicted $PM_{2.5}$ concentrations over the Apennine mountains in the center of the country, and higher $PM_{2.5}$ concentrations around Rome and Naples in the South, and the southern regions of the Po valley, at the northern most edge of the domain. Models that used AOD55 (1 and 3) have a similar pattern of missing data as seen in Figure 2.

Table 5 displays the descriptive statistics of predicted $PM_{2.5}$ across all five models. These values are taken from the output of the Bayesian Hierarchical Downscaling training model. Descriptive statistics of inputs for this training model can be seen in Table 3.

Figure 5 shows the spatial distribution of annual averages of standard deviation of predicted $PM_{2.5}$ values across the domain. The same pattern of missing data from Figure 2 can be clearly seen in models 1 and 3. However models using AOD55.GF (2,4 and 5) have unusually uniform distributions of standard deviation values across the central and upper portions of the target domain compared to the models utilizing AOD55.

Figure 6 shows the monthly averages of observed and predicted $PM_{2.5}$ concentrations. Figure 7 shows the daily averages of observed and predicted $PM_{2.5}$ concentrations.

An enlarged image of the spatial distribution of annual averages of predicted $PM_{2.5}$ values over Rome can be found in the supplementary materials (S1).

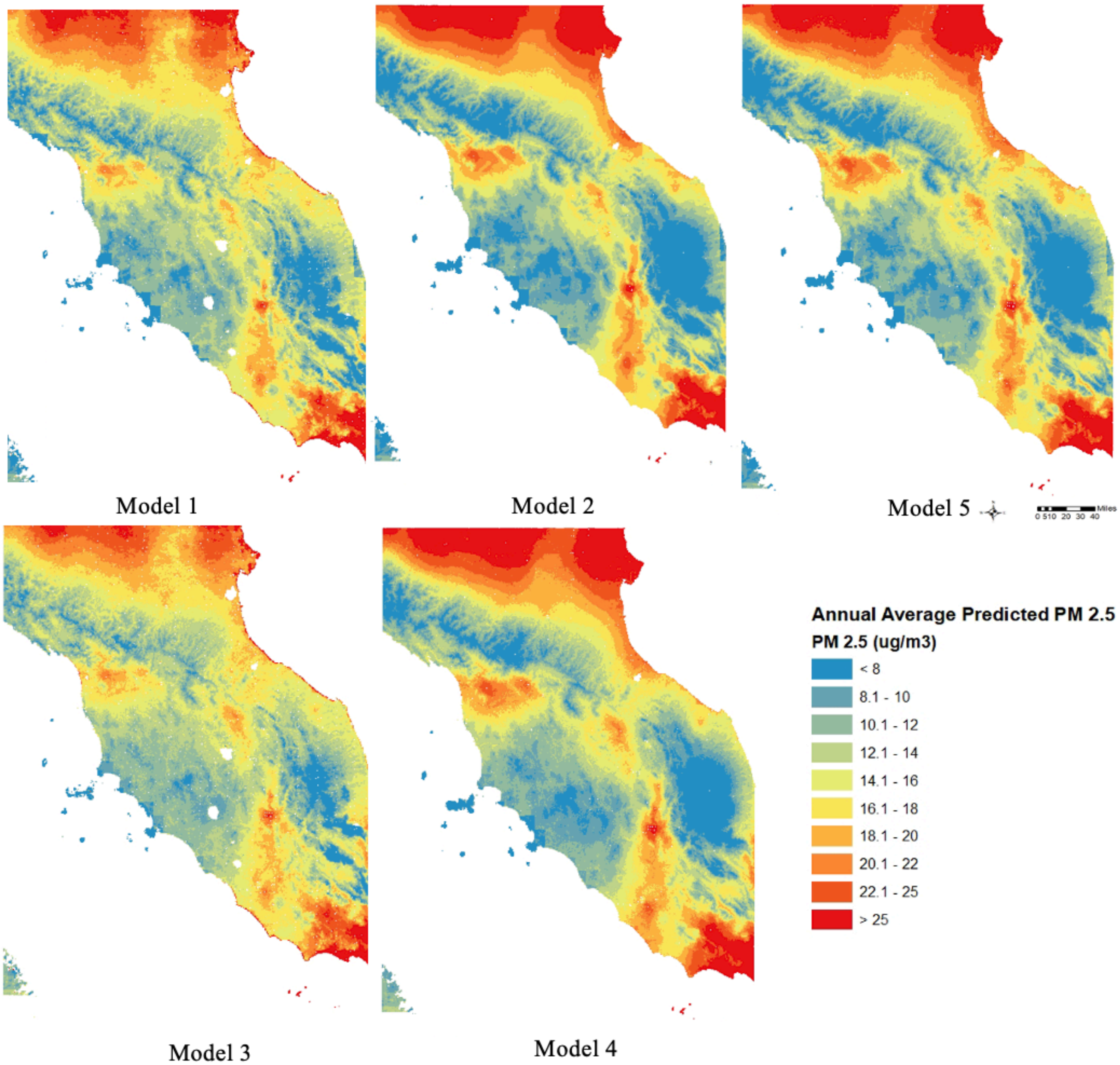


Figure 4: Annual Average Concentration of Predicted $PM_{2.5}$ ($\mu g/m^3$)

Table 5: Observed and predicted $PM_{2.5}$ concentrations ($\mu g/m^3$)

	Mean	**SD**	**Percentile**				
			5	***25***	***50***	***75***	***95***
Observed	18.3	14.2	5	10	14	21.4	46
Predicted – Model 1	18.3	12.24	4.8	10	15	22.6	42.3
Predicted – Model 2	18.3	12.26	4.6	10	16	22.9	41.7
Predicted – Model 3	18.3	14.21	4.7	10	16	23	41.6
Predicted – Model 4	18.3	12.21	4.7	10	15	23	41.6
Predicted – Model 5	18.3	12.26	4.6	10	16	22.9	41.7

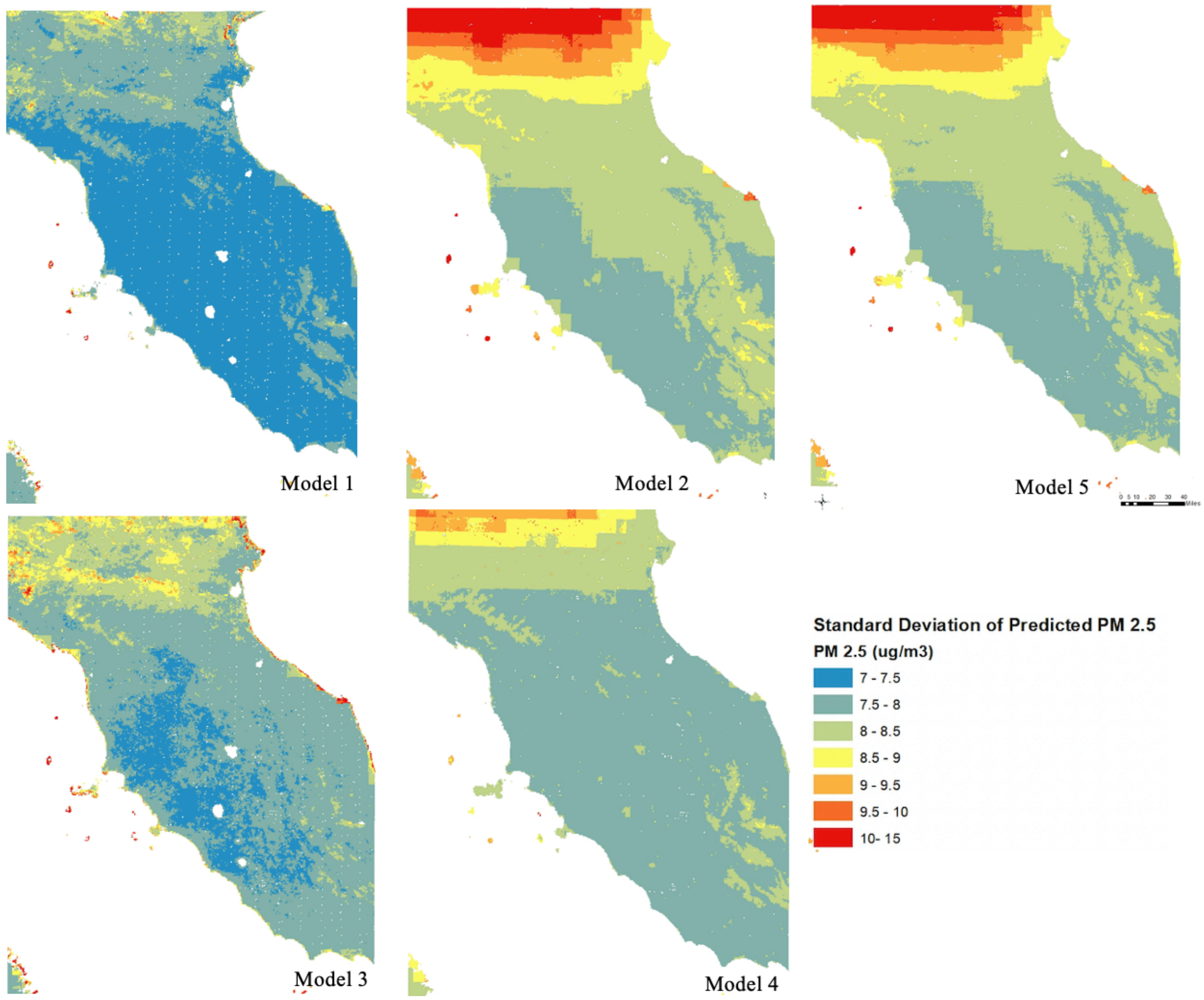


Figure 5: Annual Average Standard Deviation of Predicted $PM_{2.5}$ Concentration ($\mu g/m^3$)

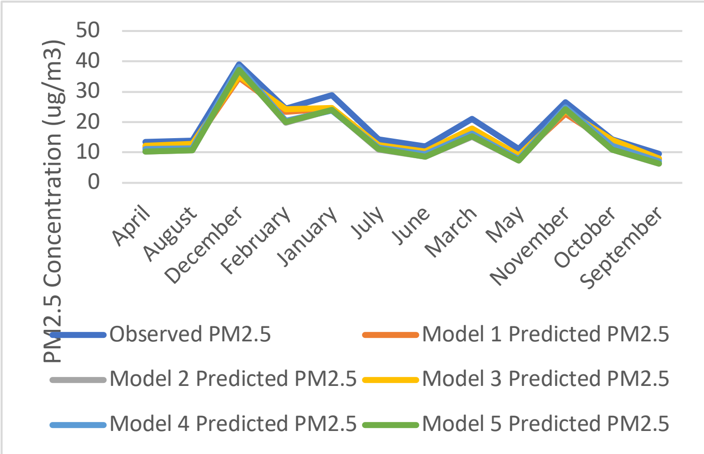


Figure 6: Monthly averages of observed and predicted $PM_{2.5}$ concentrations.

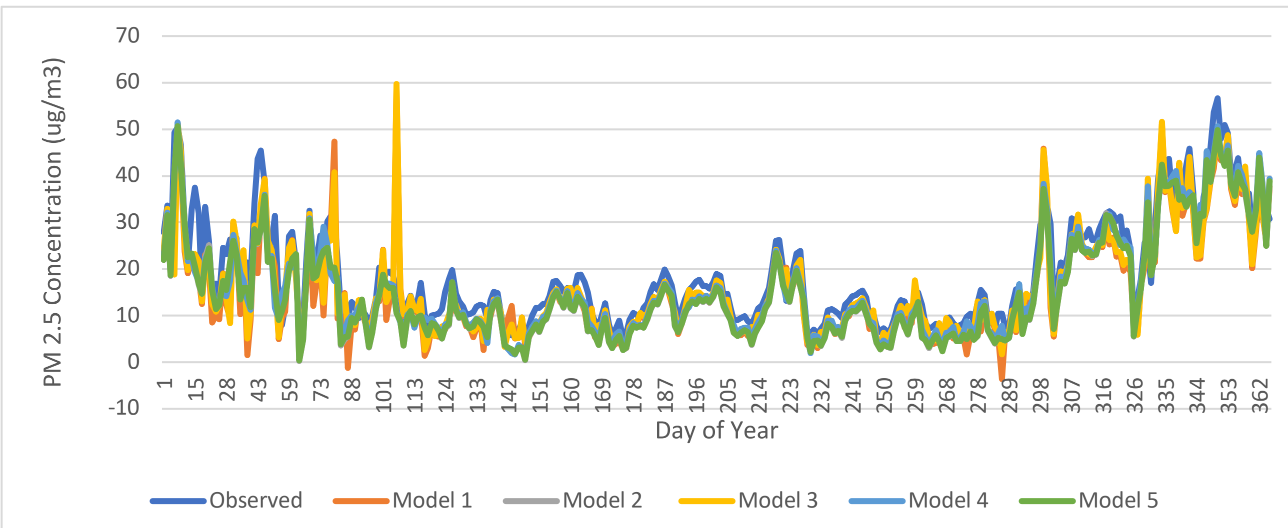


Figure 7: Daily averages of observed and predicted $PM_{2.5}$ Concentrations.

4. Discussion

This paper applies a Bayesian hierarchical downscaling model to produce $PM_{2.5}$ predictions across a large portion of Italy. Working within a domain identified by NASA's MAIA team, we compared different versions of the downscaling model with various spatiotemporal and spatial predictor inputs. We then compared these model variations to each other, and previous work done in Italy by Stafoggia et. al in 2019. Our results show that the downscaling model can provide reliable $PM_{2.5}$ predictions across the domain. Each model explained 71 to 72% of the overall variability of $PM_{2.5}$ in left out monitors during cross validation as seen in Table 5. Additionally, the daily average predicted $PM_{2.5}$ mirrored daily averages concentrations of monitoring stations, showing that predictions from this model can be used on a fine temporal scale in future exposure research.

When we compare the models to each other, we see that there is very little variation of R^2 among the models, and the regression plots for the models have slopes close to unity and negative intercepts of -.21 or -.22 (Figure 3). Notably the results of temporal cross validation showed lower R^2 values and higher RMSE and SD values than the spatial cross validation results. These findings may suggest that the model performs best on the spatial scale, although temporal performance is still adequate. The most notable result across each model is the similar performance between models despite large differences in Z vector contents (Table 1). Each Z vector was chosen from Italy data or MAIA AGP data in accordance with MAIA guidance of utilization of spatiotemporal and spatial predictors when predicting $PM_{2.5}$ (provided in the MAIA AGP). The performance of the model is similar when using data provided from Italian collaborators (models 1 and 2) compared to using data from the MAIA AGP (models 3 and 4). This suggests that MAIA AGP data could successfully be used to predict $PM_{2.5}$ in the context of our downscaling model, which could eliminate the need for large amounts of local data collection. These results also highlight that the X predictor is the main determinate of predictive ability when utilizing the downscaling model. We used extinction (AOD/PBL) as our X predictor. The PBL used in each model remained constant and was contributed by our Italian collaborators. Given the importance of the X variable in our model structure, it is logical that we observed the largest amount of variation in cross validation results between models using AOD55 versus gap filled AOD55. When gap filled AOD55 was used as the main predictor, the R^2 increased (average of .01). This is not surprising as the training sets for models using gap filled AOD55 were larger than non-gap filled AOD55 (Table 2), which likely resulted in the increased the R^2. However, as Figure 5 shows, standard deviation of the predicted $PM_{2.5}$ concentrations increased when using gap filled AOD55, which is a result that must be investigated further.

As mentioned above, this analysis utilized and built upon work done by Italian scientists (Stafoggia et al., 2019). In their 2019 paper, Stafoggia et al. utilized a machine learning model to generate PM predictions across all of Italy. To accomplish this, they used many spatiotemporal predictors and spatial predictors (approximately 39). Before using small scale predictors to improve PM predictions, their model generated an R^2 of .81 and a RMSE of 6.39 across the entire domain of Italy using training data from 229 stations in 2015. While their model remains superior to ours, with higher R^2 values, our model serves as useful compendium to their previous findings. Despite only using 5 to 6 variables in the Z vector for each model, and utilizing training data from 109 PM monitoring stations, our model was able to generate robust R^2 values. These results indicate that our model can be utilized in resource limited settings where generating a myriad of spatiotemporal and spatial predictors may not be an option. Our model can make predictions of $PM_{2.5}$ with limited training data and sparse ground monitoring measurements, making it a useful addition to the field of PM modeling. Additionally, our data augments that provided by our Italian collaborators and other ensemble machine learning models by its ability generate uncertainty measurements, which are important when considering the use of this model to generate predictions for future epidemiologic research.

Limitations of this study include the presence of data gaps within all prediction maps. These missing predictions are largely due to the lack of X predictor values within a given grid cell, however a small subset of missing data persist when using gap filled AOD measurements in the models using MAIA data. The lack of AOD55 data across the domain was a significant limitation of this project and interfered with our model's ability to yield PM predictions across the entire domain when using AOD55 in the X predictor. Additionally our model occasionally yields negative PM predictions, which is a limitation of its design.

5. Conclusion

By comparing multiple models with different predictor variables, we showed that our Bayesian Hierarchical Downscaling model can provide robust predictions of $PM_{2.5}$ concentrations across a large domain with limited monitor data and spatiotemporal predictors. Additionally, the model can provide uncertainty measurements, which distinguishes it from previous models generating PM predictions over the Italian domain. The model's adequate performance on a spatial and temporal scale, and ability to utilize MAIA AGP data without sacrificing prediction quality suggest it could be a useful tool to great PM exposure maps across Italy as well as domains with limited ground monitoring and meteorological data.

6. Acknowledgements

The work of G. Geng, Y. Liu, H. Chang, and X. Meng was partially supported by the NASA Applied Sciences Program (grant NNX16AQ28G, PI: Y. Liu). Special thank you to Danlu Zhang for her help planning and executing the early stages of this project.

7. References

Acharya, P., Barik, G., Gayen, B. K., Bar, S., Maiti, A., Sarkar, A., Ghosh, S., De, S. K., & Sreekesh, S. (2021). Revisiting the levels of Aerosol Optical Depth in south-southeast Asia, Europe and USA amid the COVID-19 pandemic using satellite observations. *Environmental Research*, *193*, 110514. https://doi.org/10.1016/j.envres.2020.110514

Amoatey, P., Sicard, P., De Marco, A., & Khaniabadi, Y. O. (2020). Long-term exposure to ambient PM2.5 and impacts on health in Rome, Italy. *Clinical Epidemiology and Global Health*, *8*(2), 531–535. https://doi.org/10.1016/j.cegh.2019.11.009

Beloconi, A., Chrysoulakis, N., Lyapustin, A., Utzinger, J., & Vounatsou, P. (2018). Bayesian geostatistical modelling of PM10 and PM2.5 surface level concentrations in Europe using high-resolution satellite-derived products. *Environment International*, *121*, 57–70. https://doi.org/10.1016/j.envint.2018.08.041

Beloconi, A., & Vounatsou, P. (2021). Substantial Reduction in Particulate Matter Air Pollution across Europe during 2006–2019: A Spatiotemporal Modeling Analysis. *Environmental Science & Technology*, *55*(22), 15505–15518. https://doi.org/10.1021/acs.est.1c03748

Brook, R. D., Rajagopalan, S., Pope, C. A., Brook, J. R., Bhatnagar, A., Diez-Roux, A. V., Holguin, F., Hong, Y., Luepker, R. V., Mittleman, M. A., Peters, A., Siscovick, D., Smith, S. C., Whitsel, L., & Kaufman, J. D. (2010). Particulate Matter Air Pollution and Cardiovascular Disease: An Update to the Scientific Statement From the American Heart Association. *Circulation*, *121*(21), 2331–2378. https://doi.org/10.1161/CIR.0b013e3181dbece1

Chang, H. H., Hu, X., & Liu, Y. (2014). Calibrating MODIS aerosol optical depth for predicting daily PM2.5 concentrations via statistical downscaling. *Journal of Exposure Science & Environmental Epidemiology*, *24*(4), 398–404. https://doi.org/10.1038/jes.2013.90

de Hoogh, K., Chen, J., Gulliver, J., Hoffmann, B., Hertel, O., Ketzel, M., Bauwelinck, M., van Donkelaar, A., Hvidtfeldt, U. A., Katsouyanni, K., Klompmaker, J., Martin, R. V., Samoli, E., Schwartz, P. E., Stafoggia, M., Bellander, T., Strak, M., Wolf, K., Vienneau, D., … Hoek, G. (2018). Spatial PM2.5, NO2, O3 and BC models for Western Europe – Evaluation of spatiotemporal stability. *Environment International*, *120*, 81–92. https://doi.org/10.1016/j.envint.2018.07.036

de Hoogh, K., Gulliver, J., Donkelaar, A. van, Martin, R. V., Marshall, J. D., Bechle, M. J., Cesaroni, G., Pradas, M. C., Dedele, A., Eeftens, M., Forsberg, B., Galassi, C., Heinrich, J., Hoffmann, B., Jacquemin, B., Katsouyanni, K., Korek, M., Künzli, N., Lindley, S. J., … Hoek, G. (2016). Development of West-European PM 2.5 and NO 2 land use regression models incorporating satellite-derived and chemical transport modelling data. *Environmental Research*, *151*, 1–10. https://doi.org/10.1016/j.envres.2016.07.005

de Hoogh, K., Héritier, H., Stafoggia, M., Künzli, N., & Kloog, I. (2018). Modelling daily PM2.5 concentrations at high spatio-temporal resolution across Switzerland. *Environmental Pollution*, *233*, 1147–1154. https://doi.org/10.1016/j.envpol.2017.10.025

De Marco, A., Amoatey, P., Khaniabadi, Y. O., Sicard, P., & Hopke, P. K. (2018). Mortality and morbidity for cardiopulmonary diseases attributed to PM2.5 exposure in the metropolis of Rome, Italy. *European Journal of Internal Medicine*, *57*, 49–57. https://doi.org/10.1016/j.ejim.2018.07.027

Dockery, D. W., Pope, C. A., Xu, X., Spengler, J. D., Ware, J. H., Fay, M. E., Ferris, B. G., & Speizer, F. E. (1993). An Association between Air Pollution and Mortality in Six U.S. Cities. *New England Journal of Medicine*, *329*(24), 1753–1759. https://doi.org/10.1056/NEJM199312093292401

Eeftens, M., Beelen, R., de Hoogh, K., Bellander, T., Cesaroni, G., Cirach, M., Declercq, C., Dėdelė, A., Dons, E., de Nazelle, A., Dimakopoulou, K., Eriksen, K., Falq, G., Fischer, P., Galassi, C., Gražulevičienė, R., Heinrich, J., Hoffmann, B., Jerrett, M., … Hoek, G. (2012). Development of Land Use Regression Models for $PM_{2.5}$, $PM_{2.5}$ Absorbance, PM_{10} and PM_{coarse} in 20 European Study Areas; Results of the ESCAPE Project. *Environmental Science & Technology*, *46*(20), 11195–11205. https://doi.org/10.1021/es301948k

Engel-Cox, J. A., Holloman, C. H., Coutant, B. W., & Hoff, R. M. (2004). Qualitative and quantitative evaluation of MODIS satellite sensor data for regional and urban scale air quality. *Atmospheric Environment*, *38*(16), 2495–2509. https://doi.org/10.1016/j.atmosenv.2004.01.039

Fattorini, D., & Regoli, F. (2020). Role of the chronic air pollution levels in the Covid-19 outbreak risk in Italy. *Environmental Pollution*, *264*, 114732. https://doi.org/10.1016/j.envpol.2020.114732

Faustini, A., Stafoggia, M., Berti, G., Bisanti, L., Chiusolo, M., Cernigliaro, A., Mallone, S., Primerano, R., Scarnato, C., Simonato, L., Vigotti, M. A., Forastiere, F., & on behalf of the EpiAir Collaborative Group. (2011). The relationship between ambient particulate matter and respiratory mortality: A multi-city study in Italy. *European Respiratory Journal*, *38*(3), 538–547. https://doi.org/10.1183/09031936.00093710

Franklin, B. A., Brook, R., & Arden Pope, C. (2015). Air Pollution and Cardiovascular Disease. *Current Problems in Cardiology*, *40*(5), 207–238. https://doi.org/10.1016/j.cpcardiol.2015.01.003

Fuller, R., Landrigan, P. J., Balakrishnan, K., Bathan, G., Bose-O'Reilly, S., Brauer, M., Caravanos, J., Chiles, T., Cohen, A., Corra, L., Cropper, M., Ferraro, G., Hanna, J., Hanrahan, D., Hu, H., Hunter, D., Janata, G., Kupka, R., Lanphear, B., … Yan, C. (2022). Pollution and health: A progress update. *The Lancet Planetary Health*, *6*(6), e535–e547. https://doi.org/10.1016/S2542-5196(22)00090-0

Geng, G., Murray, N. L., Chang, H. H., & Liu, Y. (2018). The sensitivity of satellite-based PM2.5 estimates to its inputs: Implications to model development in data-poor regions. *Environment International*, *121*, 550–560. https://doi.org/10.1016/j.envint.2018.09.051

Geng, G., Murray, N. L., Tong, D., Fu, J. S., Hu, X., Lee, P., Meng, X., Chang, H. H., & Liu, Y. (2018). Satellite-Based Daily PM $_{2.5}$ Estimates During Fire Seasons in Colorado. *Journal of Geophysical Research: Atmospheres*, *123*(15), 8159–8171. https://doi.org/10.1029/2018JD028573

Hoff, R. M., & Christopher, S. A. (2009). Remote Sensing of Particulate Pollution from Space: Have We Reached the Promised Land? *Journal of the Air & Waste Management Association*, *59*(6), 645–675. https://doi.org/10.3155/1047-3289.59.6.645

Hu, X., Waller, L. A., Lyapustin, A., Wang, Y., Al-Hamdan, M. Z., Crosson, W. L., Estes, M. G., Estes, S. M., Quattrochi, D. A., Puttaswamy, S. J., & Liu, Y. (2014). Estimating ground-level PM2.5 concentrations in the Southeastern United States using MAIAC AOD retrievals and a two-stage model. *Remote Sensing of Environment*, *140*, 220–232. https://doi.org/10.1016/j.rse.2013.08.032

Just, A. C., Wright, R. O., Schwartz, J., Coull, B. A., Baccarelli, A. A., Tellez-Rojo, M. M., Moody, E., Wang, Y., Lyapustin, A., & Kloog, I. (2015). Using High-Resolution Satellite Aerosol Optical Depth To Estimate Daily PM $_{2.5}$ Geographical Distribution in Mexico City. *Environmental Science & Technology*, *49*(14), 8576–8584. https://doi.org/10.1021/acs.est.5b00859

Kloog, I., Chudnovsky, A. A., Just, A. C., Nordio, F., Koutrakis, P., Coull, B. A., Lyapustin, A., Wang, Y., & Schwartz, J. (2014). A new hybrid spatio-temporal model for estimating daily multi-year PM2.5 concentrations across northeastern USA using high resolution aerosol optical depth data. *Atmospheric Environment*, *95*, 581–590. https://doi.org/10.1016/j.atmosenv.2014.07.014

Kloog, I., Nordio, F., Coull, B. A., & Schwartz, J. (2012). Incorporating Local Land Use Regression And Satellite Aerosol Optical Depth In A Hybrid Model Of Spatiotemporal PM $_{2.5}$ Exposures In The Mid-Atlantic States. *Environmental Science & Technology*, *46*(21), 11913–11921. https://doi.org/10.1021/es302673e

Lee, M., Kloog, I., Chudnovsky, A., Lyapustin, A., Wang, Y., Melly, S., Coull, B., Koutrakis, P., & Schwartz, J. (2016). Spatiotemporal prediction of fine particulate matter using high-resolution satellite images in the Southeastern US 2003–2011. *Journal of Exposure Science & Environmental Epidemiology*, *26*(4), 377–384. https://doi.org/10.1038/jes.2015.41

Liu, Y., Franklin, M., Kahn, R., & Koutrakis, P. (2007). Using aerosol optical thickness to predict ground-level PM2.5 concentrations in the St. Louis area: A comparison between MISR and MODIS. *Multi-Angle Imaging SpectroRadiometer (MISR) Special Issue*, *107*(1), 33–44. https://doi.org/10.1016/j.rse.2006.05.022

Liu, Y., Park, R. J., Jacob, D. J., Li, Q., Kilaru, V., & Sarnat, J. A. (2004). Mapping annual mean ground-level PM $_{2.5}$ concentrations using Multiangle Imaging Spectroradiometer aerosol optical thickness over the contiguous United States: MAPPING SURFACE PM $_{2.5}$ USING MISR AOT. *Journal of Geophysical Research: Atmospheres*, *109*(D22), n/a-n/a. https://doi.org/10.1029/2004JD005025

Lyapustin, A., Wang, Y., Korkin, S., & Huang, D. (2018). MODIS Collection 6 MAIAC algorithm. *Atmospheric Measurement Techniques*, *11*(10), 5741–5765. https://doi.org/10.5194/amt-11-5741-2018

Lyapustin, A., Wang, Y., Laszlo, I., Kahn, R., Korkin, S., Remer, L., Levy, R., & Reid, J. S. (2011). Multiangle implementation of atmospheric correction (MAIAC): 2. Aerosol algorithm. *Journal of Geophysical Research*, *116*(D3), D03211. https://doi.org/10.1029/2010JD014986

Raaschou-Nielsen, O., Andersen, Z. J., Beelen, R., Samoli, E., Stafoggia, M., Weinmayr, G., Hoffmann, B., Fischer, P., Nieuwenhuijsen, M. J., Brunekreef, B., Xun, W. W., Katsouyanni, K., Dimakopoulou, K., Sommar, J., Forsberg, B., Modig, L., Oudin, A., Oftedal, B., Schwarze, P. E., … Hoek, G. (2013). Air pollution and lung cancer incidence in 17 European cohorts: Prospective analyses from the European Study of Cohorts for Air Pollution Effects (ESCAPE). *The Lancet Oncology*, *14*(9), 813–822. https://doi.org/10.1016/S1470-2045(13)70279-1

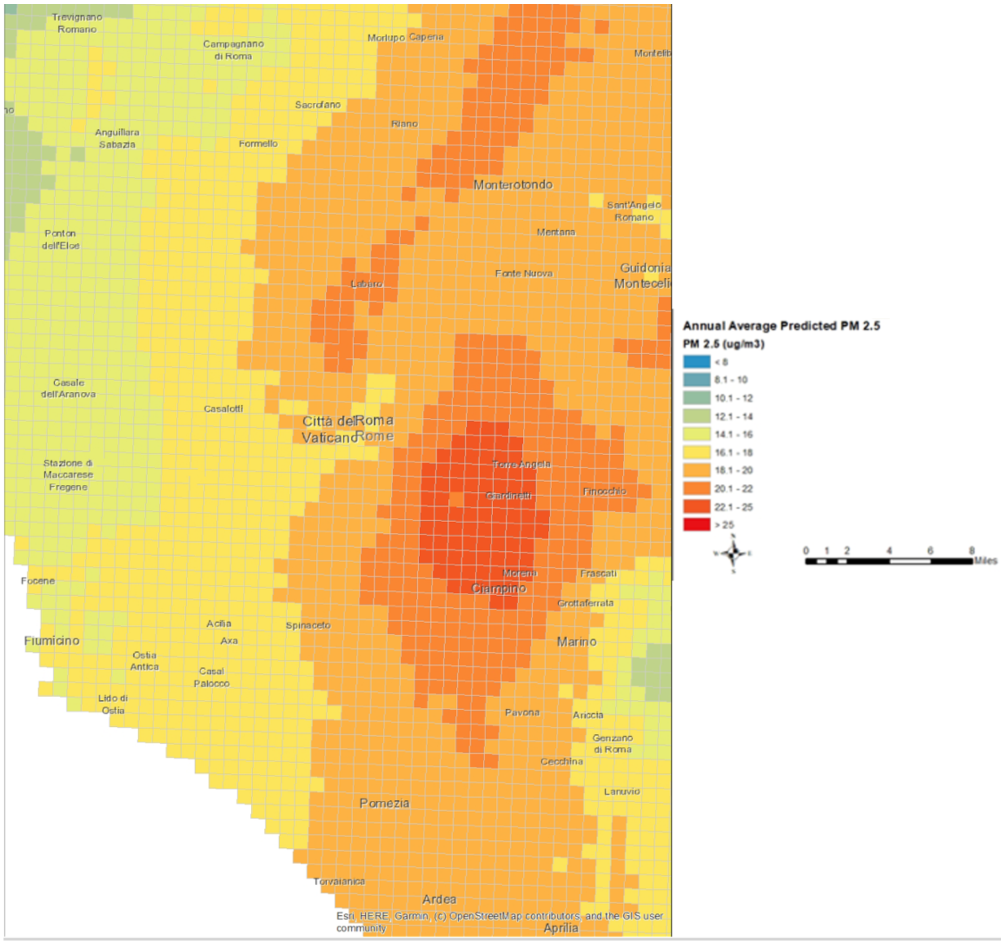
Shtein, A., Kloog, I., Schwartz, J., Silibello, C., Michelozzi, P., Gariazzo, C., Viegi, G., Forastiere, F., Karnieli, A., Just, A. C., & Stafoggia, M. (2020). Estimating Daily PM $_{2.5}$ and PM $_{10}$ over Italy Using an Ensemble Model. *Environmental Science & Technology*, *54*(1), 120–128. https://doi.org/10.1021/acs.est.9b04279

Stafoggia, M., Bellander, T., Bucci, S., Davoli, M., de Hoogh, K., de' Donato, F., Gariazzo, C., Lyapustin, A., Michelozzi, P., Renzi, M., Scortichini, M., Shtein, A., Viegi, G., Kloog, I., & Schwartz, J. (2019). Estimation of daily PM10 and PM2.5 concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model. *Environment International*, *124*, 170–179. https://doi.org/10.1016/j.envint.2019.01.016

Stafoggia, M., Schwartz, J., Badaloni, C., Bellander, T., Alessandrini, E., Cattani, G., de' Donato, F., Gaeta, A., Leone, G., Lyapustin, A., Sorek-Hamer, M., de Hoogh, K., Di, Q., Forastiere, F., & Kloog, I. (2017). Estimation of daily PM10 concentrations in Italy (2006–2012) using finely resolved satellite data, land use variables and meteorology. *Environment International*, *99*, 234–244. https://doi.org/10.1016/j.envint.2016.11.024

Turner, M. C., Krewski, D., Pope, C. A., Chen, Y., Gapstur, S. M., & Thun, M. J. (2011). Long-term Ambient Fine Particulate Matter Air Pollution and Lung Cancer in a Large Cohort of Never-Smokers. *American Journal of Respiratory and Critical Care Medicine*, *184*(12), 1374–1381. https://doi.org/10.1164/rccm.201106-1011OC

Supplementary Materials:



S1: Annual Average $PM_{2.5}$ Predictions over the region of Rome. Prediction data generated by Model 2.