VIETNAM NATIONAL UNIVERSITY, HANOI

UNIVERSITY OF ENGINEERING AND TECHNOLOGY

PHAM VAN HA

MODELING OF AIR POLLUTION IN VIETNAM USING MULTI-SOURCE AND MULTI-RESOLUTION DATA

Major: Information Systems

Code 9480104.01

DOCTOR OF PHILOSOPHY IN INFORMATION TECHNOLOGY

DISSERTATION

Hanoi - 2023

The work was completed at: VNU University of Engineering and Technology, Vietnam National University, Hanoi

Scientific supervisor:

Prof. Dominique Laffly

Assoc.Prof. Nguyen Thi Nhat Thanh

Reviewer: .....

.....

Reviewer: .....

.....

Reviewer: .....

.....

The thesis will be defensed before the doctoral admission Board of Vietnam National University, Hanoi meeting at .....

at h date month year

The thesis can be found at:

- National Library of Vietnam
- Library and Information Center Vietnam National University, Hanoi

PREFACE

Background and motivation of the study

The twenty-six session of the Conference of the Parties (COP26), The Prime Minister of Vietnam pledged to the international community to reduce stream emissions to zero by 2050. This is a strong commitment that demonstrates Vietnam's high political determination in the fight against climate change and environmental protection. For Vietnam, environmental pollution and climate change are very worrying. According to Yale University's 2020 ranking, Vietnam ranks 141st out of 180 countries in the Environmental Efficiency Index (EPI), of which Vietnam's air quality ranks 115th. Vietnam's rating is lower than the regional average and the world's.

In Vietnam, air quality monitoring is carried out at the central level by the Ministry of Natural Resources and Environment (MONRE). However, up to now, many stations have been shut out for a long time, many stations are not operating continuously. Due to the limited number of automatic and continuous air monitoring stations in Vietnam, $PM_{2.5}$ figures are also much less than other countries in the region. This problem also causes difficulties for monitoring and assessing the status of $PM_{2.5}$.

From these current conditions in Vietnam, the PhD student decided to choose the topic ***"Air quality modeling in Vietnam using multi-source, multi-resolution data".***

The Research objectives

The main objective of the dissertation is to study methods of air quality modeling in Vietnam using multi-source, multi-resolution data. Stemming from this objective, the specific objectives are including:

(i) Research and propose a method to estimate $PM_{2.5}$ map in Vietnam using numerical model
(ii) Research and propose a method to estimate $PM_{2.5}$ map in Vietnam using statistical model
(iii) Research and propose the methods for optimization of processing and modeling process

Dissertation contents

In addition to the introduction and conclusion, the composition of the dissertation is divided into 4 chapters: Chapter 1: Scientific background; Chapter 2: PM modeling over Vietnam using numerical model; Chapter 3: PM modeling over Vietnam using statistic model; Chapter 4: High Performance Computing Application in Pre-Processing and Modeling.

Dissertation Contributions

The contributions of the dissertation include:

First, the method of modeling PM maps from the proposed numerical model and evaluation using the WRF/Chem model. The main contributions of this section are presented in Chapter 2 and in the scientific publications **[PVH1**] and **[PVH5]**.

Second, the dissertation studied and evaluated georeference methods for satellite imagery and proposed the best method to apply to data preprocessing steps. The main content of this section is presented in chapter 3 and in **[PVH3]** and **[PVH8]**. This method is also applied to the input data in **[PVH6].**

Thrist, a multi-source aerosol data fusion approach is proposed to address the lack of data on satellite imagery due to factors such as cloud cover. The main contents of this section are presented in Chapter 3 and in scientific publication **[PVH7]**, this method is also applied in the process of processing input data to build models in **[PVH2].**

Finally, the dissertation has studied and applied techniques to improve data processing efficiency and modeling. The main results of this section are presented in Chapter 4 and some of them have been published in **[PVH3]**.

Chapter 1: SCIENTIFIC BACKGROUND

1.1. Introduction

PM pollution is determined based on measurements and calculations of particulate matter concentration (PM) to make an assessment of air quality. PM is a mixture of solid or liquid particles of different sizes and compositions that exist in the atmosphere. PM can be classified into PM_1, $PM_{2.5}$ or PM_{10} based on their aerodynamic diameter.

In this chapter, the dissertation will synthesize and re-system the basic knowledge of PM (the main research object in the dissertation), the effects of PM, PM monitoring methods from stations and from satellite images and methods and techniques for building PM maps from multi-source data. The contents will be presented from Sections 1.2 to 1.5 of chapter 1, some content is selected and presented in **[PVH4]** and **[PVH9].**

1.2. Particulate Matter (PM) Pollution overview

1.2.1. PM pollution in the world

1.1.2. PM pollution in Vietnam

1.3. Effects of PM pollution

PM pollution can have an impact on public health, the environment and climate change.

1.3.1. Impact on the environment and climate change

1.3.2. Impact on the public health

1.4. PM monitoring

1.4.1. PM monitoring from ground-base station

1.4.2. PM monitoring from satellite imagery

The use of satellite imagery data in PM pollution monitoring is a promising and much studied approach in recent years.

1.3.2.1. Satellites and aerosol products

1.3.2.2. Relationship between PM concentration and aerosol

1.5. PM map estimation techniques

The mapping methods are divided into 2 main groups: 1) using numerical models and 2) using a statistical model

1.5.1. Numerical model

The numerical model (specifically the CTM Chemical Transport Model) for building air quality maps is commonly used in a variety of ranges from global to regional levels.

1.5.2. Statistical model

For statistical methods, commonly used statistical models include: Linear Regression, Basic Component Analysis (PCA), Autoregressive Integrated Moving Average (ARIMA), Neural Network, Vector-Assisted Machine Learning, Deep Learning, etc.

1.6. Summary

This chapter presents the scientific literature, systematizes the theoretical basis of PM, PM impacts, PM monitoring methods and methods of building PM maps from numerical and statistical models. In the next chapter, the dissertation will present the contents of PM map modeling using numerical models.

Chapter 2: PM MODELING OVER VIETNAM USING NUMERICAL MODEL

2.1. Introduction

In this chapter, the dissertation will present an overview of the WRF/Chem model, related studies, the proposed method and the results of evaluating the WRF/Chem model to build PM maps. Methods for processing and updating emission data sets are proposed. The results of model evaluation and analysis of seasonal variations in PM concentration and fire effects are also detailed in contents from 2.2 to 2.7 of Chapter 2.

2.2. WRF/Chem air quality model

The Weather Research and Forecasting Model (WRF) is a next-generation medium-sized meteorological forecasting model system designed to cater to both atmospheric research and weather forecasting needs.

2.3. Related studies

2.3.1. Related studies in the world

2.3.2. Related studies in Vietnam

In Vietnam, the modeling method is now more commonly used, the problem of environmental pollution has been of concern, but the network of monitoring environmental measurement factors is not strong enough, so scientists have a lot of difficulty in analyzing and evaluating the status quo as well as forecasting environmental impacts due to pollution.

2.4. Problem statement

This dissertation presents the method of air quality simulation in the northern region of Vietnam using the wrf/chem model. This model was designed to simulate between January, February, March and June 2019. The emissions datasets used include HTAPV2 and REAS32. Because these emission datasets were developed for the base years of 2010 and 2015, the method of updating emissions data is applied using eclipse datasets. The simulation results are compared and evaluated with monitoring station data to assess the model's capabilities as well as the quality of different emission sets. Then, the best quality simulation results are used to assess the air quality variability above the seasonal area. The effects of monsoons and biomass burning on air quality are also considered.

2.5. Study area and data

In this dissertation, the WRF-Chem photochemical model is used to simulate $PM_{2.5}$ concentrations and other gas components orver Vietnam.

2.6. Proposed method

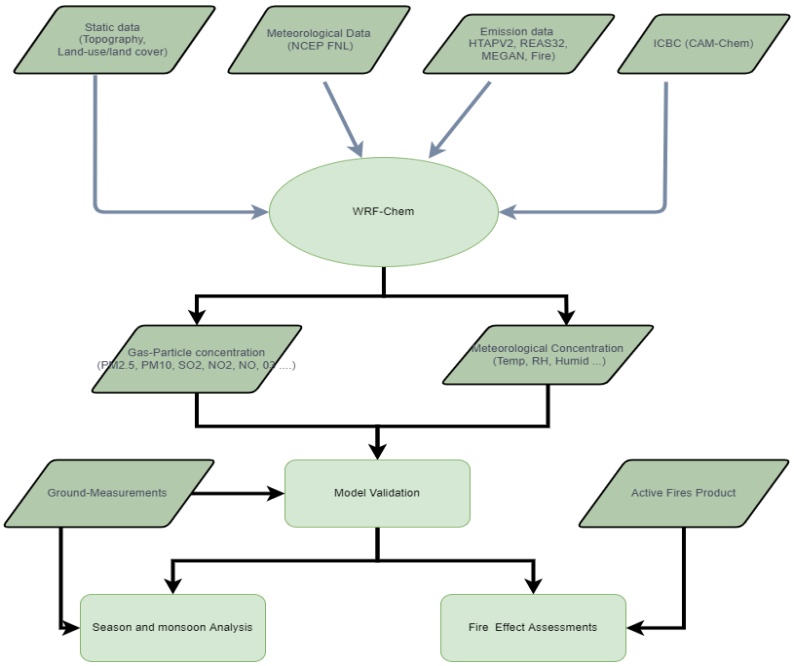

Figure 1. Modeling PM concentration using WRF-Chem model

2.6.1. Data preprocessing

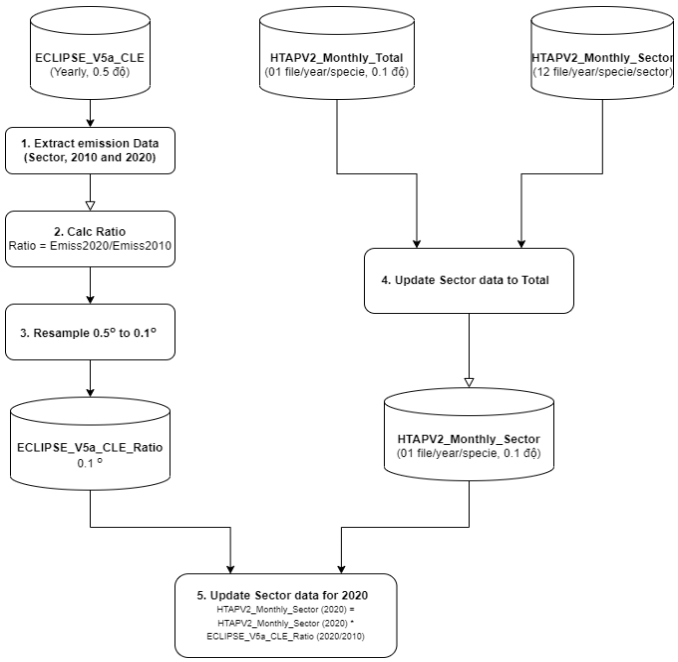

Figure 2. Emission update method

HTAPV2 data was developed for the base year of 2010, while the REAS3.1 data developed for the base year was 2015. Since ECLIPSE emissions scenario data are data sets every five years, the study used the ratio of 2020 to baseline years to update emissions data for simulated time in 2019.

2.6.2. WRF/Chem simulation

2.6.3. Model Evaluation

To carry out the quality assessment of PM simulation, PM parameters from the ground monitoring station in the northern region and at 03 fixed monitoring stations (Hanoi, Quang Ninh and Phu Tho) under the management of the Northern Center for Environmental Monitoring (NCEM), 02 monitoring stations of Hanoi Environmental Protection Agency. Data at the stations was collected during the simulation model period (January, February, 03 and June 2019).

2.6.4. Seasonal Analysis

To consider whether the model assessed seasonal changes in concentration, the average concentration in January and June 2019 from stations and model data was calculated. Simulation results from the REAS31_Update emission dataset are used to average the value from the model.

2.6.5. Fire impact assessment

Fire point products from MODIS and VIIRS NPP satellites during the simulation period were collected and assessed to affect $PM_{2.5}$ concentrations. Time-based analyses (daily and weekly) as well as spatial analyses are performed to assess the effects of fires.

2.7. Experimental results and discussion

2.7.1. Evaluation results

2.7.1.1. Evaluation of model simulation quality for different PM and gas components

Thus, considering the conditions of MFE and MFB errors, the WRF/Chem model simulates relatively well PM components such as PM_{10} and $PM_{2.5}$ while lower quality simulations for CO, NO and very poor simulations for substances such as NO_2, O_3 and SO_2.

2.7.1.2. Evaluation of model simulation quality for different emission dataset

From the above comments, it can be concluded that the REAS32 emissions dataset gives better simulation results than the HTAPV2 emissions dataset, which helps the MRE error decrease at most stations while the correlation is negligible.

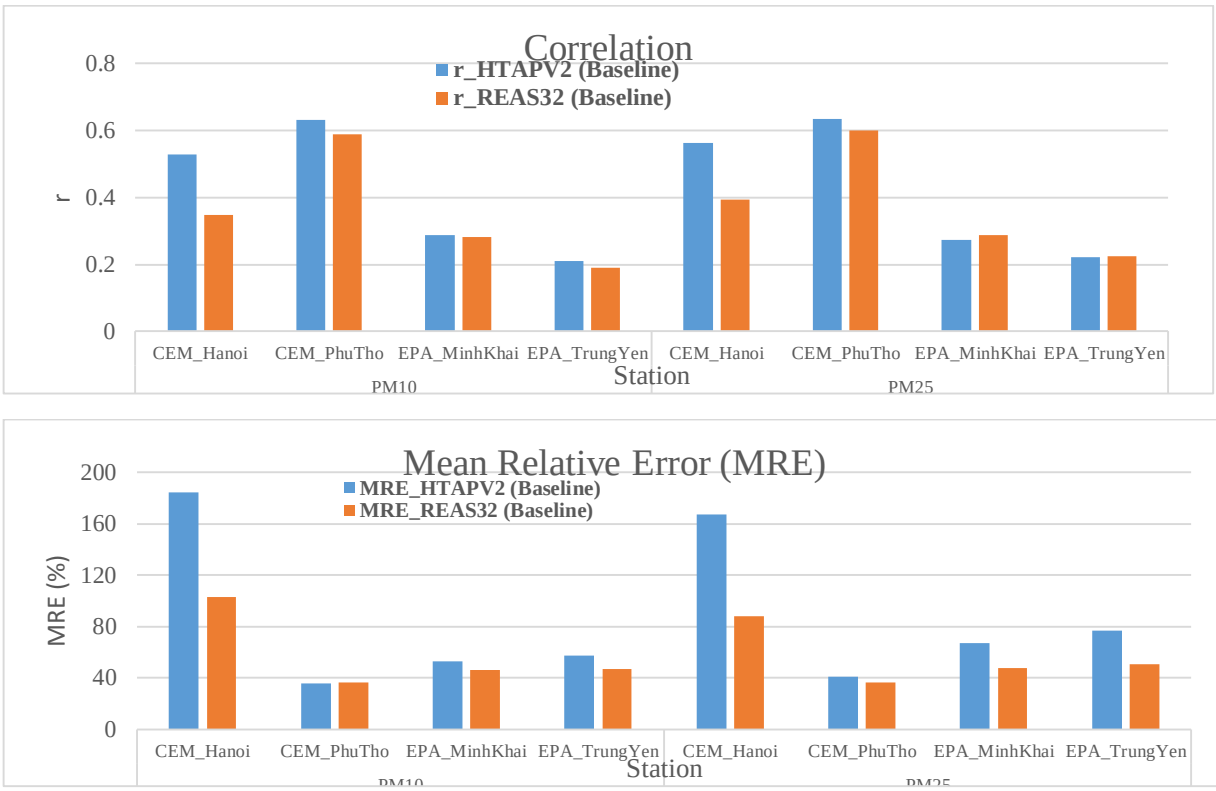

Figure 3. Correlation and relative error (MRE) of PM_{10} and $PM_{2.5}$ at stations and WRF/Chem simulation results

2.7.1.3. Evaluation of model simulation quality for updated emission dataset

As such, the above results show that the method of updating emissions data is remarkably effective with the HTAPV2 dataset when simulating PM_{10} and $PM_{2.5}$ concentrations. Meanwhile, the REAS32 dataset showed no improvement when applying the updated method, even making some substances inferior such as NO_2.

2.7.2. Seasonal variation and monsoon effect analysis

2.7.2.1. Results of seasonal variation analysis

As such, the results show that the simulation results from the WRF/Chem model both reflect the seasonal fluctuations and differences in which the dry season has high PM concentrations and the rainy season the PM concentration is lower (for both PM_{10} and $PM_{2.5}$).

2.7.2.2. Results of monsoon analysis

In terms of developments, PM concentrations tend to be low when there is no monsoon, high when there is a monsoon and low again when the monsoon is over, this statement is similar to the results of previous research of [8].

2.7.3. Active fire effects analysis

As such, analysis of PM concentration and the number of fire spots in the months shows that the distribution by space has a marked variation. The results showed that the effect of the fire point by space also differed between months. The results of the analysis also showed that $PM_{2.5}$ concentrations from the WRF/Chem model could fully reflect and explain the effects of fire both in space and time.

2.8. Summary

The simulation results of the model showed that WRF-Chem was able to capture changes in surface meteorological variables and PM concentrations in the atmosphere. The model meets the simulated target proposed by the US EPA for pm levels most at PM measuring stations. Considering the conditions of MFE and MFB errors, the WRF/Chem model simulates relatively well PM components such as PM_{10} and $PM_{2.5}$ while lower quality simulations for CO, NO and very poor simulations for substances such as NO_2, O_3 and SO_2.

Chapter 3: PM MODELING USING STATISTIC MODEL

3.1. Introduction

This chapter of the dissertation will present the contents related to 3 main issues: Geo-referencing satellite images, integrating multi-source aerosol data and building PM maps using statistical models. The detailed contents include related studies, data used and study area, proposed method, experimental results and evaluation. The contents are arranged and presented from sections 3.2 to 3.6 of this chapter.

3.2. Related studies

3.2.1. Georeferencing

Generally, there are two classes of rectification approaches. The parametric and the non-parametric approaches (Hemmleb and Wiedemann, 1997). Whereas for the parametric approach the knowledge of the interior and exterior orientation parameters are required, non-parametric approaches require just ground control points (GCPs). Non-parametric approaches include polynomial transformation, projective transformation.

3.2.2. Data fusion

There are two types of methods for data fusion data from various sensors. The first type of method can only combine values at one pixel when there is at least one sensor with an observed value, including the optimal interpolation method, the arithmetic mean method, and the weighting method, which estimates the maximum likelihood of MLE. In the second group of methods, the main technique applied is the Universal Kriging (UK) interpolation method. The UK shows strengths in geostatistical problems, interpolating values at one point based on the spatial correlation function for other locations.

3.2.3. PM modeling

The use of aerosol imaging data in PM pollution monitoring is a new and promising approach. Aerosol images can be integrated with PM monitoring data in pollution estimation models to increase the computational and predictive quality of these models. PM estimation methods vary widely, from Linear Regression (LR), Multiple Linear Regression to SVR (Support Vector Regression) and SOM (Self Organizing Map - SOM (Gupta and Christopher, 2009b; Hirtl et al., 2013; Yahi et al., 2011).

3.3. Problem statement

This chapter presents the method of PM estimation using statistical model in Vietnam. First, the georeference method is studied and tested on surface reflections from MODIS sensors to determine the best georeference method applicable to aerosol data. Next, aerosol data integration methods are studied and aerosol data integration methods from MODIS Terra, MODIS Aqua and VIIRS NPP satellite sensors are evaluated.

3.4. Study Area and Data

3.4.1. Study area

The area of study in this section is the whole of Vietnam.

3.4.2. Datasets

All examined datasets are summarized in **Error! Reference source not found.**

3.4.2.1. Georeferencing datasets

First, the assessment of several georeferencing methods is examined on MODIS Terra/Aqua and VIIRS NPP satellite radiances images over Vietnam areas. In addition, the effect of georeferencing methods on quality of satellite AOD data was examined using MODIS Terra/Aqua AOD product at 3 km and VIIRS NPP AOD at 6 km and AERONET AOD at 3 stations in Vietnam.

3.4.2.2. Fusion datasets

In this study, AOD data from the MODIS Terra/Aqua and VIIRS NPP satellites were used for integration to enhance coverage. In addition, AOD data from AERONET monitoring stations across Vietnam is also used to evaluate and verify satellite imagery products before and after integration. The data was collected between 2012 and 2016.

3.4.2.3. PM modeling datasets

Hourly average $PM_{2.5}$ concentration data was collected at nine continuous fixed monitoring stations nationwide between 2012 and 2016. Detailed information about the stations is presented in Table 2.1.

3.5. Proposed method

The method of building a PM concentration map using a statistical model is described in Figure 4.

Figure 4. Flow-chart of PM modeling using statistic model

3.5.1. Georeferencing method

In the first part, we use satellite radiance data from MODIS Terra/Aqua to assess the accuracy of different georeferencing method. GDAL has been used to implement georeferencing methods on radiance data. There are several transformation methods possible in GDAL including: Polynomial function, Thin Plate Spline and geolocation. In the second part, we implement georeferencing methods on MODIS and VIIRS AOD data. After that, the georeferenced images are validated to the AERONET AOD product at 3

stations in Vietnam including Nghia Do, Bac Lieu, Nha Trang. The validation method is described in our previous studies (Thanh T N Nguyen et al., 2015b; Vinh et al., 2018).

3.5.2. Data fusion method

Three different methods were used to combine aerosol data including Maximum Likelihood Estimation (MLE), Linear Regression and Geographically Weighted Regression (GWR).

3.5.2.1. Maximum Likelihood Estimation (MLE)

3.5.2.2. Linear Regression

3.5.2.3. Geographically Weighted Regression (GWR).

Similar to the Terra linear regression method, we experimentally integrated the data using the geo-weighted regression method. Since the study area is the entire territory of Vietnam, the relationship between AOD image pairs may be spatially altered due to changing conditions including meteorology and topography. Therefore, the geographically weighted regression method is experimentally compared with the linear regression method.

3.5.3. PM modeling method

In this section, a daily $PM_{2.5}$ concentration distribution map in Vietnam is based on the Mixed Effect Model using different data sources for the period 2012 – 2020. The methodology and results are summarized from the study of colleagues in the same research group of authors (Truong et al. 2022). Input data includes ground monitoring, satellite imagery products (AOD), meteorological maps (temperature, humidity, pressure, PBLH ...), land use maps (population density, traffic density, plant index).

3.6. Results and discussions

3.6.1. Georeferencing results

3.6.1.1. Accuracy assessment of georeferencing methods

The result shows that GDAL Polynomial function has low accuracy for both MODIS Terra and MODIS Aqua images. In another side, it could be seen that all of the other methods including GDAL TPS, GDAL Geoloc, and MRT have a good result. It could be seen that GDAL Geoloc has lower accuracy than GDAL TPS and MRT. GDAL TPS is the better method for widely processing different image sources.

3.6.1.2. Effect of georeferencing methods on data quality

Fig. 4 illustrates the correlation of selected MODIS and VIIRS NPP AOD images for each day. The result shows that MODIS and VIIRS AOD data processed by the TPS function has a better correlation than the P2 function for almost days.

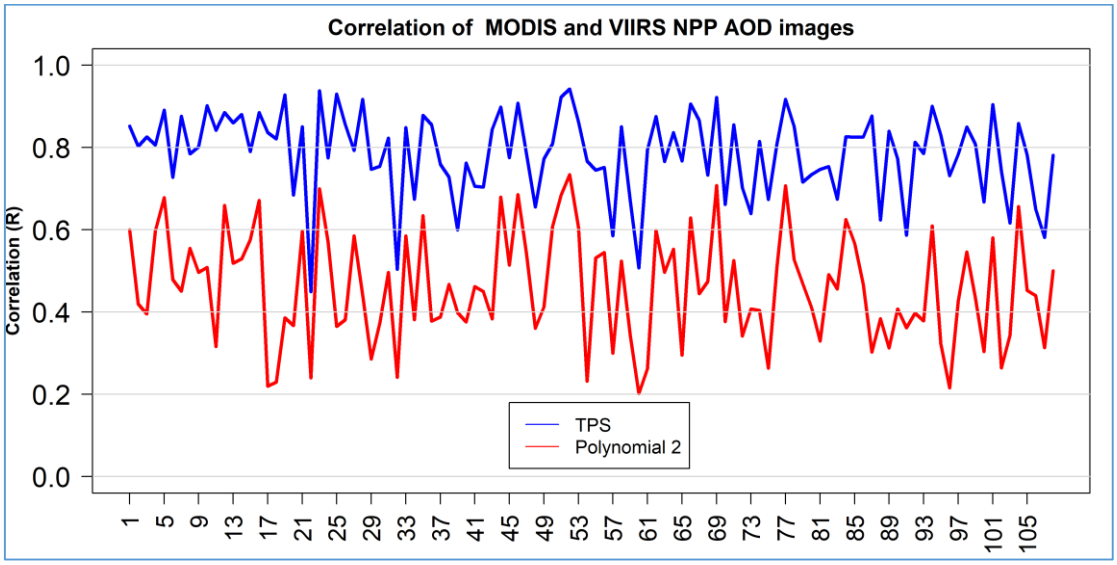

Figure 5. Correlation of selected MODIS and VIIRS NPP AOD images.

3.6.2. Data fusion results

3.6.2.1. Coverage ratio assessment

The rate of increase in data coverage of the combined image compared to the MODIS Aqua/Terra image is quite high by about 31-50%, compared to the low VIIRS image by about 1-7% (As shown in Table 4). Although the rate of increase in data coverage compared to VIIRS is not significant, if the data quality increases more than VIIRS, the integrated method still makes sense.

3.6.2.2. Evaluation of data quality

The results of the image evaluation integrated with the AERONET station of each method had the same sample number (680), the regression method of Terra data was more correlated than the MLE method. Specifically, the regression method to Terra has the highest coefficient R2 (0.81), the lowest RMSE error (0.219), the lowest nearest RE error (reaching 71,998). Hence choose this as the best day photo integration method of the methods.

Table 1. Quality of images before integration and after integration

Data	Samples	R2	RMSE	RE	MFB	MFE
Aqua	550	0.5824	0.2806	73.9229	-0.3506	0.4246
Terra	495	0.7594	0.2379	74.3250	-0.3400	0.4314
VIIRS	680	0.6018	0.2667	99.2505	-0.4113	0.5253
MLE	680	0.7511	0.2455	80.5099	-0.3343	0.4599
Terra Regression	680	0.8118	0.219	71.998	-0.3503	0.4435
GWR	680	0.72	0.302	38.290	-0.34	0.46

3.6.2. PM modeling results

Based on the selected parameters data, the MEM model was built to estimate the daily $PM_{2.5}$ map, including the first model that used the AOD parameter and the auxiliary model that did not use the AOD parameter (Truong, 2022). The above two MEM models are also performed 10 fold Cross Validation. Cross-examination results are lower than model results. The main model has cross-verification results with Pearson coefficient r reaching 0.7, R2 reaching 0.49, RMSE reaching 18.14 µg / m³, MRE reaching 52.71%. The auxiliary model also had better cross-examination results than the main model with Pearson coefficient r reaching 0.75, R2 reaching 0.56, RMSE reaching 16.64 µg/m³, MRE reaching 54.2% (Table 2).

Table 2. Results of model evaluation and cross-verification (Truong, 2022)

		N	r	R^2	RMSE (µg/m³)	MRE (%)
Validation	Main model	10,614	0.83	0.68	14.14	40.99
	Support model	34,208	0.81	0.65	14.8	48.58
10-fold cross validation	Main model	1,061	0.7	0.49	18.14	52.71
	Support model	3,420	0.75	0.56	16.64	54.2

3.7. Summary

In this study, different georeferencing have been examined on MODIS Aqua/Terra and VIIRS NPP radiance images over the Vietnam region. GADM data were utilized to assess the accuracy of these georeferencing methods. The assessment results show that TPS, Geolocation, and MRT are better than Polynomial transform. However, TPS is the best method for widely processing different satellite image sources. Secondly, several data fusion methods were applied on MODIS and VIIRS AOD images in Vietnam. The results show that the geographically weighted regression (GWR) method gives the best accuracy results while helping to increase the data coverage significantly and still ensure data quality.

Finally, daily $PM_{2.5}$ maps are estimated using two MEM models (one primary and one auxiliary) developed on a 10-year dataset (2012-2021) including $PM_{2.5}$ measured at ground stations, satellite AOD, meteorology, and land use maps. The main model had evaluation results with Pearson r = 0.83, R2 = 0.68, RMSE = 14.14 $\mu g/m^3$, MRE = 40.99% on 10614 samples.

Chapter 4: HIGH-PERFORMANCE COMPUTING APPLICATION IN PREPROCESSING AND MODELING

4.1. Introduction

In our research process, from the preprocessing data for the construction of a model, its execution involves many complicated steps and requires considerable implementation time. Therefore, in this chapter, the dissertation uses certain methods to improve the performance of a number of steps, including optimization of WRF/Chem models running on HPC, gereferencing of VIIRS optimization and data fusion optimization.

4.2. WRF/Chem Model Optimization on HPC

4.2.1. Objective

We install and run the model on HPC in order to test and select the optimal plan to improve the performance of the model.

4.2.2. Experimental simulation

4.2.3. Experimental results

As the number of MPI increases, the calculation time will not decrease but will increase. The reason for this is that when the number of MPI is small, the time needed to calculate the model will be much greater than the communication time between MPI. When a certain number is reached, the calculation time will not change while the communication time between the MPI will increase.

4.3. Georeferencing of VIIRS Optimization

4.3.1. Objective

In this section, we experiment with sampling VIIRS images in order to improve performance and maintain the required image accuracy.

4.3.2. Sampling method

Therefore, in this section, we experiment with sampling based on two different distributions: regular distribution and random distribution.

4.1.1. Evalution results

The results show that the error gradually decreases as the number of GCPs increases. Basically, the random sampling method gives smaller errors when the number of GCP points is greater than 1000 points. Meanwhile, with a GCP number of 144 corresponding to the number of GCPs on the MODIS image, the error of the reference method is about 6 km, which corresponds to the size of the VIIRS pixel. Thus, we can conclude that, by simply sampling 144 GCP points (corresponding to the number of GCP images of

MODIS images), the error of VIIRS images is acceptable (in the size of 1 pixel).

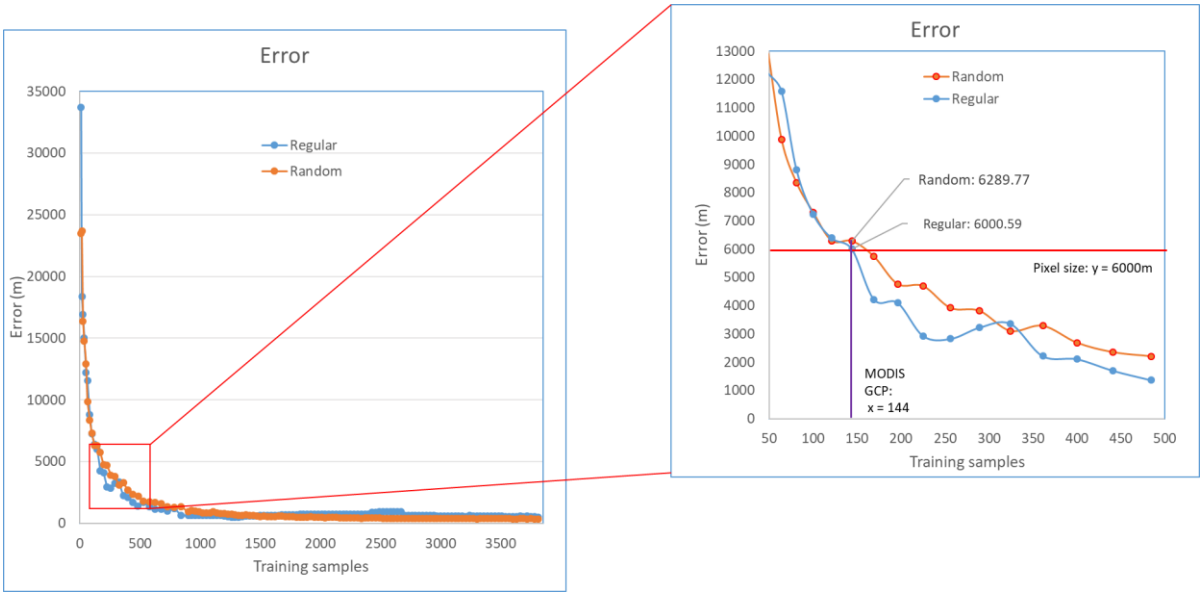

Figure 6. Georeferencing error of different training sample size

4.4. Building GWR on HUPI cloud platform

4.4.1. Objective

In this section, the cloud-based GWR algorithm is built and evaluated for efficiency compared to some published GWR libraries.

4.4.2. Methodology

As can be seen in the Golden Search optimization method, the steps of building and estimating the model are performed for each data point sequentially. Therefore, when the number of data points is large, the calculation time of the model is very long. Parallelization methods are proposed to parallelize the process of model construction and estimation.

4.4.3. Experimental result

Thus, the GWR algorithm on the HUPI platform does not show an improvement in runtime, but there is an improvement in the resources used to help the algorithm successfully run with large sample sets (From 100K and above).

4.5. GWR optimization using Genetic Algorithm

4.5.1. Objective

In the above section, the GWR algorithm is built on HUPI based on the Golden Search algorithm to find the optimal bandwidth. This algorithm will look for the bandwidth value in 1 interval [a, b] so that the Target Cross Validation Score function is the smallest. Instead of applying Golden Search, in this section the dissertation will evaluate the effectiveness of using genetic algorithms to be able to find the optimal bandwidth.

4.5.2. Problem statement

Input:

Given the geo-weighted regression model G = GWR(bw, X_n, Y_n) where:

The accuracy of the model is evaluated by the CV_Core (Cross-validation Core) parameter calculated as follows:

$$CV_{bw} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Output:

Apply genetic algorithms to search for bw bandwidth in the range [a, b] such that the value CV_Core is minimal.

4.5.3. Methodology

GAs are general techniques that help solve problems by simulating the evolution of humans or organisms in general (based on Darwin's theory of all species evolution), under predetermined conditions of the environment. The goal of GAs is not to provide the optimal exact solution but to provide a relatively optimal solution.

4.5.4. Experimental result

Detailed CV values for each image, each set of parameters are represented as the figure below.

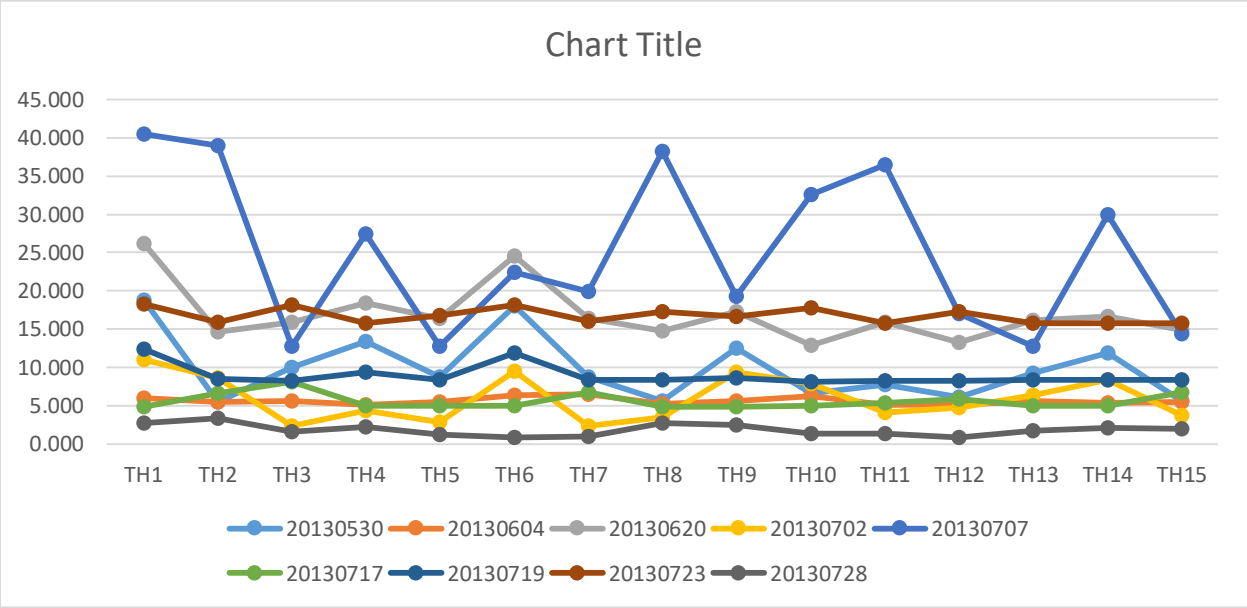

Figure 7. CV Score for different images

It can be seen that the best set of parameters is TH5, TH2, TH4 and TH6 also have relatively good results.

4.6. Summary

This chapter presents proposed solutions to optimize data processing and modeling. WRF/Chem model reviews on HPC have shown the optimal number of MPIs to shorten the best model runtime. The sampling method on VIIRS images shows that with an evenly spaced sampling method with a certain number of GCP, the georeference method fully meets the accuracy and significantly reduces processing time. The GWR algorithm on the HUPI platform also indicates that it solves the resource problem when running large numbers of samples while other libraries do not. Finally, the application of genetic algorithms to optimize the GWR model has shown the parameter set of genetic algorithms with the smallest CV_Score on the test set to help the algorithm converge. Some results of this section have been published in [PVH3].

CONCLUSION

The contribution of the dissertation

The dissertation presented a systematic study on PM modeling in Vietnam using multi-source, multi-resolution data.

Chapter 1 present the scientific literature, systematizes the theoretical basis of PM, PM impacts, PM monitoring methods and methods of building PM maps from numerical and statistical models. These

contents are a solid theoretical basis that addresses the research questions of the dissertation. Some content is selected and presented in [PVH4] and [PVH9].

Chapter 2 presents a method of modeling PM maps from the proposed numerical model and evaluation using the WRF/Chem model. The simulation results of the model showed that WRF-Chem was able to capture changes in surface meteorological variables and PM concentrations in the atmosphere. The model meets the simulated target proposed by the US EPA for pm levels most at PM measuring stations. Looking at the emission datasets, the REAS32 emission dataset gives better simulation results than the HTAPV2 emission dataset. The results of updating emission data show that the method of updating emissions data is remarkably effective with the HTAPV2 dataset when simulating PM_{10} and $PM_{2.5}$ PM concentrations. On the other hand, the results of seasonal analysis show that simulations from the WRF/Chem model both reflect seasonal fluctuations and differences in which the dry season has high PM concentrations and the rainy season lower PM concentrations (for both PM_{10} and $PM_{2.5}$). In addition, PM concentration is also affected by the monsoon, PM concentration tends to be low when there is no monsoon, high when there is monsoon and low when the monsoon is over. Finally, the analysis of the effect of the fire point showed that the correlation between the fire point in the Northern region was higher than that of the whole country, the weekly correlation was higher by the day. $PM_{2.5}$ concentrations from the WRF/Chem model can fully reflect and explain the effects of fire both in space and time. The main contributions of this section are presented in the scientific publications [PVH1] and [PVH5]. The product output of the model has also been used as the input of the PM estimation model in [PVH2].

Chapter 3 of the dissertation will present the contents related to 3 main issues: Geo-referencing satellite images, integrating multi-source aerosol data and building PM maps using statistical models. The detailed contents include related studies, data used and study area, proposed method, experimental results and evaluation. First, different georeferencing has been examined on MODIS Aqua/Terra and VIIRS NPP radiance images over the Vietnam region. The assessment results show that TPS, Geolocation, and MRT are better than Polynomial transform. However, TPS is the best method for widely processing different satellite image sources. TPS function gives a better quality on both MODIS Aqua/Terra and VIIRS AOD georeferenced images than P2 function. This study highlights the effect of preprocessing methods on quality of satellite data and give a suggestion on selecting a suitable georeferencing method to pre-process data before performing other studies. The main content of this section is presented in chapter 3 and in **[PVH3]** and **[PVH8]**. This method is also applied to the input data in **[PVH6].** Second, several data fusion methods were applied on MODIS and VIIRS AOD images in Vietnam. Data fusion methods generally increase data coverage and improve data quality. The GWR regression model shows the best correlation while MLE gives the lowest relative error. Therefore, depending of the application so that we can choose the appropriate data fusion method. The main contents of this section are presented in scientific publication **[PVH7]**, this method is also applied in the process of processing input data to build models in **[PVH2].**

Chapter 4 present proposed solutions to optimize data processing and modeling. WRF/Chem model reviews on HPC have shown the optimal number of MPIs to shorten the best model runtime. The sampling method on VIIRS images shows that with an evenly spaced sampling method with a certain number of GCP, the georeference method fully meets the accuracy and significantly reduces processing time. The GWR algorithm on the HUPI platform also indicates that, although the algorithm does not achieve runtime efficiency, it solves the resource problem when running large numbers of samples while other libraries do not. Finally, the application of genetic algorithms to optimize the GWR model has shown the parameter set of genetic algorithms with the smallest CV_Score on the test set to help the algorithm converge. The main results of this section have been published in **[PVH3]**.

The limitation of the Dissertation:

Some limitations can be mentioned are:

- PM modeling using numerical models only uses available global transmitters with rough analysis and little updates. Emission inventory in Vietnam from different sources are not used in the model.
- The results of the new model are moderate; the data assimilation techniques are not applied to improve the quality of the model using a combination of many different data sources.
- The developed HUPI-based GWR algorithm is not efficient in terms of execution time, it needs to be optimized to shorten the running time.

Future works:

From the results achieved in the dissertation as well as the remaining limitations, there are some research directions for the future works:

– Try to apply data assmilation techniques to improve modeling quality.

– Try to process and include emission inventory in Vietnam to model to update emission data.

– Optimization of GWR on HUPI cloud to shorten performance time.

LIST OF PUBLICATIONS

[PVH1] Van Ha Pham, Xuan Truong Ngo, Hieu Dang Trung Phan , Tuan Vinh Tran, Hoang Anh Le, Astrid Jourdan , Dominique Laffly , Thi Nhat Thanh Nguyen (2022). Emission comparison for WRF-Chem model and impact assessment of monsoon and active fire on air quality in the Northern Vietnam. Aerosol and Air Quality Research (In Review).

[PVH2] Truong X Ngo, **Ha V. Pham**, Hieu D.T. Phan, Anh T.N. Nguyen, To Thi Hien, Thanh T.N. Nguyen "A daily and fully fine Particulate Matter concentration dataset derived from space observations for Vietnam in the period of 2012-2020", Science of Total Environment (In Review)

[PVH3] Pham , V.H., Nguyen , T.N.T. and Laffly , D. (2020). Remote Sensing Case Studies. In TORUS 2 – Toward an Open Resource Using Services, D. Laffly (Ed.). doi:10.1002/9781119720553.ch7

[PVH4] Pham , V.H., Luu , V.H., Phan , A., Laffly , D., Bui , Q.H. and Nguyen , T.N.T. (2020). Remote Sensing Products. In TORUS 2 – Toward an Open Resource Using Services, D. Laffly (Ed.). doi:10.1002/9781119720553.ch4

[PVH5] Do, T.N.N., Ngo, X.T., **Pham, V.H**. et al. Application of WRF-Chem to simulate air quality over Northern Vietnam. Environ Sci Pollut Res (2020). https://doi.org/10.1007/s11356-020-08913-y

[PVH6] Thanh T.N. Nguyen, **Ha V. Pham**, Kristofer Lasko, Mai T. Bui, Dominique Laffly, Astrid Jourdan, Hung Q. Bui (2019). Spatiotemporal analysis of ground and satellite-based aerosol for air quality assessment in the Southeast Asia region. Environmental Pollution. https://doi.org/10.1016/j.envpol.2019.113106.

[PVH7] Pham Van Ha, Ngo Xuan Truong, Astrid Jourdan, Dominique Laffly, Nguyen Thi Nhat Thanh. Evaluation of Maximum Likelihood Estimation and regression methods for fusion of multiple satellite Aerosol Optical Depth data over Vietnam. The 11th International Conference on Knowledge and Systems Engineering (KSE 2019)

[PVH8] Pham Van Ha, Nguyen Thi Nhat Thanh, Bui Quang Hung, Pascal Klein, Astrid Jourdan, Dominique Laffly, Assessment of georeferencing methods on MODIS Terra / Aqua and VIIRS NPP satellite images in Vietnam. Proceeding of The 10th International Conference on Knowledge and Systems Engineering (KSE 2018), 01-03 November 2018, Ho Chi Minh City, Vietnam.

[PVH9] Thi Nhat Thanh Nguyen, Hoang Anh Le, Thi Minh Tra Mac, Thi Trang Nhung Nguyen, **Van Ha Pham**, Quang Hung Bui (2018). "Current Status of $PM_{2.5}$ Pollution and its Mitigation in Vietnam". Global Environmental Research, vol.22, no.1&2.