amU Aix Marseille Université

NNT : 2024AIXM0001

THÈSE DE DOCTORAT

Soutenue à Aix-Marseille Université

le 04 Décembre 2024 par

Elliot CHEVET

Modeling of atmospheric pollution due to marine traffic in Marseille

Discipline
Sciences pour l'ingénieur

Spécialité
Mécanique et physique des fluides

École doctorale
ED 353 Sciences pour l'ingénieur : mécanique, physique, micro et nanoélectronique

Laboratoire/ Partenaires de recherche
IRPHE
AtmoSud

Composition du jury

Chantal STAQUET Professeure, Université Grenoble Alpes	Présidente du jury
Luminita DANAILA Professeure, Université de Rouen Normandie	Rapporteuse
Riccardo MEREU Professeur Assistant, Politecnico di Milano	Rapporteur
Pietro SALIZZONI Professeur, École Centrale de Lyon	Examinateur
Olivier BOIRON Professeur, École Centrale Méditerranée	Directeur de thèse
Fabien ANSELMET Professeur, École Centrale Méditerranée	Co-directeur de thèse
Alexandre ARMENGAUD Responsable de la coopération internationale et scientifique, AtmoSud	Membre invité
Guillaume FELISAZ Conseiller technique, Association Cap au Nord Marseille	Membre invité

Centrale Méditerranée

AtmoSud
Inspirer un air meilleur

cnrs

Affidavit

I, undersigned, Elliot Chevet, hereby declare that the work presented in this manuscript is my own work, carried out under the scientific direction of Olivier Boiron and Fabien Anselmet, in accordance with the principles of honesty, integrity and responsibility inherent to the research mission.

The research work and the writing of this manuscript have been carried out in compliance with both the french national charter for Research Integrity and the Aix-Marseille University charter on the fight against plagiarism.

This work has not been submitted previously either in this country or in anyother country in the same or in a similar version to any other examination body.

Marseille, October 09, 2024



Liste de publications et participation aux conférences

1) Liste des publications et/ou brevets réalisées dans le cadre du projet de thèse :

1. Chevet, E., Boiron, O., and Anselmet, F. Modeling of air pollution due to marine traffic in marseille. Atmospheric Environment, 329:120542, 2024

2) Participation aux conférences et écoles d'été au cours de la période de thèse :

1. Conférence «Sustainable Solutions at Times of Transition», July 2022, Nisyros, Grèce
2. Conférence «Annual meeting of the American Physical Society Division of Fluid Dynamics», Novembre 2023, Washington DC

Résumé

La pollution de l'air est un enjeu persistant à Marseille et dans ses environs depuis des décennies, le trafic maritime étant de plus en plus reconnu comme l'un des principaux responsables de ce problème de santé publique. La proximité du port avec le centre-ville, associée aux émissions des nombreux navires à quai, y compris certains des plus grands navires de croisière du monde, a suscité de vives inquiétudes et un grand débat public. Les principaux polluants, à savoir les particules fines (PM_{10}, $PM_{2.5}$), le dioxyde d'azote (NO_2), le dioxyde de soufre (SO_2) et l'ozone (O_3), sont à l'origine de diverses pathologies, notamment l'irritation des voies respiratoires, les maladies cardiopulmonaires et le cancer du poumon.

Le travail de thèse présenté ici s'inscrit alors dans ce contexte, avec pour premier objectif la mise en place d'un modèle numérique permettant d'évaluer la contribution des navires à la pollution atmosphérique à Marseille. Le modèle utilisé pour représenter les nombreux phénomènes météorologiques et chimiques en interaction dans l'atmosphère est le modèle WRF-Chem. Afin d'assurer une représentation précise des émissions des navires, des données de trafic en temps réel provenant des autorités portuaires sont utilisées, contrairement aux modèles basés sur des données de trafic maritime annualisées. La configuration géographique unique de Marseille, entourée de vallées et bordée par la mer Méditerranée, est influencée par des phénomènes météorologiques spécifiques qui ont un impact direct sur la pollution de l'air dans la région. Une meilleure compréhension de ces mécanismes constitue le deuxième objectif de cette recherche, qui permettrait d'identifier les conditions météorologiques typiques favorisant des épisodes de forte pollution pendant les périodes de trafic maritime intense.

Fournir des prévisions quotidiennes sur la pollution atmosphérique est essentiel pour prendre des décisions au moment opportun afin de limiter les risques et informer la population directement touchée par la pollution. Cependant, le coût de calcul des simulations à haute résolution représente un défi pour les prévisions opérationnelles. Le troisième objectif du doctorat est de proposer une approche utilisant des modèles de Machine Learning (ML), supervisées par les résultats des simulations numériques du modèle WRF-Chem. En raison des grandes dimensions des données à traiter, les modèles classiques de ML sont peu efficaces. Des techniques de modélisation d'ordre réduit (ROM) ont donc été employées pour réduire les dimensions des données tout en conservant les informations essentielles. Les techniques de décomposition orthogonale propre (POD) et les autoencodeurs ont été combinées à un réseau neuronal récurrent (RNN) pour prédire les concentrations de $PM_{2.5}$.

Les principaux résultats de cette recherche doctorale sont : (i) La contribution des navires à la pollution de l'air à Marseille a été estimée entre 6 et 7% pour les $PM_{2.5}$, 4% pour les PM_{10} et 18% pour le NO_2; (ii) les particules fines sont transportées vers l'intérieur des terres, tandis que les particules plus grossières restent à proximité

des zones portuaires en raison de la sédimentation gravitationnelle; (iii) les vents faibles et la dynamique de la brise de mer favorisent l'accumulation des polluants au-dessus de la ville; et (iv) le développement d'un modèle multivarié POD-LSTM a significativement amélioré la précision des prévisions à long terme par rapport à l'approche univariée.

Mots clés : Dispersion de polluants, Modélisation numérique, Pollution maritime, Apprentissage automatique.

Abstract

Air pollution has been a persistent issue in Marseille and its surroundings for decades, with maritime traffic increasingly recognized as a key contributor to this public health concern. The port's proximity to the city center, combined with emissions from numerous ships at berth, including some of the world's largest cruise ships, has raised significant concerns and sparked extensive public debate. The pollutants of interest in this work, which include particulate matter (PM_{10}, $PM_{2.5}$), nitrogen dioxide (NO_2), sulfur dioxide (SO_2) and ozone (O_3), are responsible for respiratory diseases such as airway irritation, cardiopulmonary and lung cancers.
The PhD work presented here takes place in this context, with the first objective of setting up a numerical model to assess ships contribution to atmospheric pollution in Marseille. The model used to represent the many meteorological and chemical phenomena interacting in the atmosphere is the WRF-Chem model. To ensure an accurate representation of ship emissions, real-time traffic data from port authorities are used, in contrast to models based on annualized maritime traffic data.
Marseille's unique geographical configuration, surrounded by valleys and bordered by the Mediterranean Sea, is influenced by specific phenomena that have a direct impact on air pollution in the region. A better understanding of these mechanisms constitutes a second objective of this research, as it would help to identify typical meteorological conditions that favor higher pollution levels during periods of intense maritime traffic.
Providing daily atmospheric pollution forecasts is crucial for timely decision-making to mitigate the risk and informing the population directly affected by pollution. However, the computational cost of high-resolution simulations presents challenges for operational forecasting. The third objective of the PhD is addressing these challenges by proposing a Machine Learning (ML) approach, supervised by numerical simulation results from the WRF-Chem model. Due to the high dimensionality of the data, traditional ML models face learning difficulties. To mitigate this, reduced-order modeling (ROM) techniques were applied to reduce data dimensions while retaining key information. Both Proper Orthogonal Decomposition (POD) and autoencoders were combined with a recurrent neural network (RNN) to predict $PM_{2.5}$ concentrations.
The main findings of this PhD research are: (i) ship pollution in Marseille The contribution of ships to air pollution in Marseille is estimated at 6 to 7% for $PM_{2.5}$, 4% for PM_{10}, and 18% for NO_2; (ii) fine particles are transported inland while coarser particles remain near port areas due to gravitational settling; (iii) weak winds and sea-land breeze dynamics contribute to pollutant accumulation over the city; and (iv) the development of a multivariate POD-LSTM model significantly enhances long-term prediction accuracy compared to the univariate method.

Keywords: Pollutant dispersion, Numerical modeling, Ship pollution, Machine Learning.

Acknowledgements

Je souhaite remercier sincèrement mes directeurs de thèse, Olivier Boiron et Fabien Anselmet, pour leur accompagnement et leurs conseils précieux tout au long de ce travail. Un grand merci à l'ensemble des membres du jury pour l'intérêt qu'ils portent à ce travail. Merci également à Sonia Oppo et Alexandre Armengaud de AtmoSud qui m'ont été d'une grande aide pour la réalisation de ce travail. Les échanges avec Guillaume Felisaz m'ont permis de mieux comprendre les enjeux liés à ma thèse, ce qui a grandement contribué à ma motivation. Je tiens aussi à exprimer ma gratitude à mes collègues et amis de IRPHE pour leur soutien, ainsi qu'à mes proches pour leur patience et leur présence constante durant ces années.

Contents

1 Introduction **1**

1.1 Atmospheric Pollution . . . 1

1.1.1 Earth's Atmosphere . . . 1

1.1.2 Pollutants Overview . . . 2

1.1.3 Global and European Impact . . . 5

1.1.4 Air Quality Guidelines . . . 6

1.1.5 Air Quality Index . . . 7

1.2 Marseille: A Mediterranean Port City . . . 9

1.2.1 Marseille and its surroundings . . . 9

1.2.2 Micro-climate . . . 10

1.2.3 A Port City . . . 13

1.2.4 Air Quality surveillance in Marseille . . . 14

1.3 Modeling Atmospheric Pollution . . . 18

1.3.1 Spatial Scales . . . 18

1.3.2 Numerical Methods . . . 19

1.3.3 WRF-Chem . . . 23

1.4 Objectives of Research . . . 31

2 Modeling of Atmospheric Pollution due to Marine Traffic with WRF-Chem **33**

2.1 Introduction . . . 33

2.1.1 Marine Traffic and Air Pollution . . . 33

2.1.2 Methodologies for Ship Emissions . . . 35

2.2 Model Set-Up . . . 38

2.2.1 Study Area . . . 38

2.2.2 Numerical Methods . . . 39

2.2.3 Parameterization . . . 43

2.2.4 Initial and Boundary Conditions . . . 46
2.3 Marine Traffic in Marseille . . . 47
2.4 Emissions . . . 49
2.4.1 Ship Emissions . . . 50
2.4.2 Background Emissions . . . 52
2.5 Model Validation . . . 57
2.5.1 Validation on Meteorology . . . 58
2.5.2 Validation on Pollution . . . 61
2.6 Real Case Study . . . 64
2.6.1 Case Study . . . 64
2.6.2 Results . . . 66
2.6.3 Conclusion on real case study . . . 72
2.7 Influence of Meteorological Conditions . . . 73
2.7.1 Model Set-Up . . . 73
2.7.2 Ship Traffic and Emission Scenario . . . 73
2.7.3 Time-slot Selection . . . 74
2.7.4 Results . . . 76
2.7.5 Conclusion on the influence of Meteorological Conditions . . . 83
2.8 Discussions . . . 84

3 Modeling of Atmospheric Pollution with Machine Learning 87
3.1 Introduction . . . 87
3.1.1 From Numeric to Machine Learning . . . 87
3.1.2 Machine Learning Forecasting . . . 87
3.1.3 Reduced Order Modeling . . . 92
3.2 POD-LSTM for Air Pollution Forecasting . . . 94
3.2.1 Methodology . . . 94
3.2.2 Test on Potential Temperature . . . 98
3.2.3 Application to $PM_{2.5}$ . . . 103
3.2.4 Multivariate Model . . . 107
3.2.5 Conclusion on POD-LSTM . . . 110
3.3 Autoencoder-LSTM . . . 113
3.3.1 Autoencoder and POD . . . 113
3.3.2 Model Architecture . . . 114
3.3.3 Results . . . 116
3.3.4 Conclusion on Autoencoder-LSTM . . . 116

3.4 Discussions 117

4 Conclusion **121**
4.1 Summary 121
4.2 Perspectives 123

Appendix A Land Use **125**

Appendix B Ship Categories **127**

Appendix C NO_2 and SO_2 Ship Contribution **129**

Appendix D Features Correlations **131**

Appendix E Publication **133**

References **135**

List of Figures

1.1 Vertical structure and zonation of the atmosphere, showing the temperature profile to 100-km altitude. Note the logarithmic decline in pressure (left axis) as a function of altitude. Source: Schlesinger and Bernhardt (2020). 2
1.2 Deaths attributable to ambient particulate matter pollution in 2015. Source: Cohen et al. (2017). 6
1.3 Left: Map of France. Right: Population density in and around Marseille. Source: geoportail.gouv.fr, INSEE data 2017. 10
1.4 Monthly average temperature in Marseille between 2018 and 2021. Data from AtmoSud Longchamp station. 11
1.5 Wind Rose of Marseille from 2019 to 2022. Data from 7 GPMM stations and 2 AtmoSud sations (Longchamp & Verneuil). 12
1.6 Density map of marine traffic in west Mediterranean sea. Source: European Commission. 13
1.7 Map of the port of Marseille-Fos. Top left: map of the west basin. Top right: map of the east basin. Bottom middle: location of the two basins. Source: marseille-port.fr. 15
1.8 Annual concentration evolution since 2000 in % for main pollutants measured at AtmoSud stations. Source: AtmoSud (2022). 16
1.9 Atmospheric processes computed in a Eulerian grid box. Source: Jacob (1999). 21
1.10 Movement of a fluid element along an air flow trajectory. Source: Jacob (1999). 22
1.11 ARW Modeling System Flow Chart. Source: Skamarock et al. (2008) 24
1.12 Eta vertical coordinate system. Source: Skamarock et al. (2008) 25
1.13 Integration of the acoustic time step in the RK3 scheme 28
1.14 Arakawa C-grid. Source: Skamarock et al. (2008) 28
1.15 Interactions between physics categories in WRF. 29

2.1 Map of the domain of interest 39

2.2 Map of the three nested domains 40
2.3 Mean elevation of vertical levels 41
2.4 Mean elevation of the first 10 levels 41
2.5 GT distribution in Marseille (in black) and Fos (in red). Y-axis is in log scale. 49
2.6 Average number of calls per hour. 50
2.7 24-hour averaged $PM_{2.5}$ estimated ship emissions from 2nd and 3rd levels for 2021-07-01 in $\mu g \cdot m^{-2} \cdot s^{-1}$ (log scale). 52
2.8 Ship emission per month. Min and Max daily average values in blue. Standard deviation in red. 53
2.9 Ship emission by category. (a) represents weekday average and (b) represents hourly average. 54
2.10 2D interpolation method using area intersection. 55
2.11 24-hour averaged $PM_{2.5}$ background emissions for 2021-04-22 in $\mu g \cdot m^{-2} \cdot s^{-1}$ (log scale). High emissions in the western port correspond to the refineries and chemical plants of Fos-sur-Mer. 55
2.12 Share of emission sectors in the DOI 57
2.13 T2 absolute difference between WRF and Arome for 2022-09-10 at 7 PM (UTC+2). 59
2.14 Time series of WRF, Arome and Observations from Marignane for 2022-09-05. ($Wind = \sqrt{U^2 + V^2}$) 60
2.15 Locations of the 27 AtmoSud monitoring stations (red dots) with terrain height in m represented by the background color. The red star indicates the Longchamp station. 62
2.16 Time series of concentrations at Longchamp station for MOZAIC (in back), MOZCART (in blue), and Obs (in red) in $\mu g.m^{-3}$ from 2021-05-31 to 2021-07-01. Shaded areas represent the standard deviation of daily concentration values. 64
2.17 Scatter plots of MOZAIC predictions vs. Observations at Longchamp station from 2021-05-31 to 2021-07-01. The red dashed line marks the 1:1 line. 65
2.18 Average wind speed (in $m \cdot s^{-1}$) and wind roses at 6 different locations on D03 at first vertical level from 2021-05-31 to 2021-07-01. 67
2.19 Ship contribution to SO_2 and NO_2 from 2021-05-31 to 2021-07-01. 69
2.20 Ship contribution to O_3 from 2021-05-31 to 2021-07-01. 69
2.21 Ship contribution to $PM_{2.5}$ and PM_{10} from 2021-05-31 to 2021-07-01. 70

2.22 Ship-emitted particles concentration (left column) and number (right column) for the 4 bins from 2021-05-31 to 2021-07-01. (a), (b), (c) and (d) share the same color scale in $\mu g.m^{-3}$. (e), (f), (g) and (h) have their own color scale in $kg^{-1}_{\text{dry air}}$. 71
2.23 Stokes number for the 4 size bins over D03. 72
2.24 24-hour averaged $PM_{2.5}$ estimated ship emissions from ship traffic scenario in $\mu g \cdot m^{-2} \cdot s^{-1}$ (log scale). 75
2.25 Windrose of 2021 from Marignane meteo station 76
2.26 72-hour averaged contribution of ships to $PM_{2.5}$ concentration in % for weak winds. 77
2.27 72-hour averaged contribution of ships to $PM_{2.5}$ concentration in % for moderate winds. 78
2.28 Differences between all time slots and **W2** in ship-emitted $PM_{2.5}$ concentration in $\mu g \cdot m^{-3}$. First and second rows correspond to weak and moderate winds respectively. 80
2.29 Top: map of the 5 shorelines; Bottom: 72-hour averaged $PM_{2.5}$ concentration levels along the shorelines for time-slot **Su1**. Borders of Fos: 4.80°E to 5.00°E ; Borders of Marseille: 5.26°E to 5.51°E 81
2.30 In black: Time series of $PM_{2.5}$ ship-emitted average concentration over Marseille; In red: Time series of average scalar product between the sea-breeze vector and wind direction over Marseille 83

3.1 Basic concept of a neuron (left) and a recurrent neuron (right) 89
3.2 Illustration of a forward pass through a LSTM Cell. The input sequence is composed of 4 steps X_{t_i}, while only one output Y_{t_3} is produced, which typically corresponds to the predicted time step in the context of time series prediction. 90
3.3 Concept of a LSTM Cell. 90
3.4 The POD-LSTM architecture. 94
3.5 Time serie of the mean perturbation potential temperature in K. Each color corresponds to a time block. 98
3.6 POD cumulative mode energy (in black) and relative projection error (in red) for θ'. Vertical red dashed line represents mode 32. 99
3.7 MSE on temporal coefficients of the 32 modes. 100
3.8 Relative prediction error for θ' (in black) and for mean field (in red). 95% confidence interval is represented in shaded grey. 101

3.9 Snapshots of original field (left column) and predicted field (right column) in K starting from 2021-08-11 12:00:00. 102
3.10 Box plots of $PM_{2.5}$ concentration in $\mu g.m^{-3}$ for each block. 103
3.11 POD cumulative mode energy (in black) and relative projection error (in red) for $PM_{2.5}$ concentration. Vertical red dashed line represents mode 32. 104
3.12 Energy spectrum of the POD modes with wavelength $\lambda = \frac{2\pi}{k}$ in km. Colors indicate the mode number. $E(k)$ for mode 32 is in red, while the vertical dashed line indicates its maxima. 105
3.13 MAE in $\mu g \cdot m^{-3}$ for each pair $(N_{\text{inp}}, N_{\text{out}})$. 106
3.14 Relative prediction error for $PM_{2.5}$ concentration (in black) and for mean field (in red). 95% confidence interval is represented in shaded grey. 107
3.15 Cumulative mode energy of features from the projection on $PM_{2.5}$ POD basis. 109
3.16 Relative prediction error for $PM_{2.5}$ concentration (in black) and for mean field (in red). 95% confidence interval is represented in shaded grey. 111
3.17 Snapshots of original field (left column) and predicted field (right column) in $\mu g \cdot m^{-3}$ starting from 2021-08-11 12:00:00. 112
3.18 Principle of an Autoencoder. 113
3.19 nMOR architecture. 115
3.20 Relative prediction error for $PM_{2.5}$ concentration (in black) and for mean field (in red). 95% confidence interval is represented in shaded grey. 117

A.1 21-MODIS land use categories on domain D03. 125

C.1 72-hour averaged contribution of ships to NO_2 concentration in % for weak winds. 129
C.2 72-hour averaged contribution of ships to NO_2 concentration in % for moderate winds. 129
C.3 72-hour averaged contribution of ships to SO_2 concentration in % for weak winds. 130
C.4 72-hour averaged contribution of ships to SO_2 concentration in % for moderate winds. 130

D.1 131

E.1 First page of publication in the journal of Atmospheric Environment. 133

List of Tables

1.1 WHO air quality guidelines (AQGs) and estimated reference levels (RLs) . . 7
1.2 ICAIRh Thresholds 8
1.3 ICAIRh Index 8

2.1 MOZAIC setup. 38
2.2 Spatial description of the WRF simulation domains. 40
2.3 Maximum stable Courant numbers for one-dimensional linear advection. Source: Wicker and Skamarock (2002). 42
2.4 General call statistics registered at the port of Marseille-Fos in 2021. Docking Time (in hour) and GT are the average for each ship type. 48
2.5 Assumptions for the ship stack height. Source: AtmoSud 51
2.6 Yearly emissions by activity sector. D03 was divided into two areas by a longitudinal line at 5.10°E. W3 and E3 zones are respectively to the west and east of the demarcation line. 56
2.7 Comparison between WRF and Arome from 2022-09-05_00:00:00 to 2022-09-18_23:00:00 59
2.8 Errors of WRF and Arome with respect to observations from Marignane from 2022-09-05 to 2022-09-18 61
2.9 MOZCART Evaluation from 2021-05-31 to 2021-07-01 using AtmoSud network. # Total and # Marseille rows gives the number of stations available for the period and the pollutant considered within the domain D03 and the sub-domain of Marseille respectively. MAE and RMSE are in $\mu g.m^{-3}$. 63
2.10 MOZAIC Evaluation from 05-31-2021 to 07-01-2021 using AtmoSud network. # Total and # Marseille rows gives the number of stations available for the period and the pollutant considered within the domain D03 and the sub-domain of Marseille respectively. MAE and RMSE are in $\mu g.m^{-3}$. 63

2.11 Call statistics registered at the port of Marseille-Fos from 2021-05-31 to 2021-07-01. Docking Time (in hour) and GT are the average for each ship type. +/- column refers to the percentage point difference with monthly average (i.e. 8.33% of total calls for 2021). 66

2.12 Contribution of ships to pollutant concentrations from 2021-05-31 to 2021-07-01. Max and %_Max are respectively the maximum value and contribution of the time-averaged ship-emitted concentrations. D03 and %_D03 are respectively the spatio-temporal ship-emitted concentration and contribution within the DOI. Mrs and %_Mrs are respectively the spatio-temporal ship-emitted concentration and contribution within the sub-domain of Marseille, which is defined in 2.1. 68

2.13 Ship traffic scenario 74

2.14 Weather conditions for each time slot over 2021 76

2.15 Contribution of ships to pollutants concentration 79

2.16 $PM_{2.5}$ Concentration levels along distances from shore for time-slot **Su1**. Tot. is for total domain. Marseille and Fos are splitting shorelines in longitude according to the borders represented in Fig.1.7. Outside refers to the remaining longitudes of the domain. 80

3.1 LSTM architecture. 97

3.2 Results of POD-LSTM with 32 POD modes. MAE and RMSE are in $\mu g \cdot m^{-3}$. E_r^{72} is the average relative prediction error over a 72h horizon. 106

3.3 Results of MPOD-LSTM with 32 POD modes. MAE and RMSE are in $\mu g \cdot m^{-3}$. E_r^{72} is the average relative prediction error over a 72h horizon. 110

3.4 Convolutional encoder (left) and decoder (right) filter sizes and strides. Source: Gonzalez and Balajewicz (2018). 115

3.5 Results of the nMOR model. MAE and RMSE are in $\mu g \cdot m^{-3}$. E_r^{72} is the average relative prediction error over a 72h horizon. 116

3.6 Computational costs required for training and inference processes for a 72-hour horizon time of the POD-LSTM and nMOR. Times are non-dimensional according to the time required for WRF-Chem simulations as $\frac{\tau_{\mathrm{ML}}}{\tau_{\mathrm{WRF\text{-}Chem}}}$. . . . 118

A.1 21-MODIS Land use categories and associated parameters (Summer and Winter values). Parameters from left to right are: albedo (ALBD %), soil moisture availability (SLMO $\times 100\%$), surface emissivity (SFEM $\times 100\%$), surface roughness length (SFZ0 $\times 10^{-2}$m), thermal inertia (THERIN $\times 4.184 \times 10^2 \mathrm{J} \cdot \mathrm{m}^{-2} \cdot \mathrm{K}^{-1} \cdot \mathrm{s}^{-1/2}$), and surface heat capacity (SFHC $\times 10^6 \mathrm{J} \cdot \mathrm{m}^{-2} \cdot \mathrm{K}^{-1}$). . . . 126

List of Abbreviations

AASQA	Association Agréée de Surveillance de Qualité de l'Air
AE	Autoencoders
AIS	Automatic Identification System
ANN	Artificial Neural Networks
AQGs	Air Quality Guidelines
AQI	Air Quality Index
ARIMA	Auto-Regressive Integrated Moving Average
Arome	Météo-France numerical weather prediction model
ARW	Advanced Research WRF
BC	Black Carbon
CAE	Convolutional Autoencoder
CENAQ	Connexion électrique des navires à quai
CFD	Computational Fluid Dynamics
CIGALE	Consultation d'Inventaires Géolocalisés Air-cLimat-Energie
CNN	Convolutional Neural Network
CTM	Chemistry Transport Model
DNN	Deep Neural Networks
DOI	Domain Of Interest

EEA ..... European Environment Agency

EMEP ..... European Monitoring and Evaluation Programme

EPA ..... Environmental Protection Agency

EQIS-PA ..... Évaluation Quantitative d'Impact sur la Santé de la Pollution de l'Air

FDDA ..... Four-Dimensional Data Assimilation

GDAS/FNL .... Global Data Assimilation System (Final)

GHG ..... Greenhouse Gas

GOCART ..... Goddard Chemistry Aerosol Radiation and Transport

GPMM ..... Grand Port Maritime de Marseille

GT ..... Gross Tonnage

ICAIRh ..... Indice Cumulé de l'Air horaire

IMO ..... International Maritime Organization

LES ..... Large Eddy Simulation

LNG ..... Liquid Natural Gas

LSM ..... Land-Surface Model

LSTM ..... Long-Short Term Memory

MAE ..... Mean Absolute Error

MDO ..... Marine Diesel Oil

MEGAN ..... Model of Emission Gases and Aerosols from Nature

MGO ..... Marine Gas Oil

ML ..... Machine Learning

MODIS ..... Moderate Resolution Imaging Spectroradiometer

MOSAIC ..... Model for Simulating Aerosol Interactions and Chemistry

MOZAIC ..... Combination of MOZART with MOSAIC

MOZART ..... Model for Ozone and Related Chemical Tracers

MOZBC ..... MOZART Boundary Conditions

MOZCART ..... Combination of MOZART with GOCART

MPOD-LSTM .. Multivariate POD-LSTM

MSE ..... Mean Squared Error

NCEP ..... National Center for Environmental Prediction

NMB ..... Normalized Mean Bias

NME ..... Normalized Mean Error

nMOR ..... Non-linear Model Order Reduction

NO_2 ..... Nitrogen dioxide

NOAA ..... National Oceanic and Atmospheric Administration

NP ..... Nominal Power

NWP ..... Numerical Weather Prediction

O_3 ..... Ozone

OC ..... Organic Carbon

PACA ..... Provence-Alpes-Côte d'Azur

PBL ..... Planetary Boundary Layer

$PM_{2.5}$ ..... Particulate matter with a diameter $\leq 2.5\ \mu m$

PM_{10} ..... Particulate matter with a diameter $\leq 10\ \mu m$

POD ..... Proper Orthogonal Decomposition

POD-LSTM .... Univariate coupling of POD and LSTM

PS ..... Pressure at Sea level

r ..... Pearson correlation coefficient

RH ..... Relative Humidity at 2 m

RK ..... Runge-Kutta

RL ..... Reference Level

RMSE ..... Root Mean Squared Error

RNN ..... Recurrent Neural Network

RO ..... Residual Oil

ROM ..... Reduced-Order Modeling

RPE ..... Relative Projection Error

RRTMG ..... Rapid Radiative Transfer Model

SECA ..... Sulfur Emission Control Area

SO_2 ..... Sulfur dioxide

SOA ..... Secondary Organic Aerosols

ST ..... Steam Turbine

SVD ..... Singular Value Decomposition

T2 ..... Temperature at 2 m

TKE ..... Turbulent Kinetic Energy

UCM ..... Urban Canopy Model

UFP ..... Ultra-Fine Particle

VOC ..... Volatile Organic Compound

WHO ..... World Health Organization

WPS ..... WRF Preprocessing System

WRF ..... Weather Research and Forecasting

WRF-Chem ..... WRF model coupled with Chemistry

1

Introduction

1.1 | Atmospheric Pollution

1.1.1 | Earth's Atmosphere

The Earth's atmosphere is a complex system of gases on the surface of the globe. It is mainly composed of nitrogen (78.09%), oxygen (20.95%), argon (0.93%) and carbon dioxide (0.04%). Structured in distinct layers according to temperature gradient and composition, the atmosphere consists of 5 main layers. The troposphere, closest to the Earth's surface, is where most of the atmosphere's mass is concentrated. The phenomena discussed in the present manuscript primarily occur in this layer. Typically, temperature decreases by approximately 6.5 K per kilometer of altitude until reaching the tropopause, the boundary separating the troposphere from the stratosphere. The latter is characterized by the presence of the ozone layer, as indicated in Fig.1.1, which absorbs ultraviolet radiation from the Sun. Beyond the stratosphere lies the mesosphere, followed by the thermosphere, and ultimately the exosphere, which represents the outermost layer that extends from about 600 km up to 10 000 km.

A sub-layer of the troposphere deserves to be introduced here, given its central role in the context of the present work. Often referred to as the Planetary Boundary Layer (PBL), it constitutes the lowest part of the troposphere and is characterized by turbulent winds due to the strong influence of friction, as well as vertical mixing. Above the PBL lies a region characterized by a more uniform vertical distribution and typically stronger winds, known as the free atmosphere. The thickness of the PBL fluctuates widely, as it is directly influenced by radiative forcing from the earth's surface. Typically, the PBL extends to $\sim$ 1 km above ground. Some physical processes at stake in the PBL are detailed thereafter (Section 2.2.3.4).

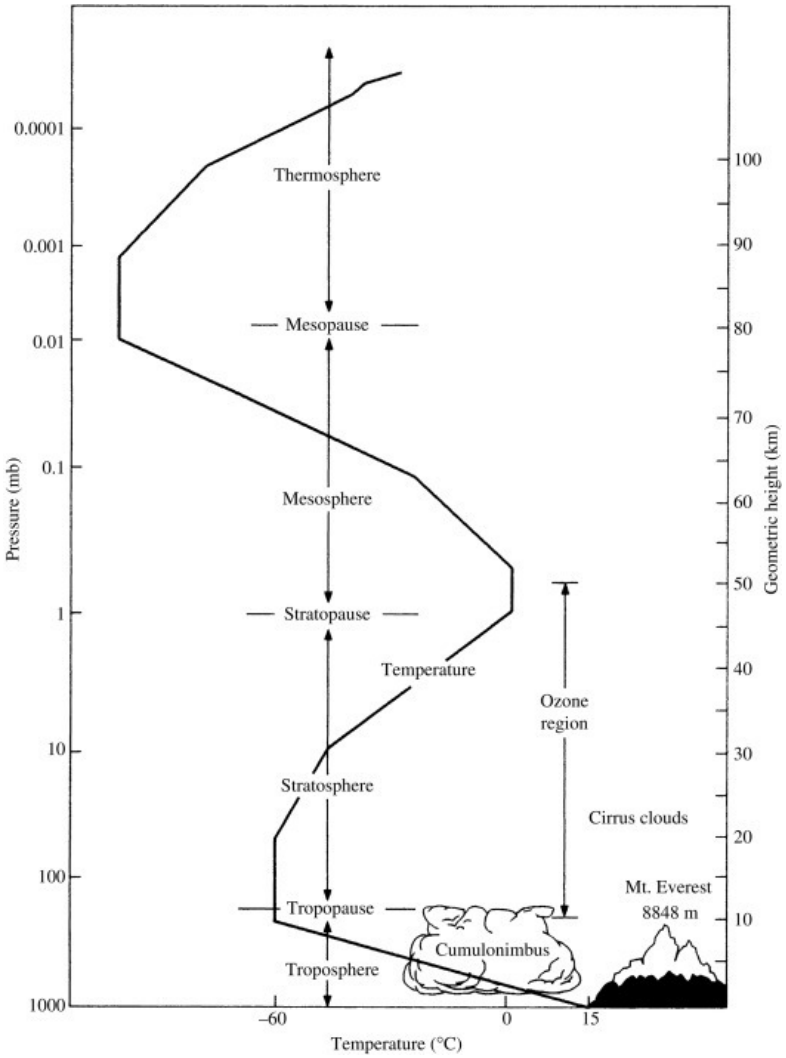

Figure 1.1: Vertical structure and zonation of the atmosphere, showing the temperature profile to 100-km altitude. Note the logarithmic decline in pressure (left axis) as a function of altitude. Source: Schlesinger and Bernhardt (2020).

1.1.2 | Pollutants Overview

There is often confusion about whether a specific substance qualifies as an air pollutant. For instance, greenhouse gases (GHGs) such as CO_2 and CH_4 might be considered air pollutants depending on the context. Here, we classify a substance as an air pollutant if it poses a direct risk to human health, thus excluding GHGs. They can either be emitted through human activities, known as anthropogenic emissions, or occur naturally, referred

to as biogenic emissions. Certain chemical processes in the troposphere can transform substances into pollutants. As these resulting pollutants are not directly emitted by a source, they are called secondary pollutants, as opposed to primary pollutants. The following sections outline the major pollutants in terms of health risks.

1.1.2.1 | Particulate Matter

Particulate matter (PM), also known as aerosols, refers to any solid or liquid particle in suspension in the air. These particles vary in size, composition and origin. They can be emitted directly into the atmosphere from sources such as vehicles, ships, industrial processes, construction activities and wildfires, or they can form through chemical reactions, known as secondary organic aerosols (SOA). They include a wide range of substances, such as dust, pollen, soot, smoke, sea salt, liquid droplets, etc... PM are categorized based on their size, with PM_{10} (particles with a diameter $\leq 10\ \mu m$) and $PM_{2.5}$ (particles with a diameter $\leq 2.5\ \mu m$) being the most documented ones. In recent years, there has been increasing attention on Ultra-Fine Particles (UFPs), characterized by a diameter $\leq 0.1\ \mu m$, as their impact on health might have been underestimated so far (Qi et al., 2024). As discussed in Section 1.1.3, PM have the strongest impact on health compared to other pollutant, causing numerous respiratory and cardiovascular diseases. For this reason, we pay particular attention to these compounds throughout this work.

1.1.2.2 | SO_2

Sulfur Dioxide or SO_2 is a gas-phase pollutant. The combustion of sulfur-containing fossil fuels, such as coal and oil, in industrial processes and power plants are generally the main sources. In port areas, ship traffic can be a significant source of SO_2 if heavy fuel is used. In addition, volcanic eruptions and certain natural processes can also release a significant amount of SO_2 into the atmosphere. Sulfur dioxide impacts both human health and the environment. When released into the atmosphere, it can react with other compounds to form sulfuric acid and sulfate aerosols, which contribute to the formation of acid rain and smog. Exposure to sulfur dioxide can irritate the respiratory system, causing symptoms such as coughing, wheezing, shortness of breath and chest tightness. It can also lead to the exacerbation of respiratory conditions such as asthma and bronchitis.

1.1.2.3 | NO_2

Nitrogen dioxide or NO_2 is a gas-phase pollutant as well, mainly resulting from the photo-chemical reaction of NO with other atmospheric compounds. As NO_2 is mostly

a secondary pollutant, we typically refer to nitrogen oxides (NO_x) rather than NO_2 when discussing emissions. NO_x emissions are primarily due to combustion processes in vehicles, ships, power plants and industrial facilities. NO_2 irritates airways and can aggravate cardiopulmonary and respiratory diseases, as well as allergic responses. NO_x emissions can lead to the formation of nitric acids as well, causing acid rain.

1.1.2.4 | Ozone

Ozone or O_3 is a gase-phase species already naturally present in large quantities, mainly in the stratosphere, as mentioned above. While ozone plays a vital role in our planet's radiative balance, it is considered a pollutant at low altitudes, due to its effects on health, when it reaches our lungs. Ozone is a secondary pollutant, mainly formed by photochemical reactions. As a result, episodes of high ozone concentration are more likely to occur on sunny days, typically in summer. Exposure to ozone can result in symptoms such as coughing and a scratchy throat. Moreover, it can exacerbate existing lung conditions like asthma, emphysema and chronic bronchitis.

1.1.2.5 | VOCs

Volatile Organic Compounds or VOCs are gas-phase pollutants. VOCs include a variety of organic molecules, such as hydrocarbons like benzene (C_6H_6) and toluene (C_7H_8), as well as other organic compounds like formaldehyde (CH_2O). They are characterized by their high vapor pressure at ambient temperature and low solubility in water. They are predominantly synthetic chemicals employed in various industries such as paint manufacturing, pharmaceuticals and refrigeration. Widely used as ingredients in household products, VOCs concentration is generally higher indoor than outdoor. Outdoor emissions can originate from vehicles, ships, industries etc... Trees and vegetation emit a variety of VOCs, with isoprene (C_5H_8) being one of the most significant. Exposure to VOCs can result in various health effects, such as irritation of the eyes, nose and throat, headaches, nausea and potential damage to the liver, kidneys or even the central nervous system.

1.1.2.6 | Others

While some air pollutants will receive less attention in this manuscript, they are still worth mentioning.
Among them, carbon monoxide or CO is known to be fatal in the event of a gas leak in a home, but is also emitted by vehicles, in the industrial and residential sectors. However,

outdoor CO concentrations are rarely reaching high levels.
Heavy metals form another class of pollutants. Characterized by their high density and potential toxicity at low concentrations, these compounds include arsenic (As), cadmium (Cd), chromium (Cr), nickel (Ni), lead (Pb) and zinc (Zn). They are both carcinogenic and toxic, with the potential to accumulate in vital organs like the liver, kidneys, and bones. Additionally, they can have adverse effects on the respiratory system.
Poly-cyclic Aromatic Hydrocarbons (PAH) are a group of organic compounds composed of multiple aromatic rings. Although there are a multitude of PAH, benzo(a)pyrene ($C_{20}H_{12}$ or BaP) is currently the only one subject to regulatory values. It is a particulate pollutant, generally measured in PM_{10}. It is mainly produced by the combustion of firewood, green waste and fossil fuels. PAH are associated with adverse health effects, including cancer, respiratory problems, and developmental abnormalities.

1.1.3 | Global and European Impact

Humans have always polluted the atmosphere. An interesting study of McConnell et al. (2018), based on measurement campaigns in Antarctica and numerical models, shows that observed variations in past lead emissions correspond to specific periods in our evolution: "European lead emissions closely varied with historical events, including imperial expansion, wars, and major plagues. Emissions rose coeval with Phoenician expansion, accelerated during expanded Carthaginian and Roman mining primarily in the Iberian Peninsula, and reached a maximum under the Roman Empire. Emissions fluctuated synchronously with wars and political instability particularly during the Roman Republic, and plunged coincident with two major plagues in the second and third centuries, remaining low for $>$ 500 years. Bullion in silver coinage declined in parallel, reflecting the importance of lead–silver mining in ancient economies.". If lead is an effective historical marker of our pollution, it is nowadays an example of the effectiveness of large-scale regulations. Since the enactment of various regulations, notably the Clean Air Act in 1970, there has been a substantial decrease in lead emissions. According to a 2016 report from the European Monitoring and Evaluation Programme (EMEP), lead emissions in Europe have dropped by 90% since 1990 (Colette et al., 2016).
Unfortunately, despite these regulations, air pollution was still the fifth-leading mortality risk factor according to the Global Burden of Desease 2015 (GBD 2015 Mortality and Causes of Death Collaborators, 2016). The global burden of ambient air pollution has even increased between 1990 and 2015, with 4.2 million estimated premature deaths from

exposure to $PM_{2.5}$ in 2015 against 3.5 millions in 1990. Figure 1.2 provides an insightful overview of the global health impacts of particulate matter. In most of the countries, the proportion of deaths attributable to PM pollution is higher than 3%, reaching over 8.6% in most of Southern and East Asia.

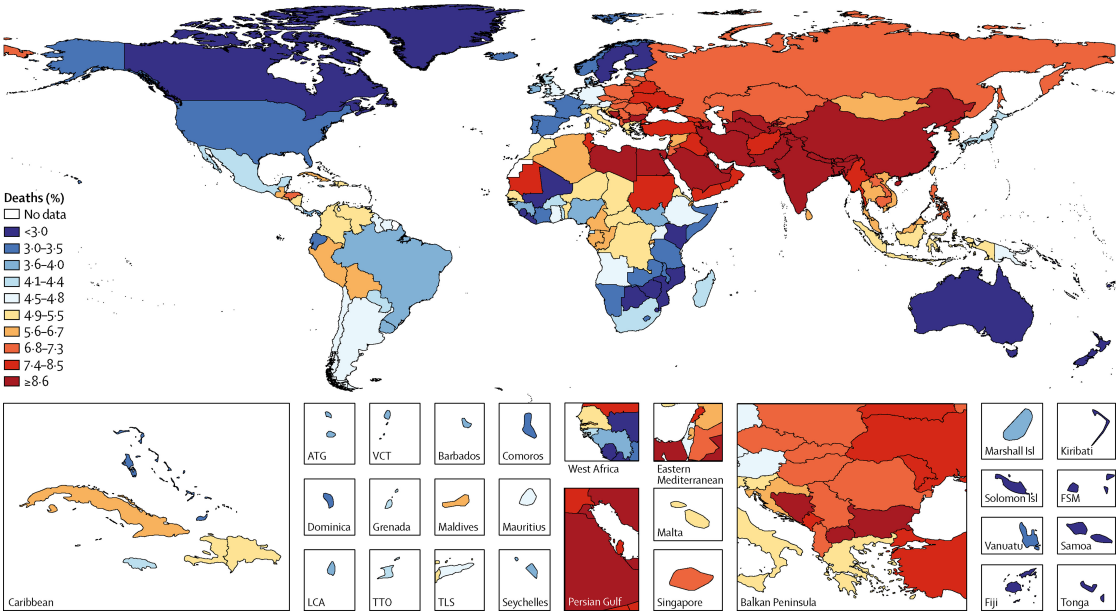

Figure 1.2: Deaths attributable to ambient particulate matter pollution in 2015. Source: Cohen et al. (2017).

In Europe in 2020, the estimated number of premature death from exposure to $PM_{2.5}$, NO_2 and O_3 is 253 000, 52 000 and 22 000 respectively, as reported by the EEA (European Environment Agency, 2022).

1.1.4 | Air Quality Guidelines

While pollution has been a persistent problem since the dawn of civilizations, large-scale air quality regulations are relatively recent in our history. The first regulations appeared in response to extreme events, such as the "Great smog" episode in London in 1952, which led to the creation of the Clean Air Act in 1956. The United States followed suit with the U.S Clean Air Act in 1963, administered by the U.S Environmental Protection Agency (EPA) since 1970. The first global air quality guidelines (AQGs) were established by the World Health Organization (WHO) in 1987. Since then, WHO has regularly updated and expanded these guidelines to take account of advances in scientific knowledge and

the evolution of air pollution-related health problems. The latest updates, displayed in Table 1.1, were published in 2021.

Table 1.1: WHO air quality guidelines (AQGs) and estimated reference levels (RLs) [a]

Pollutant	Averaging period	AQG	RL	Comments
PM_{10}	1 day	45 $\mu g.m^{-3}$	-	3-4 exceedance days per year
	calendar year	15 $\mu g.m^{-3}$	-	
$PM_{2.5}$	1 day	15 $\mu g.m^{-3}$	-	3-4 exceedance days per year
	calendar year	5 $\mu g.m^{-3}$	-	
O_3	Maximum daily 8-h mean	100 $\mu g.m^{-3}$	-	3-4 exceedance days per year
	peak season [b]	60 $\mu g.m^{-3}$	-	
NO_2	1 hour	200 $\mu g.m^{-3}$	-	
	1 day	25 $\mu g.m^{-3}$	-	3-4 exceedance days per year
	calendar year	10 $\mu g.m^{-3}$	-	
SO_2	10 minutes	500 $\mu g.m^{-3}$	-	
	1 day	40 $\mu g.m^{-3}$	-	3-4 exceedance days per year
CO	1 hour	30 $mg.m^{-3}$	-	
	Maximum daily 8-hour mean	10 $mg.m^{-3}$	-	
	1 day	4 $mg.m^{-3}$	-	3-4 exceedance days per year
BaP	calendar year	-	0.12 $ng.m^{-3}$	
C_6H_6	calendar year	-	1.7 $\mu g.m^{-3}$	
Pb	calendar year	0.5 $\mu g.m^{-3}$	-	
As	calendar year	-	6.6 $ng.m^{-3}$	
Cd	calendar year	5 $ng.m^{-3}$ [c]	-	
Ni	calendar year	-	25 $ng.m^{-3}$	

[a] As WHO has not set an AQG for BaP, C_6H_6, As and Ni, the RL was estimated assuming an acceptable risk of additional lifetime cancer risk of approximately 1 in 100 000.

[b] Average of daily maximum 8-hour mean concentration in the six consecutive months with the highest six-month running average O_3 concentration.

[c] AQG set to prevent any further increase of Cd in agricultural soil, likely to increase the dietary intake of future generations.

1.1.5 | Air Quality Index

Air quality has become a daily feature of most weather applications. The Air Quality Index (AQI) is a common way of communicating air quality to the population. Actually, the definition of the AQI differs from one institution to another. However, the common idea is to propose an index based on the concentrations of the main pollutants, often PM_{10}, $PM_{2.5}$, O_3, NO_2, SO_2, in order to provide a universally understandable assessment of air quality. In this report, we will take as reference the AQI of the Région Sud Provence-

Alpes-Côte d'Azur (PACA), of which Marseille is the main city. This AQI was defined by AtmoSud, the regional AASQA (Association Agréée de Surveillance de Qualité de l'Air). The AQI ICAIRh (Indice Cumulé de l'Air) is an hourly index, based on thresholds S_x for each pollutants showed in Table 1.2. A sub-index I_p is defined per pollutant p:

$$I_p = x + \frac{[P] - 0.6 - S_{\mathrm{x}}}{S_{\mathrm{x+1}} - S_{\mathrm{x}}} \tag{1.1}$$

with $[P]$ the concentration of pollutant p, and x such that $S_x < [P] < S_{x+1}$ ($x = 5$ if $[P] > S_5$). The AQI is then calculated:

$$ICAIRh = max(I_{PM_{10}}, I_{PM_{2.5}}) + I_{NO_2} + I_{O_3} \tag{1.2}$$

An assessment of air quality and a color is associated with the index value, as shown in Table 1.3

Table 1.2: ICAIRh Thresholds

Threshold	S_0 $\mu g.m^{-3}$	S_1 $\mu g.m^{-3}$	S_2 $\mu g.m^{-3}$	S_3 $\mu g.m^{-3}$	S_4 $\mu g.m^{-3}$	S_5 $\mu g.m^{-3}$
$PM_{2.5}$	0	20	40	50	100	150
PM_{10}	0	40	80	100	200	300
NO_2	0	40	90	120	230	340
O_3	0	50	100	130	240	380

Table 1.3: ICAIRh Index

ICAIRh	color	
0		Good
2		Fair
3		Moderate
4		Poor
10		Very poor
20		Extremely poor

1.2 | Marseille: A Mediterranean Port City

1.2.1 | Marseille and its surroundings

1.2.1.1 | History

The founding myth of Massalia, the ancient name for Marseille, tells of the encounter between a Greek sailor named Protis and a Gallic princess called Gyptis. Her father, Nannos, who was the king of the Segobriges, gifted the land of Massalia to them as a wedding present. From its founding around 600 BC until its capture by Caesar in 49 BC, Massalia was an independent Greek city. Although the city lost its political independence to the Romans, Greek culture persisted over time. It served as a major hub for redistributing Mediterranean products along the coasts of Gaul and Iberia. It wasn't until the end of the 15th century that the town became part of the kingdom of France. The port of Marseille served as a valuable asset for the kings of France, who exploited it for their expeditions to Italy. Under the reigns of Louis XIII and Louis XIV, the town tripled in surface area. The Second Empire triggered a significant urban boom in Marseille, marked by the development of railroads, the creation of new harbor basins in the north, and major urban planning projects. By the turn of the 20th century, Marseille's expansion was evident as the city opened up to the Mediterranean and beyond, fueled by steam-powered shipping companies and its connections with the colonial world. History reveals the evolution of a city built around its central lifeline: the port. This remains true today, with the port's growing activity making it the 6th largest in Europe by tonnage.

1.2.1.2 | Geography

Marseille, located in southeastern France, is the principal city of the Provence-Alpes-Côte d'Azur region. Covering a surface area of 240.62 km^2, it is one of France's largest cities. The Mediterranean Sea borders Marseille to the south and east. About 15 km northeast of Marseille lies Fos-sur-Mer, an important industrial zone in the region.

1.2.1.3 | Demography

In 2021, Marseille has a population of 873 076, making it the second most populous city in France, behind Paris. Fig.1.3 shows population density in the area. The overall population density in Marseille is 3617 $hab \cdot km^{-2}$, (INSEE 2020).

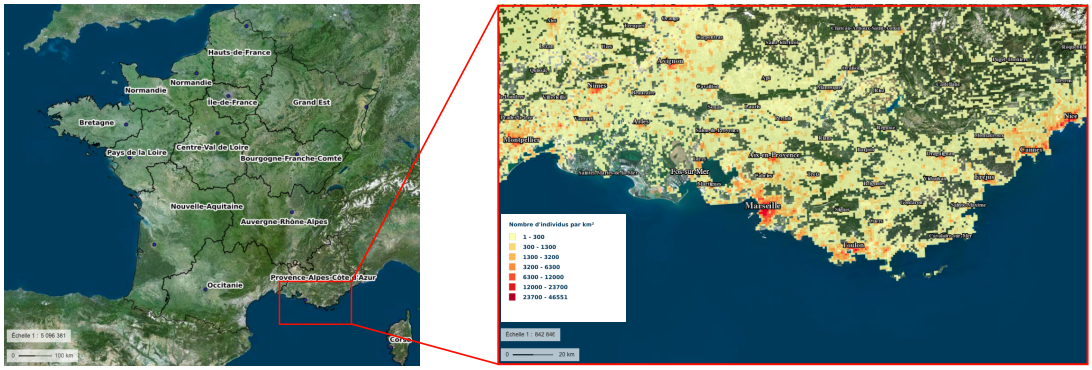

Figure 1.3: Left: Map of France. Right: Population density in and around Marseille. Source: geoportail.gouv.fr, INSEE data 2017.

1.2.2 | Micro-climate

1.2.2.1 | Topography

Marseille's weather is largely influenced by the surrounding topography. The city is located in an area where the effects of two main valleys converge. The Rhône valley to the north, wedged between the Massif Central and the Alps, and the Durance valley to the northeast (Bastin et al., 2005). On a more local scale, Marseille is surrounded by a multitude of hills, such as Mont Puget, Marseilleveyre, Saint-Cyr, Sainte-Baume, Garlaban, L'étoile and L'Estaque. The valley of l'Huveaune between Saint-Cyr and Garlaban hills might also influence the local weather.

1.2.2.2 | A Mediterranean Climate

Marseille's climate has its own particular characteristics. Firstly, Marseille is the sunniest city in France, receiving on average 2898 hours of sunshine annually (infoclimat.fr). Summers in Marseille are characterized by hot and dry weather, with temperatures averaging over $21°C$ from June to September, as shown in Fig.1.4. Summer rainfall in Marseille is among France's lowest, typically totaling just 1 to 3 days of rain per month from June to August (meteofrance.fr). However, the region is prone to Mediterranean episodes, characterized by thunderstorm systems and heavy rainfall. These episodes are generated when a large mass of warm, moisture-laden air migrates inland from the Mediterranean Sea (Duffourg and Ducrocq, 2011). On average, these episodes occur three to six times a

year. Winters are generaly mild, averaging around $10°C$ during the months of December, January, and February.

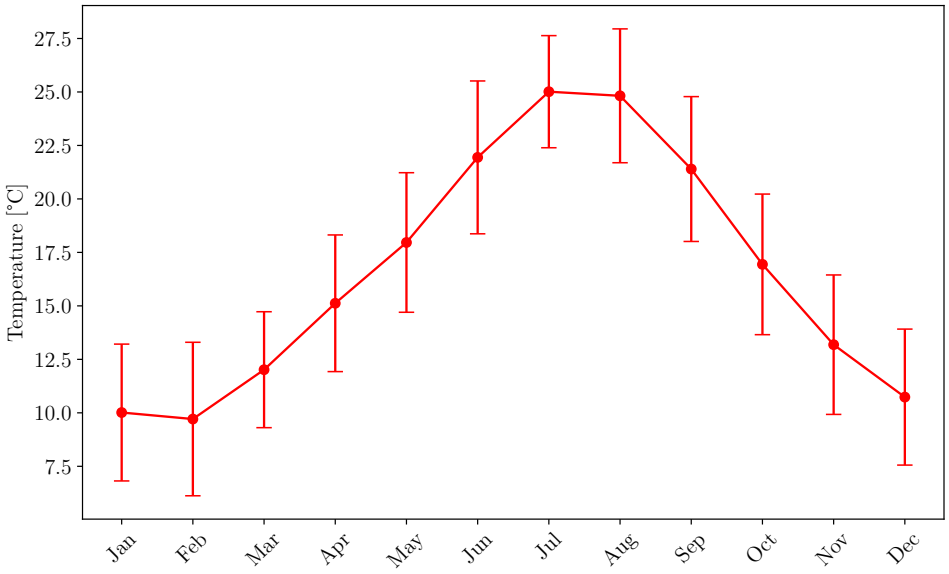

Figure 1.4: Monthly average temperature in Marseille between 2018 and 2021. Data from AtmoSud Longchamp station.

1.2.2.3 | Wind Regimes

The presence of distinct wind patterns is one of the main characteristics of Marseille's weather. The predominant wind is the Mistral, which originates from the Rhône Valley, as previously mentioned. This regime is clearly observed from the wind rose of Marseille in Fig.1.5. Such winds are typically cold and dry, with violent gusts reaching up to 100 $km \cdot h^{-1}$.

The southeast winds are generally attributed to the Marin. Often generated by a depression in the Gulf of Lion, the Marin is typically very humid and can sometimes lead to episodes of heavy rainfall, such as the Mediterranean episodes mentioned earlier. These winds are more frequently observed in spring and autumn. Less frequently, southeast winds sometimes originate from the Sahara, known then as the Sirocco. These episodes are easily identifiable, as the air is laden with fine sand particles.

Sea-land breeze is observed in Marseille when synoptic winds are weak. They occur due to temperature differences between the land and the sea. During the day, the sun heats

up the land faster than the sea. As the land gets hotter, the air above it warms up and rises as warm air is lighter. This process generates a low-pressure zone over the land, which induces the inflow of cooler air from the sea to replace the rising warm air. This movement of air from the sea to the land is known as sea breeze. At night, the process reverses, as the land cools down faster than the sea, making the air above the sea warmer than the air above the land. A land breeze is then generated. The bay of Marseille forms a semi-enclave with a SW opening. Thus, sea breezes are from the SW, and land breezes are from the NE. These breezes are well represented on the Wind Rose in Fig.1.5, mainly at low speeds.

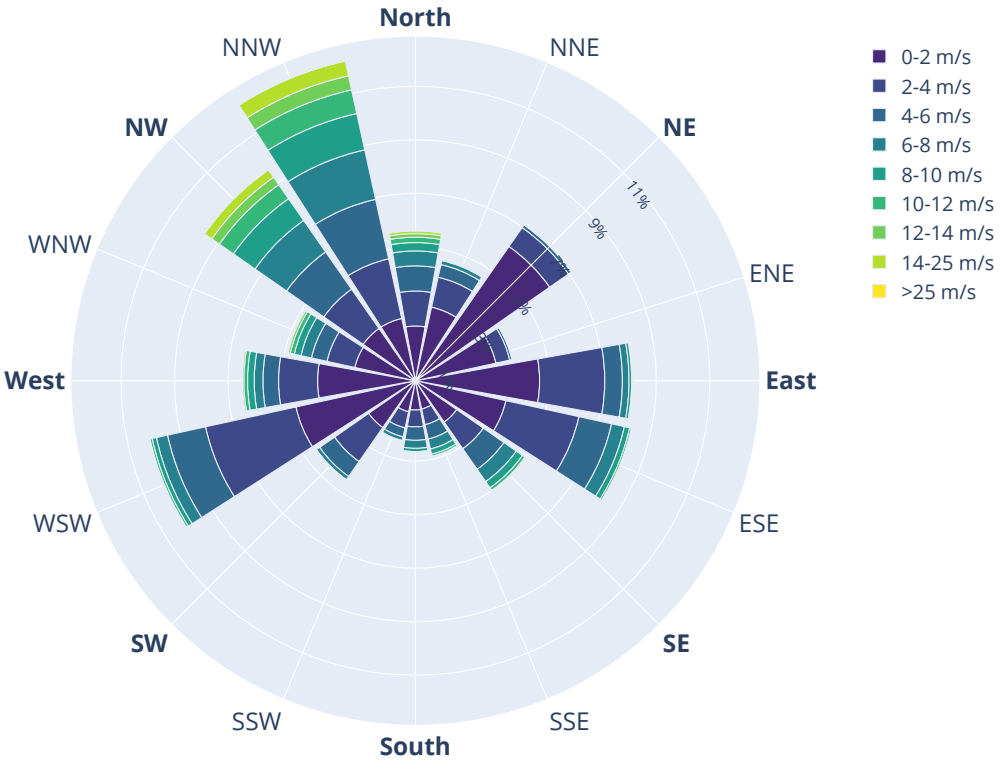

Figure 1.5: Wind Rose of Marseille from 2019 to 2022. Data from 7 GPMM stations and 2 AtmoSud sations (Longchamp & Verneuil).

1.2.3 | A Port City

1.2.3.1 | Le Grand Port Maritime de Marseille

Le Grand Port Maritime de Marseille (GPMM) had a major influence on the development of the city, as mentioned earlier. Nowadays, it stands as France's leading port in terms of passenger traffic and ranks second in terms of merchandise traffic, with about 4 million passengers and 72 Mt of merchandise in 2023 (marseille-port.fr). Its location at the crossroads of major shipping routes makes it an important hub for international maritime trade, as represented in Fig.1.6. Cruise tourism is also highly developed in the Mediterranean Sea. Marseille's central position makes it an ideal destination. In 2023, the GPMM ranked second among Mediterranean cruise ports, with 2 600 000 passengers (medcruise.com), which represent a 40% increase compared to 2019 before Covid-19.

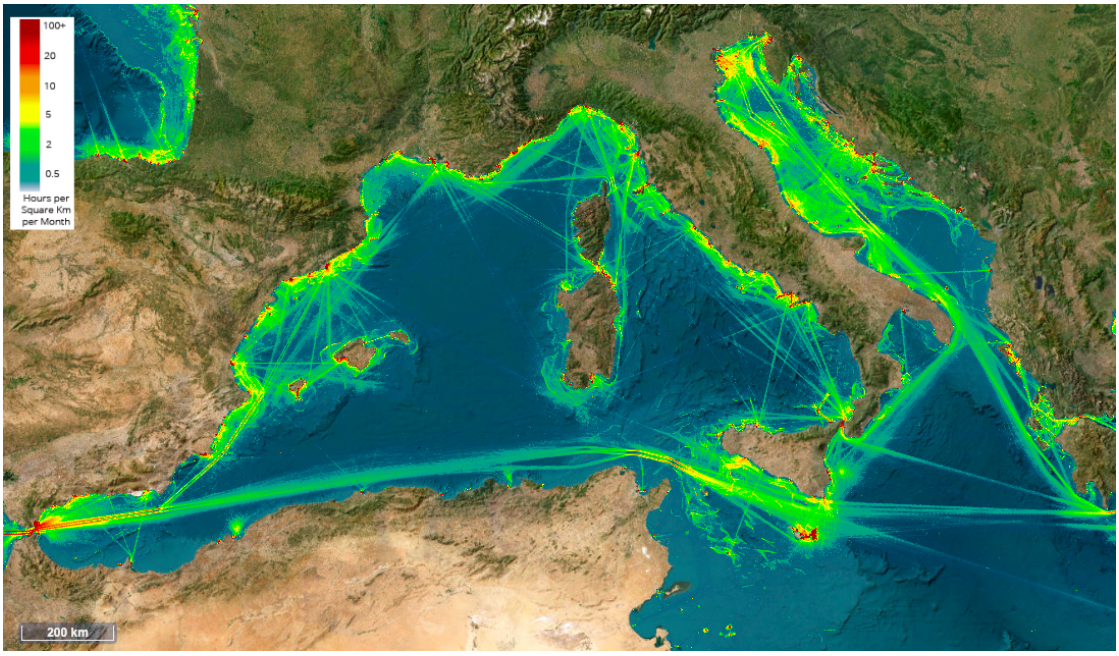

Figure 1.6: Density map of marine traffic in west Mediterranean sea.
Source: European Commission.

The port is divided into two basins, the East basin is in the city of Marseille while the West basin is in the industrial zone of Fos, as represented in Fig.1.7.

1.2.3.2 | East Basin

The East basin comprises 10 terminals spread over 400 ha, from the Vieux Port to l'Estaque. A 7 km-long dike shields the boats from the sea, with two entrances to the port located at the northern and southern ends of the structure. The later is restricted to boats less than 200 m in length. One of the activities of this basin is to welcome cruise ships. The northern section, which receives the largest cruises, has a maximum capacity of 6 ships. Ferries are also operated, most of the time connecting Marseille to Corsica, Sardinia, Algeria and Tunisia. Other terminals handle commercial activities, bringing in containers, dry bulk and liquid bulk carriers. The East Basin is also known for its 10 repair forms, including Form 10, the third largest repair form in the world, situated in the center part of the basin.

1.2.3.3 | West Basin

The Gulf of Fos and its surroundings extend from the mouth of the Rhône River to Cap Couronne. Along with the port of Fos, it includes three ancient ports, the port of Saint Louis du Rhône and the ports of Bouc and Lavéra. The Etang de Berre communicates with the Golf de Fos via the Canal de Caronte. The Fos-sur-Mer area can accommodate tankers of over 400 000 tons at the oil terminal. The latter is also used for the import and export of refined products and chemicals.
Other terminals specialize in LNG (Liquid Natural Gas) carriers, solid fuel imports and other products such as methanol, methylene, vinyl chloride, etc...
Numerous terminals serve containers, with more than 1000 ships a year.

1.2.4 | Air Quality surveillance in Marseille

1.2.4.1 | AtmoSud

In every region of France, there is an organization responsible for monitoring air quality. These organizations are called Associations Agréées de Surveillance de la Qualité de l'Air (AASQA). In the PACA region, AtmoSud assumes this role. Daily air quality is displayed on their website (atmosud.org) with a colormap showing ICAIRh index values all over the region. Data from monitoring stations is also shared on a daily basis. A network of 66 permanent monitoring stations are installed all over the region, of which 7 are in Marseille and about 20 in the area of Fos and Etang de Berre, known for its poor air quality.
Apart from this operational surveillance of the regional air quality, AtmoSud publishes

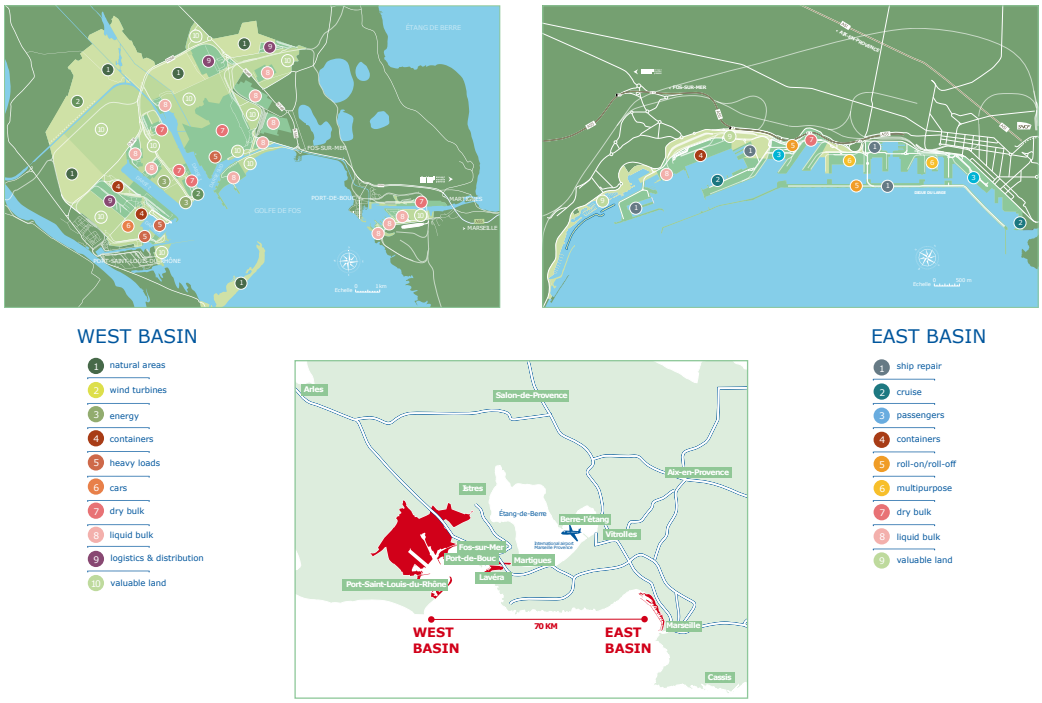

Figure 1.7: Map of the port of Marseille-Fos. Top left: map of the west basin. Top right: map of the east basin. Bottom middle: location of the two basins. Source: marseille-port.fr.

annual reports, as well as specific reports on measurement campaigns, recent pollution events or impact assessments.

AtmoSud regularly participates to large scale projects, often European Union (EU) projects, such as the SCIPPER project (scipper-prohect.eu), which aims at assessing the impact of shipping emissions on air quality in 5 EU ports (Mamarikas et al., 2023). AtmoSud is also a partner in a number of academic thesis projects, such as this one.

1.2.4.2 | Air Quality in the Region

A notable decrease in the concentration values of the main pollutants, with the exception of ozone, is observed, as shown in Fig.1.8. 2021 AtmoSud annual report states that emissions have decreased by about 30% for NO_x and PM_{10}, and up to 80% for SO_2. These emission reductions result from environmental regulations, such as the Plan de Protection de l'Atmosphère (PPA), which applies to any urban area with over 250 000

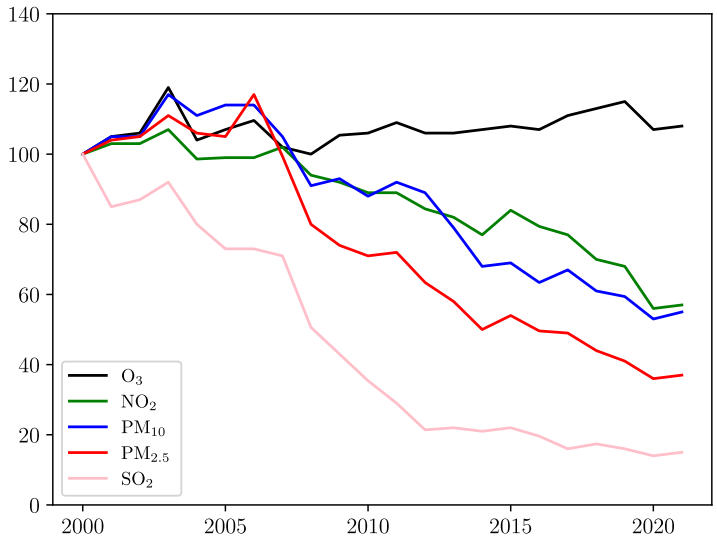

Figure 1.8: Annual concentration evolution since 2000 in % for main pollutants measured at AtmoSud stations. Source: AtmoSud (2022).

inhabitants. This environmental legislation follows the European directives of 2008. Fuel quality has improved, particularly with reduced sulfur content in the industrial and the maritime transport sector. Additionally, technical advancements have improved motor efficiency and filtration systems, significantly contributing to this reduction. Despite these improvements, the number of people exposed to levels above 2021 WHO limits remains alarmingly high. Still according to the same report, in 2019, the proportion of the exposed population was 79%, 80% and 100% for NO_2, PM_{10} and $PM_{2.5}$ respectively.

1.2.4.3 | Local Associations

In response to this overexposure, some inhabitants of Marseille have created associations to express their concerns about air pollution. Most of these associations are in the northern neighborhoods of the city, such as "Cap au Nord", "Mieux Respirer !" and "Action Estaque Environnement". This area is particularly exposed to pollution from cruise ships due to the proximity of the cruise terminals. These associations united around a project called "Opération Capteurs" to gather evidence on the impact of ships on air pollution by installing multiple micro-sensors along the port. These connected micro-sensors send real-time data to the PurpleAir website, where the collected measurements

are accessible at purpleair.com.

1.2.4.4 | Political Context

Other activist groups are informing and demonstrating against Cruises and port activities. On June 14, 2022, the "Stop Croisières" collective blocked access to the port using canoes to prevent the entry of the Wonder of the Seas, the largest cruise ship at the time. Their demands include the progressive cessation of cruise activities for environmental and health reasons, and transparency in the GPMM's action plans.
As these claims and criticisms gain more attention, the port announces several measures for a "green port". This plan includes the following points:

- The installation of electric onshore power supply:
  These installations are called CENAQ for "Connexion électrique des navires à quai". Up to this date, four were installed to ensure Marseille-Corsica lines, and two were added for Marseille-Maghreb lines. A legal action is currently underway in the Administrative Court of Marseille to hold the public authorities (the Marseille City Hall and the Prefecture of Bouches-du-Rhône) and the GPMM authorities accountable for setting an insufficient number of CENAQ and for the delay in implementing these installations.

- Encourage the use of Liquid Natural Gas (LNG) for ships:
  The use of LNG instead of heavy fuel significantly reduces emissions from ships. However, the LNG solution is criticized due to the environmental impact of its extraction methods and methane leaks throughout its lifecycle.

- The use of low-sulfur fuel oils (0.1% m/m):
  The GPMM has announced plans to implement this measure in January 2024, one year ahead of the International Maritime Organization (IMO), which will introduce a Sulfur Emission Control Area (SECA) zone in the Mediterranean in 2025 to enforce the same low-sulfur limit.

- Reduce speed to 10 knots in the port regulation zone.

The Marseille City Hall has launched a petition calling for the prohibition of port access for the most polluting ships, particularly cruise ships, during air pollution peaks. The concerned associations are rather critical of this petition, as they believe that the city hall has public authority and could take such measures on its own, as the port authority

of Nice did on June 16, 2022, at the request of the city hall, to order a polluting cruise ship to leave the port.

1.3 | Modeling Atmospheric Pollution

To better understand air pollution, there are two complementary approaches: measurement and modeling. Using sensors is the most precise way to evaluate the concentration of a pollutant at a given location. However, a reliable sensor over a long period is very expensive, which significantly limits the resolution of a sensor network in a given area. Modeling, therefore, allows for providing an estimate with much finer resolution and for covering much larger areas, as it is only limited by numerical resources. Another advantage of numerical models is their forecasting capability, which allows for predicting air pollution levels for the coming days with a certain degree of accuracy. Nevertheless, numerical models would lack meaning without sensor data, as it is essential for validation purposes.

1.3.1 | Spatial Scales

Air pollution modeling requires a thorough comprehension of meteorological factors, as it will be discussed later (see Section 1.3.2). Therefore, scales for pollution and meteorological modeling are closely related. Atmospheric mechanisms and phenomena cover a vast spectrum of spatial and temporal scales. The spectrum is usually split in 4 categories:

- The Microscale:
  It studies the smallest meteorological events from few centimeters to about one kilometer. This involves surface heat and moist exchanges, local turbulence as well as small structures of clouds. These phenomena occurs on short temporal scales, usually less than a day.

- The Mesoscale:
  Commonly referred to as the regional scale, it covers phenomena occurring between one kilometer and several hundred kilometers, that last for one day to a week. These include weather fronts, large thunderstorms, sea/land breezes and heat island effect.

- The Synoptic Scale:
  Also called the Macroscale, it studies weather mechanisms from several hundred to few thousand kilometers. At this scale, mechanisms are generally governed by

pressure on large areas. Therefore, cyclones are a typical weather event studied at this scale. These phenomena can last over several weeks.

- The Global Scale:
  Sometimes referred to as the Megascale, it pertains to phenomena observed on a global scale that persist for months or years, occurring continuously or with a certain periodicity, at the boundary between weather and climate. Among these are global winds (trades and westerlies) and Hadley, mid-latitude and polar cells. The Oceano-atmospheric event of El Niño also occurs at this scale.

In this work, the scale of interest lies somewhere between the Mesoscale and the Microscale. This scale allows us to include the city of Marseille as well as the industrial area of Fos, while maintaining a sufficient horizontal resolution ($<$ 1km) to keep an adequate number of grid cells over Marseille. The idea is to be able to distinguish pollution levels by neighborhood at the very least. Details of the domain size, time and spatial resolution are discussed in Section 2.2.1. Moreover, the sea-land breeze is effectively represented at this scale, which is especially relevant given Marseille's coastal location. Accurately capturing this phenomenon is crucial for understanding and predicting local air quality patterns.

1.3.2 | Numerical Methods

The numerical methods discussed are not an exhaustive overview of what exists in the literature, but they represent the most commonly used ones for air pollution modeling.

- Gaussian Plume Models:
  The Gaussian plume model is one of the oldest atmospheric dispersion models and remains widely used. It aims to predict the distribution of air pollutants emitted from a constant point or surface source, typically a chimney, under steady wind conditions. It assumes that the pollutants disperse following a Gaussian distribution both horizontally and vertically. Such models are computationally very efficient. However, they are usually quite restrictive due to the assumptions of steady wind flow, constant emission flux and deposition is often considered negligible. Some complex Gaussian-based pollutant dispersion models include chemical processes such as SIRANE (Soulhac et al., 2011) or AERMOD (Cimorelli et al., 2004). Gaussian plume models are typically applied in microscale studies.

- Chemistry Transport Models:
  Chemistry Transport Models (CTMs) are models generally composed of numerous physical and chemical schemes, with the aim of providing a comprehensive representation of atmospheric composition in time and space. The general equation solved by CTMs is the atmospheric transport and diffusion equation, which is typically a form of the continuity equation for chemical species in the atmosphere:

$$\frac{\partial C_i}{\partial t} + \underbrace{\nabla \cdot (\mathbf{u} C_i)}_{\text{Advection}} = \underbrace{\nabla \cdot (K \nabla C_i)}_{\text{Turbulent Diffusion}} + \underbrace{R(C_i)}_{\text{Chemical Reactions}} + \underbrace{E_i}_{\text{Emissions}} - \underbrace{D(C_i)}_{\text{Sink}} \quad (1.3)$$

  where:

  C_i : Concentration of the i-th chemical species (mass per unit volume)

  $\mathbf{u}$: Wind velocity vector

  K : Eddy diffusion tensor

  R : Net rate of production or loss due to chemical reactions

  E_i : Emission rate of the i-th species

  D : Deposition rate

  An important part of the work involved in modeling air pollution is determining the E_i term. Identifying the various sources within the simulation domain can be particularly challenging for regional studies. Emission inventories, which are databases specifically designed for this purpose, provide essential data. However, these inventories typically offer annual and coarse data, introducing a significant source of uncertainty into the overall model. The work carried out on emission data for the present study is presented in Section 2.
  Solving Eq.1.3 requires knowledge of various meteorological variables. The advection term is directly determined by the wind vector $\mathbf{u}$. The chemical reaction term is influenced by temperature, which affects reaction rates according to the Arrhenius law, and by solar radiation, which drives photochemical processes. The emission term may be affected by surface wind speed, which can introduce additional dust and sea salt into the atmosphere. Lastly, the sink term is affected by humidity, precipitation, and wind, all of which play a role in modeling the wet and dry deposition of pollutants. CTMs are often coupled with Numerical Weather Prediction (NWP) models. This coupling can be "off-line," where meteorological data is computed first and then used for chemical and transport calculations, or "online," where the NWP and CTM models interact and influence each other in

real-time, with both being computed simultaneously. The model used in this work called WRF-Chem is a "online" coupling between the NWP model WRF, the gas-phase chemistry model MOZART and an aerosol scheme (GOCART or MOSAIC), as detailed in Section 1.3.3.

There are two main approaches to calculate C_i, Eulerian and Lagrangian. Eulerian models use spatially fixed grid cells, where each cell defines a "box" representing an atmospheric parcel. Physical and chemical processes are computed within these boxes, while the transfer of substances between boxes is calculated at their boundaries, as illustrated in Fig.1.9. On the other hand, the Lagrangian approach

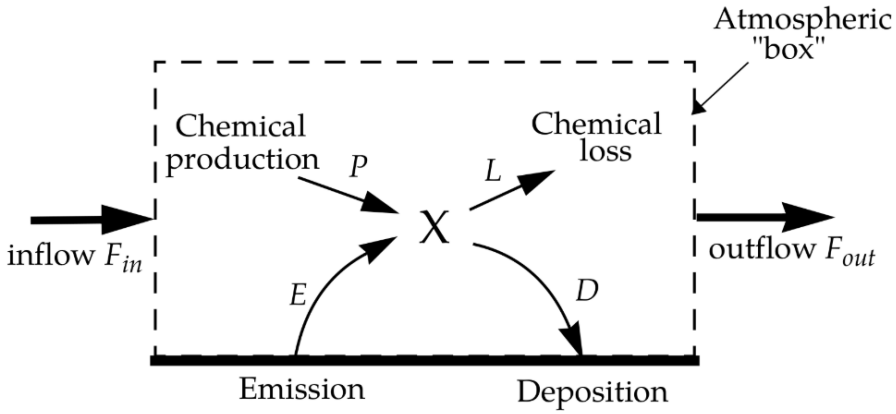

Figure 1.9: Atmospheric processes computed in a Eulerian grid box.
Source: Jacob (1999).

tracks the movement of individual air parcels or particles through the atmosphere. It calculates concentrations by following these parcels or particles along their trajectories, taking into account their interactions with the surrounding environment. In the context of atmospheric pollution modeling, a fluid element is defined as a volume of air, represented in Fig.1.10, that is small enough for all points within it to move uniformly with the same velocity, yet large enough to encompass a statistically representative collection of molecules. While Eulerian models are well suited for regional and large-scale simulations, Lagrangian models are more effective for capturing detailed processes and tracking the movement of pollutants at smaller scales and over specific trajectories. WRF-Chem operates within the Eulerian framework, which is better suited for representing regional-scale mechanisms relevant to our study.

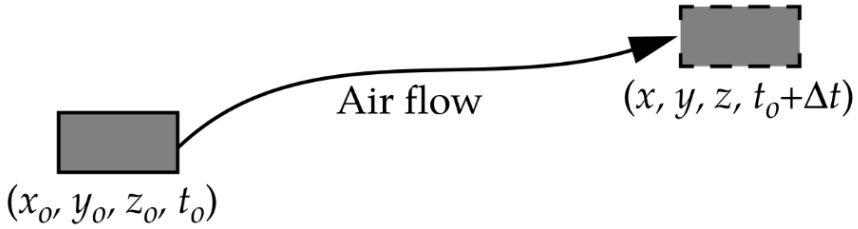

Figure 1.10: Movement of a fluid element along an air flow trajectory.
Source: Jacob (1999).

- Machine Learning:
  Machine learning has become a widely used tool for predicting air quality, with a significant rise in scientific publications since the 2010s (Méndez et al., 2023). This surge is largely due to the recent proliferation of available air pollution data, including ground measurements from monitoring stations, numerical simulations, and satellite data, particularly from the MODIS instrument. In parallel, advancements in machine learning have led to the development of new efficient techniques for various applications such as predictive modeling, source apportionment, downscaling, data imputation and anomaly detection.
  Notably, ML has proven to be a powerful tool for predictive modeling, especially with the use of Deep Neural Networks (DNN). However, most studies propose one-dimensional prediction models, typically at specific measurement station locations. In this manuscript, we focus on spatio-temporal forecasting models, which provide predictions in the form of 2D or 3D scalar fields. Predicting the spatio-temporal behavior of a concentration field using machine learning remains a challenging task due to the high dimensionality of the data and the complex nonlinear relationships between meteorological features and pollutant concentrations. Deep learning has shown particular promise in this area, as it is highly effective for managing large datasets and capturing complex nonlinear mechanisms. Different types of architecture and algorithms were used for spatio-temporal air pollution forecasting with deep learning. Chen et al. (2021) predicted $PM_{2.5}$ concentration over China with a DNN and an attention mechanism based on satellite and meteorological data. Other models such as SIDLM model (Yan et al., 2021) use a combinasion of DNN and a linear regression model. Zang et al. (2021) added a tree-based architecture to SIDLM in order to predict Ozone in China. An autoencoder model was

implemented by Cheng et al. (2022) to model ozone concentration in China as well.

1.3.3 | WRF-Chem

This section introduces the WRF-Chem model used in this work. WRF-Chem is a coupling of the NWP model WRF, the gas-phase chemistry model MOZART and an aerosol scheme. Throughout this PhD project, two distinct aerosol schemes, GOCART and MOSAIC, were employed, both of which are presented in this section.

1.3.3.1 | Weather Research and Forecasting

The Weather Research and Forecasting (WRF) model is a meteorological forecasting model developed by the National Oceanic and Atmospheric Administration (NOAA) in the United States, in partnership with the National Center for Atmospheric Research (NCAR) and over 150 international research centers. The research code is open source, making it one of the most widely used weather models in atmospheric research. It includes two dynamical cores: the Advanced Research WRF (ARW) and the Non-hydrostatic Mesoscale Model (NMM). Since ARW is the solver used in this work, we will focus solely on describing it. Fig.1.11 shows the different components of ARW. As we are only interested in representing real weather scenario, the Real program is used instead of the Ideal program. Real requires terrestrial data, typically integrated into WRF, and gridded data used for initial and boundary conditions for weather. In this study, the gridded data utilized are GDAS/FNL 0.25°, which initially consist of coarse simulated global-scale data corrected with ground measurements and satellite data by the National Center for Environmental Prediction (NCEP). Bellow is a brief description of the main components of the Real ARW framework:

- WPS:
  WPS (WRF Preprocessing System) program has three major functions: defining the simulation domains and interpolating terrestrial data into them, converting meteorological data files, and finally interpolating these data to the calculation domains. These functions are performed by the WPS sub-programs Geogrid, Ungrib and Metgrid respectively. The parameters are configured by editing the `namelist.input` file, which is used by all three algorithms.

  Geogrid preprocesses the WRF Terrestrial Data (see Fig.1.12) and creates the domains and the grids according to the information given by the user in the

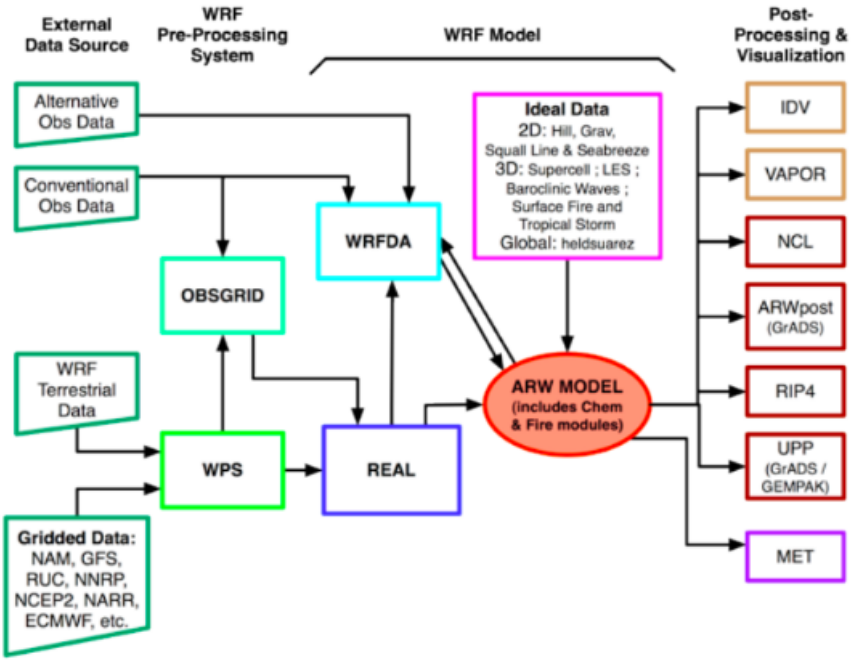

Figure 1.11: ARW Modeling System Flow Chart. Source: Skamarock et al. (2008)

`namelist.input` file. Various types of projection are available in Geogrid. The Lambert projection was selected for this project, as it is best suited to mid-latitudes.

Then, Ungrib processes GDAS/FNL 0.25° data, formatting them to be compatible with the WRF code.

Finally, Metgrid interpolates the data horizontally on the simulation grid, providing initial and boundary conditions to the ARW solver.

- Real:
  Real's main function is to compute the vertical interpolation of the input data. The Eta vertical coordinate system is used. Eta levels are represented on Fig.1.12 and are defined as follows:
  $$\eta = \frac{p_h - p_{ht}}{p_{hs} - p_{ht}} \tag{1.4}$$
  With p_{hs} the hydrostatic pressure at the surface and p_{ht} the hydrostatic pressure at the highest level of the simulation domain. Such vertical coordinates are said to be terrain-following, as p_h, the hydrostatic pressure, directly depends on ground elevation. The system is termed hybrid due to the treatment of pressure as isobaric beyond a specified threshold value η_c In our case, $\eta_c = 0.2$.

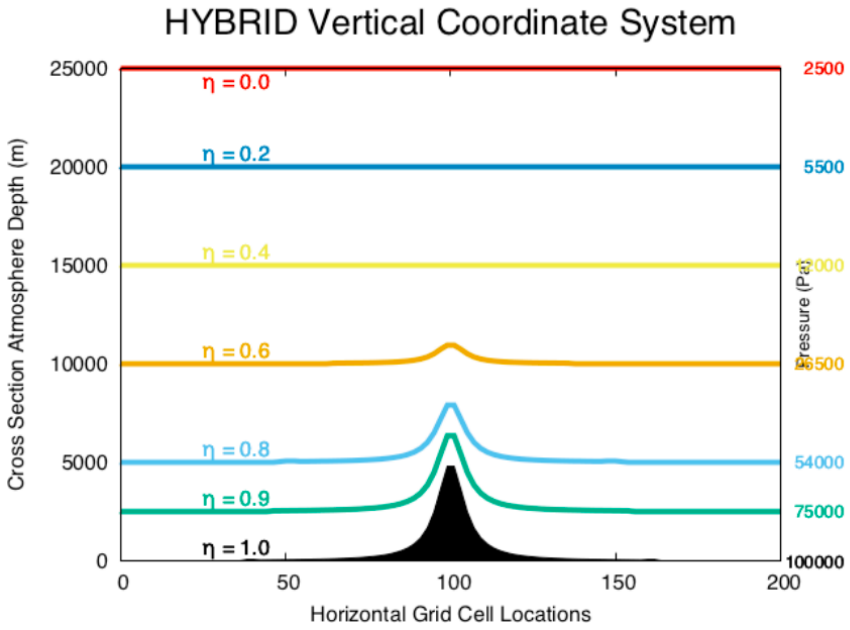

Figure 1.12: Eta vertical coordinate system. Source: Skamarock et al. (2008)

- ARW:
  The ARW model is non-hydrostatic, meaning that vertical convective motions are accounted for during calculations. The model uses the vertical coordinate system introduced in Eq.1.4.

  The surface column mass is defined as:

$$\mu = p_{hs} - p_{ht} \tag{1.5}$$

  The equations solved by ARW are in the flux form. These are written as:

$$\mathbf{V} = \mu\mathbf{v} = (U, V, W),\ \Omega = \mu\dot{\eta},\ \mathbf{\Theta} = \mu\theta \tag{1.6}$$

  Where $\mathbf{v} = (u, v, w)$ are the covariant velocities and $\omega = \dot{\eta}$ is the contravariant vertical velocity (vertical in terms of hybrid coordinates) and θ is the potential temperature. Also appearing in the equations are $\phi = gz$ the geopotential, p the pressure, and $\alpha = 1/\rho$ the specific volume.

  To integrate humidity into the dynamic equations, the coordinate system for a dry atmosphere is defined as:

$$\eta = \frac{(p_{dh} - p_{dht})}{\mu_d} \tag{1.7}$$

where μ_d represents the dry air column mass, and p_{dh} and p_{dht} are the hydrostatic pressures of the dry atmosphere. The flux variables can be rewritten as:

$$\mathbf{V} = \mu_d \mathbf{v},\ \Omega = \mu_d \dot{\eta},\ \Theta = \mu_d \theta \tag{1.8}$$

Using these notations, the dynamic equations are written as:

$$\partial_t U + (\nabla \cdot \mathbf{V}u) + \mu_d \alpha \partial_x p + \left(\frac{\alpha}{\alpha_d}\right) \partial_\eta p \partial_x \phi = F_U \tag{1.9}$$

$$\partial_t V + (\nabla \cdot \mathbf{V}v) + \mu_d \alpha \partial_y p + \left(\frac{\alpha}{\alpha_d}\right) \partial_\eta p \partial_y \phi = F_V \tag{1.10}$$

$$\partial_t W + (\nabla \cdot \mathbf{V}w) - g(\partial_\eta p - \mu) = F_W \tag{1.11}$$

$$\partial_t \Theta + (\nabla \cdot \mathbf{V}\theta) = F_\Theta \tag{1.12}$$

$$\partial_t \mu_d + (\nabla \cdot \mathbf{V}) = 0 \tag{1.13}$$

$$\partial_t \phi + \frac{(\mathbf{V} \cdot \nabla \phi) - gW}{\mu_d} = 0 \tag{1.14}$$

$$\partial_t Q_m + (\nabla \cdot \mathbf{V}q_m) = F_{Q_m} \tag{1.15}$$

In the above equations,

$$\nabla \cdot \mathbf{V}a = \partial_x(Ua) + \partial_y(Va) + \partial_\eta(\Omega a) \tag{1.16}$$

and

$$\mathbf{V} \cdot \nabla a = U\partial_x a + V\partial_y a + \Omega \partial_\eta a \tag{1.17}$$

with a being a generic variable. The right-hand side terms $F_U, F_V, F_W, F_\Theta, F_{Q_m}$ are the forcing terms from physical parameterization, spherical projections, and Earth's rotation.

α_d denotes the specific volume of dry air and α the specific volume of total air such that $\alpha = \alpha_d(1 + q_v + q_c + q_r + q_i + \ldots)^{-1}$ with q_* being the mixing ratios of vapor, cloud, rain, snow, etc...

These equations are supplemented by the diagnostic equation for the specific volume of dry air:

$$\partial_\eta \phi = -\alpha_d \mu_d \tag{1.18}$$

and the equation of state:

$$p = p_0 \left(\frac{R_d \theta_m}{p_0 \alpha_d}\right)^\gamma \tag{1.19}$$

with $\gamma = \frac{c_p}{c_v}$ the ratio of specific heats for dry air, R_d the gas constant for dry air, and $p_0 = 10^5$ Pascal the reference pressure, $\theta_m = \theta(1+\frac{R_v}{R_d})$, and $Q_m = \mu_d q_m$; $q_m = q_v, q_c, q_r, \ldots$

In practice, these equations are not integrated in this form by the ARW algorithm. To minimize truncation errors, the variables are decomposed into a reference state and a perturbation. This involves modifications to the equations that are not written in this document, but are available in Section 2.5 of the ARW technical document (Skamarock et al., 2008).

- Temporal Discretisation

  ARW uses a two-time-step method. The first time step is defined to integrate low-frequency modes, representing most meteorological phenomena. The second time step is used for high-frequency modes, generally related to the propagation of acoustic waves. This separation ensures numerical stability. The third-order Runge-Kutta (RK3) scheme is used for calculating low-frequency phenomena. Let Φ be the variable to predict, and the associated equation $\Phi_t = R(\Phi)$. The RK3 scheme is structured as follows:

$$\Phi^* = \Phi^t + \frac{\Delta t}{3} R(\Phi^t) \tag{1.20}$$

$$\Phi^{**} = \Phi^t + \frac{\Delta t}{2} R(\Phi^*) \tag{1.21}$$

$$\Phi^{t+\Delta t} = \Phi^t + \Delta t R(\Phi^{**}) \tag{1.22}$$

  Δt is the time step for low-frequency modes. A second, smaller time step $\Delta\tau$ is then defined, such that the predictions of the variables involved in the acoustic equations are integrated within the RK3 scheme. Fig.1.13 shows an example configuration where $\Delta\tau = \Delta t/4$, which is the ratio used in this work. R_{LF} and R_{HF} refers to the resolution of low and high frequencies respectively, such that $\Phi_t = R_{LF}(\Phi) + R_{HF}(\Phi)$.

- Spatial Discretization:

  Spatial discretization is not described in details here, see Skamarock et al. (2008) for more information. Only horizontal and vertical structure of the grid is discussed and represented in Fig.1.14.

  This structure is known as the Arakawa C-grid. The points at the center of each grid cell, where θ is located, are referred to as mass points. At these points,

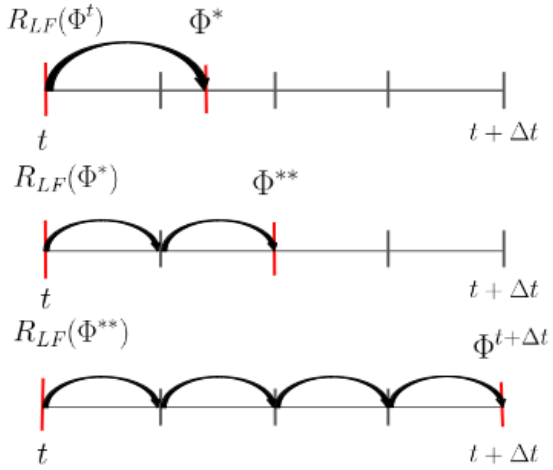

Figure 1.13: Integration of the acoustic time step in the RK3 scheme

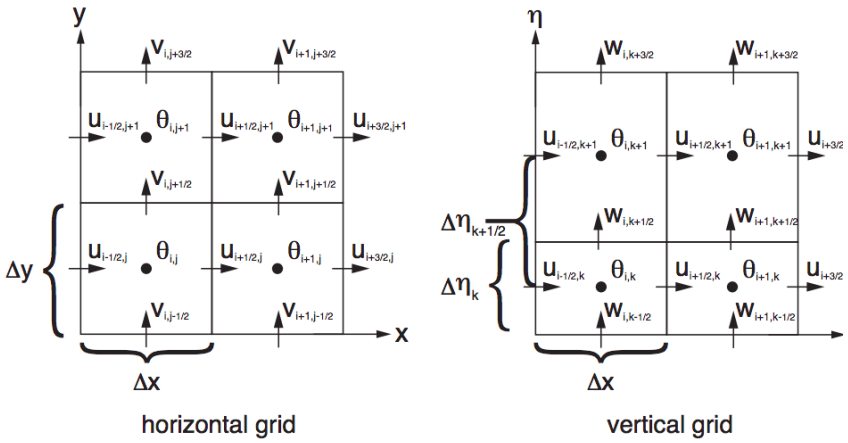

Figure 1.14: Arakawa C-grid. Source: Skamarock et al. (2008)

scalar variables such as concentration C_i, potential temperature θ, column mass μ, mixing ratios q_m, pressure p, and specific volume α are calculated. These variables represent the average values within a grid cell. In the horizontal grid, the velocity U is computed at the midpoints of the edges normal to the x-axis, while the velocity V is evaluated at the midpoints of the edges normal to the y-axis. In the vertical grid, the velocity W and the vertical velocity Ω are calculated at the midpoints of the horizontal edges. The geopotential ϕ is also positioned on these edges, as W appears in the equation $\phi = g \cdot W$. Mass point levels are referred to as unstaggered

levels.

- Parameterization:
  ARW allows for multiple physic schemes to represent subgrid-scale processes. The physics processes are microphysics, cumulus parameterization, planetary boundary layer (PBL), land-surface model and radiation. For each process, several schemes are available. Interactions between the different physics categories are summarized in Fig.1.15. The choices made in this work regarding physics are detailed in Section 2.2.3.

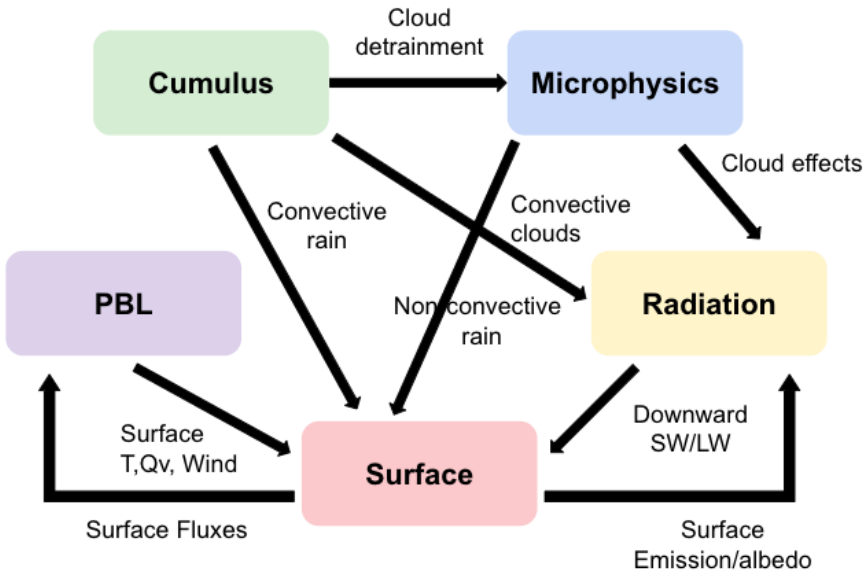

Figure 1.15: Interactions between physics categories in WRF.

1.3.3.2 | WRF-Chem

WRF-Chem is an extension of the WRF model that incorporates chemical processes. It allows for the simulation of the transport, dispersion, and chemical transformation of trace gases and aerosols within the atmosphere. To model these phenomena, the MOZART (Model for Ozone and Related Chemical Tracers) atmospheric chemistry model is coupled with an aerosol scheme. The aerosol and chemistry components are entirely compatible with the NWP aspect of the model. They share the same horizontal and vertical grid, the same transport systems, and utilize the same physics schemes for subgrid-scale transport.

- MOZART:
  MOZART describes the evolution of 85 chemical species in the gas phase involved in 157 chemical reactions and 37 photolysis reactions. The full list of the reactions and the description of chemical mechanisms are present in Emmons et al. (2010).

  Transport of chemical species is based on the equation of conservation of scalar variables:

  $$\partial_t(\mu_d q_i) + (\nabla \cdot \mathbf{V}\mu_d q_i) = 0 \tag{1.23}$$

  where q_i is the mixing ratio of species i. Note that Eq.1.23 and Eq.1.15 are the same if $F_{Q_m} = 0$.
  For dry deposition velocities, the resistance-based representation of Wesely (1989) is used. For photochemical reactions, photolysis rates are calculated using the FTUV (Fast Tropospheric Ultraviolet-Visible) model described in Tie et al. (2003). The partitioning of the spectrum into 17 bins in the range $[121 - 700]$ nm allows for an efficient computation of the photolysis rates.

- GOCART:
  The GOCART (Goddard Chemistry Aerosol Radiation and Transport) model, developed by Chin et al. (2000), simulates the main aerosols present in the troposphere including sulfate, dust, organic carbon (OC), black carbon (BC), and sea salt. GOCART is a bulk scheme so the particle size distribution is not resolved, except for dust and sea-salt, for which 5 and 4 sectional bins are considered respectively. Secondary organic aerosols (SOA) are not modeled in GOCART. This aerosol scheme, while less accurate than more complex schemes such as MOSAIC, is computationally less costly, making it suitable for non-real case studies. This scheme was used in Section 2.7, and for generating the dataset on which the ML model was trained. The combination of MOZART with GOCART is referred to as MOZCART.

- MOSAIC:
  The MOSAIC (Model for Simulating Aerosol Interactions and Chemistry) model was developed by Zaveri et al. (2008). It includes sulfate (SULF = SO_4^{2-} + HSO_4^-), methanesulfonate (CH_3SO_3), nitrate (NO_3), chloride (Cl), carbonate (CO_3), ammonium (NH_4), sodium (Na), calcium (Ca), black carbon, primary organic mass, liquid water and other inorganic mass. MOSAIC uses 4 aerosol size sections (bins) for all species, ranging from $39nm$ to $10\mu m$. This scheme operates with the VBS (Volatility Basis Set) approach for the computation of SOA. The MOSAIC scheme

was used in Section 2.6. The combination of MOZART with MOSAIC is referred to as MOZAIC.

1.4 | Objectives of Research

This PhD project focuses on two main objectives. The first is directly related to the issue of pollution from maritime traffic in Marseille, as introduced in Section 1.2. Most of the models currently used for air quality prediction account for annual averaged maritime emissions, which do not capture high pollution events caused by increased maritime traffic. Developing a numerical model to predict air pollution based on real-time traffic data would allow for an accurate assessment of the impact of maritime pollution in the bay of Marseille. The model Set-Up is presented in Section 2.2, while model validation can be found in Section 2.5. A real case study is then carried, to assess the contribution of marine traffic to pollution. Then, a high traffic scenario is studied under different meteorological conditions, to better understand the influence of weather on ship pollution in Marseille.

The complexity of such a model represents a significant computational cost, posing a challenge for operational use. One way to address this problem is through the use of machine learning. Thus, the second objective of this thesis is to develop a machine learning model capable of predicting air pollution in Marseille. The high dimensionality of the data for both inputs and outputs, as well as the strong non-linearity of the mechanisms involved in the atmosphere, makes it a real challenge. Reduced order modeling techniques are applied to this specific problem. A first approach is presented, using Proper Orthogonal Decomposition (POD) as a reduction method, coupled with a Recurrent Neural Network (RNN). This approach is then generalized to a non-linear reduction method, by using Auto-Encoders instead of POD. These recent techniques are still in development in the field of CFD (Computational Fluid Dynamics), and have not been applied to spatio-temporal air quality prediction so far.

2

Modeling of Atmospheric Pollution due to Marine Traffic with WRF-Chem

2.1 | Introduction

2.1.1 | Marine Traffic and Air Pollution

The seaways are increasingly frequented, both for development of international trade and for tourism. Although maritime activity has experienced a recent slowdown due to COVID-19, projections for 2050 anticipate a growth in global international maritime trade, with estimates ranging from a 25% to 250% increase (DNV, 2022). Besides, 70% of emissions related to maritime transport occur within 400 km from the coast (Endresen et al., 2003). In fact, most cruise ships continue to run their engines at berth, directly affecting the health of populations living near port areas. The proximity of large ports to densely populated areas makes these emissions a real threat to the health of the many residents of these port cities. Controlling pollution from maritime transport in port cities is therefore of utmost importance, hence the significance of building a model capable of accurately estimating the impact of port activity on air pollution using comprehensive and reliable data.

2.1.1.1 | Global and Mediterranean Impact

The impact of boat pollution on health has been the subject of numerous studies over the past couple of decades. Mueller et al. (2023) provide an extensive literature review on health impact assessments. At a global scale, it has been estimated that ship pollution was responsible for 60 000 premature deaths worldwide in 2001/2. Most of these deaths occurred along major shipping routes in the European/Mediterranean region and East Asia, while South East America was the least affected (Corbett et al., 2007). A similar

study by Partanen et al. (2013) estimated that in 2010, approximately 50 000 premature deaths from lung cancer (LC) and cardio-pulmonary diseases (CP) were attributable to ship pollution. Sofiev et al. (2018) projected ∼265 000 LC and CP premature deaths, with 6.4 M childhood asthma cases.
In the Mediterranean Sea, only two studies were found on the health impact of ship pollution. Viana et al. (2020) assessed the health impact of ship pollution in eight cities (Nicosia, Brindisi, Genoa, Venice, Msida, Mellila, Athens, and Barcelona). They found that ship-sourced $PM_{2.5}$ accounted for 430 (95% CI: 220–650) premature deaths per year. Nunes et al. (2021) estimated that shipping emissions caused 1944 (95% CI: 1294–2528) deaths due to $PM_{2.5}$ and 1536 (95% CI: 1,907–2,132) deaths due to NO_2 in the Iberian Peninsula (Portugal + Spain).

2.1.1.2 | Previous studies on ship pollution

Various air quality models have been applied worldwide to determine the impact of marine traffic on air quality: CALPUFF in Napoli (Murena et al., 2018; Toscano et al., 2021); TAPM in Hamburg (Ramacher et al., 2020); CAMx in Europe (Jiang et al., 2020); AERMOD in Turkey (Kuzu et al., 2021) etc... WRF-Chem has been used extensively to assess the impact of maritime traffic on air pollution and has demonstrated high efficiency. In China, WRF-Chem is commonly used in the recent studies for that purpose (Chen et al., 2018; Cheng et al., 2023; Huang et al., 2022; Ma et al., 2022; Wang et al., 2019). WRF-Chem was also applied to study maritime pollution in the Arctic seas by Marelle et al. (2016) and Raut et al. (2022). However, the application of WRF-Chem to estimate the impact of maritime traffic on air quality has not yet been implemented in the western Mediterranean. Toscano (2023) has recently produced a detailed review of the numerous studies on the contribution of ship emissions to air quality in the main ports in the Mediterranean sea. Among the 32 studies referenced in the review, only one is carried in France, indicating a lack of studies of this type in the country. As the port of Marseille (GPMM) has the highest number of calls in France, it has been the subject of a few studies when it comes to its impact on the city's air quality. Project APICE in 2012 (Salameh et al., 2018) followed by Project CAIMANS in 2015 (Piga and Armengaud, 2015) produced a comprehensive report comparing the impact of ships on air pollution in the main Mediterranean ports, including GPMM. However, simulations in these projects were based on annual emission inventories, and therefore does not capture high and short emission events. In 2020, a significant measurement campaign was carried out in Marseille as part of the SCIPPER project. Boikos et al. (2024) have sim-

ulated the dispersion of a ship plume in the port of Marseille with a steady RANS-CFD approach on a domain of 2 x 1.2 km^2. This fine-scale study incorporates the complex geometry of buildings. However, pollutants were considered chemically inert due to the short timescales of the simulation relative to those of chemical processes.

2.1.1.3 | Regulation

Until recently, heavy fuel oil (HFO), with its high sulfur and nitrogen content, was the most cost-effective and commonly used fuel, which caused significant air pollution during combustion. With the International Maritime Organization (IMO) now imposing stricter sulfur limits, there is a shift towards using cleaner fuels such as marine diesel oil (MDO), marine gas oil (MGO), and liquefied natural gas (LNG), which are anticipated to greatly reduce air pollution levels (Mueller et al., 2023). Since 2020, the IMO has implemented measures called "IMO-2020" to curb ship fuel emissions under Annex VI of the MARPOL (Marine Pollution Convention) (IMO, 2016), imposing a sulphur content limit of 0.5% m/m (mass to mass). The Mediterranean Sea is under consideration for designation as a SECA zone in 2025, which would lower the sulphur content limit down to 0.1% m/m. A recent study conducted in the context of the Baltic Seas, where the SECA is already in effect, assessed the benefits of adopting such a policy, resulting in an estimated reduction in premature deaths attributable to maritime traffic of about 35% (Barregard et al. (2019)). Still under the same Annex VI, since 2013, new ships must comply with the Energy Efficiency Design Index (EEDI) and the Ship Energy Efficiency Management Plan (SEEMP), which focuses on the reduction of GHG emissions from ships. European directives also play a role in regulations. The first directive (Directive EU 2014/94/EU) mandates the installation of LNG refueling points and the provision of onshore electric power in all Trans-European Transport Network (TEN-T) ports by 2025. Another directive (Directive EU 2016/802) imposes a sulfur content limit of 0.1% m/m for ships at berth in all EU ports if they are docked for more than two hours. Despite these efforts to regulate fuel pollutant content, elevated levels of sulfur oxides (SO_x), nitrogen oxides (NO_x), and fine particles (PM_{10} and $PM_{2.5}$) persist in ship emissions.

2.1.2 | Methodologies for Ship Emissions

As outlined in the research objectives, a crucial aspect of this work is incorporating real-time ship emission data into the model. In this context, real-time data refers to time-resolved data that aligns with the model's other time steps. Hourly data is deemed

sufficient, as meteorological inputs are interpolated on an hourly basis. Most emission inventories, such as EDGAR-HTAP at a global scale or EMEP in Europe, rely on annual maritime traffic statistics, some of which are over a decade old. Therefore, these databases have been excluded from this study for ship emission estimation. This aspect is all the more crucial as it aligns with the long-term goal of an operational use of the framework developed in this study.

2.1.2.1 | Top Down v.s Bottom Up

Ship emissions can be estimated using either "bottom-up" or "top-down" methodologies. The top-down method estimates emissions based on total fuel consumption, while the bottom-up approach utilizes raw ship traffic data. Studies have shown that top-down estimations are generally less accurate (Miola and Ciuffo, 2011). Consequently, the bottom-up methodology is adopted in this study.

2.1.2.2 | Real-Time Ship Traffic Data

Ship traffic data is typically derived from the Automatic Identification System (AIS), which provides real-time information for a wide range of vessels worldwide (Van der Tak and Cotteleer, 2011; Winther et al., 2014; Yang et al., 2015). However, AIS data is not open source and requires either purchasing access through platforms such as Marine Traffic or participating in the data system by installing a receiver. Alternatively, access to traffic data can be obtained from port authorities. GPMM authorities agreed to share for the first time their traffic data in October 2022, which gives access to all reported ship calls in the past years. These data are detailed in Section 2.3.

2.1.2.3 | From Traffic to Emissions

The general method to evaluate emissions from a ship is based on two main characteristics: the gross tonnage (GT), and the type of the boat. Correlations have indeed been established for each boat type to determine the nominal power (NP) based on the GT (EMEP-Table 3-12). From these data it is possible to compute emissions according to the following formula:

$$E_{i,b,m,p} = \sum_{e} P_{b,e} * LF_{b,e,p} * EF_{i,e,m}/S \tag{2.1}$$

with $E_{i,b,m,p}$ the emission flux (in $g \cdot km^{-2} \cdot h^{-1}$) of the pollutant i from ship type b (Ro Ro Cargo; Bulk Carriers; Container; Passenger; General Cargo; Tugs; Tanker etc...)

consuming fuel type m (Residual Oil (RO); Marine Diesel Oil (MDO); Marine Gas Oil (MGO)) during phase p (cruise; manoeuvring; hotelling). $P_{b,e}$ is the nominal power (in kW), $LF_{e,b,p}$ is the load factor for engine type e (Propulsion; Auxiliary; Boiler), $EF_{i,e,m}$ is the emission factor (in $g \cdot kWh^{-1}$), and S the surface of a grid cell (in km^2).
As shown in the above equation, the total emission from a boat is the sum of the emissions from the Main (Propulsion), Auxiliary and Boiler engines. Methodologies typically vary in their emission factors, GT/NP correlations, and assumptions about the ship's age. In this work, the applied methodology will be specified for each study, as there may be slight differences.

2.2 | Model Set-Up

This section describes the primary configuration of the WRF-Chem model used throughout this study. This setup is referred to as MOZAIC, reflecting its specific chemistry and aerosol configurations. Section 2.7 discusses a setup with minor variations from the main MOZAIC configuration, which will be clearly identified. This variant is referred to as the MOZCART setup.
Table 2.1 provides a summary of the key information regarding the main setup.

Table 2.1: MOZAIC setup.

Model setup	Values
DOI	Marseille and its surroundings
Horizontal resolution	0.8x0.8 km^2
Vertical resolution	51 eta levels up to 50 hPa
DOI size	102 × 102 cells
Land Use	IGBP MODIS Noah 21-categories
Meteorological I&BC	GDAS/FNL 0.25°
Chemical I&BC	WACCM
Nudging	On
Physics	**Scheme**
Microphysics	Morrison
Cumulus	Grell-3
Shortwave Radiation	RRTMG
Longwave Radiation	RRTMG
PBL	MYNN
Horizontal Diffusion	2D Smagorinsky
Surface Layer	MYJ/ETA
Land-Surface	Noah LSM
Urban Canopy	UCM
Chemistry	**Scheme**
Gas phase chemistry	MOZART
Aerosols	MOSAIC 4-bin
Photolysis	FTUV
Biogenic Emissions	MEGAN

2.2.1 | Study Area

The studied area is centered on Marseille, and includes both the EAST and WEST basins of the port of Marseille-Fos as depicted in Fig.2.1. The domain of interest (DOI) extends

from 4.80°E to 5.80°E and from 42.89°N to 43.61°N, i.e. approximately 80×80 km^2. A post-simulation sub-domain of about 23×20 km^2 (from 5.26°E to 5.51°E and from 43.18°N to 43.39°N) is defined in blue to specifically focus the analyses on Marseille.

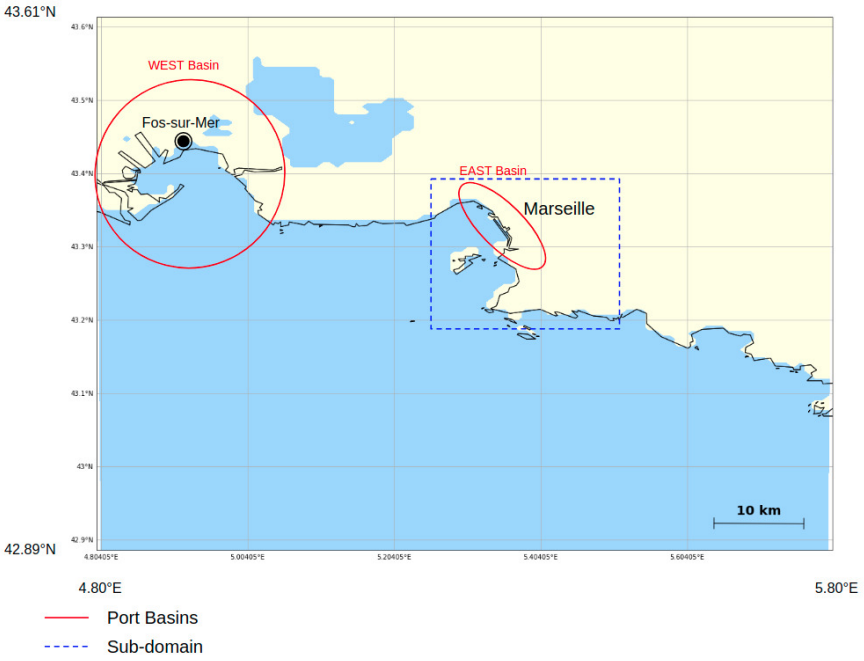

Figure 2.1: Map of the domain of interest

2.2.1.1 | Nesting

Nesting in the WRF model involves simulating at much higher resolutions than the initial and boundary condition data (e.g., GDAS/FNL 0.25°). This technique employs a parent domain interacting with fully nested child domains, which have finer spatial and temporal resolutions. Boundary conditions for the finer grids are derived through interpolation from the coarser grid at each time step. This nesting process can be iterated by nesting domains within other nested domains, thereby progressively refining the grid to focus on specific areas of interest. In this work, we use a three-level nesting configuration, as illustrated in Fig.2.2. Starting from GDAS/FNL data at 0.25° resolution, which translates to approximately 27×27 km^2, the nested grids progressively refine to achieve a final horizontal resolution of 0.8×0.8 km^2 over the domain of interest.

2.2.2 | Numerical Methods

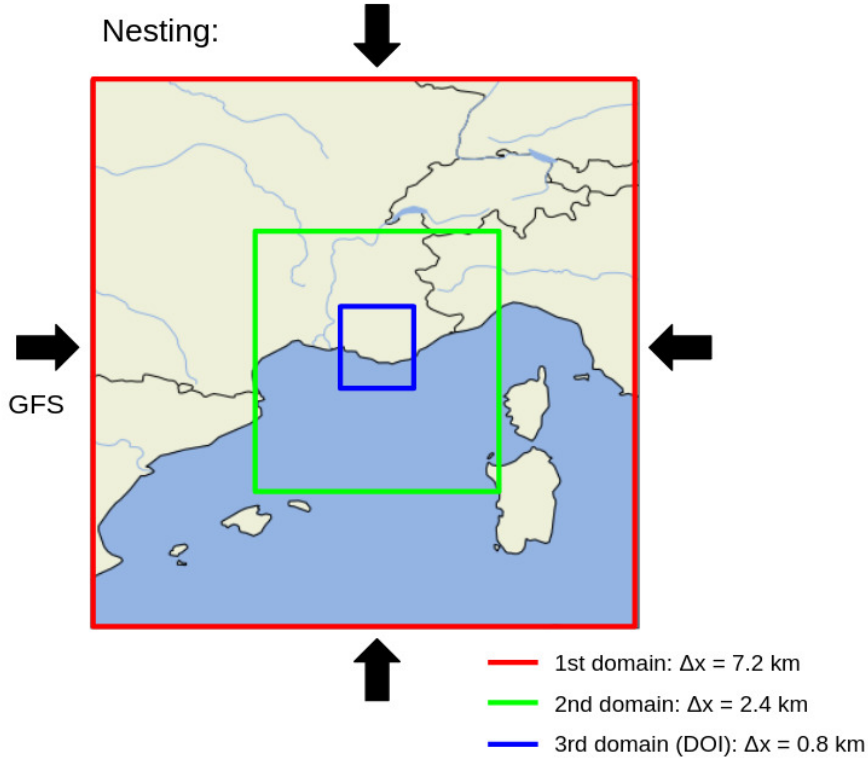

Figure 2.2: Map of the three nested domains

2.2.2.1 | Horizontal Resolution

Horizontal Resolutions for the three domains are summarized in Table 2.2. All three domains have the same number of cells in latitude and longitude. A grid ratio of 3 was set from a parent to a child domain in order to minimize interpolation calculations.

Table 2.2: Spatial description of the WRF simulation domains.

	Δx	Number of cells	Domain Length
D01	7.2 km	124×124	892.8 km
D02	2.4 km	114×114	273.6 km
D03	0.8 km	102×102	81.6 km

2.2.2.2 | Vertical Resolution

The simulations are run on 51 vertical levels, from ground level to 5000 Pa (≈ 19.3 km). Figure 2.3 shows the distribution of the 51 vertical levels, while Figure 2.4 provides a

close-up of the first 10 levels. The grid is also staggered in the vertical direction, with most of the scalar variables being computed at mass points. Therefore, only 50 vertical levels are used for these computations. The first level for mass point computation is 25 m.

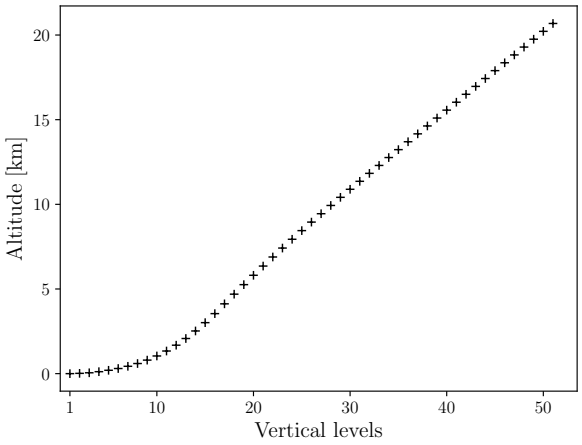

Figure 2.3: Mean elevation of vertical levels

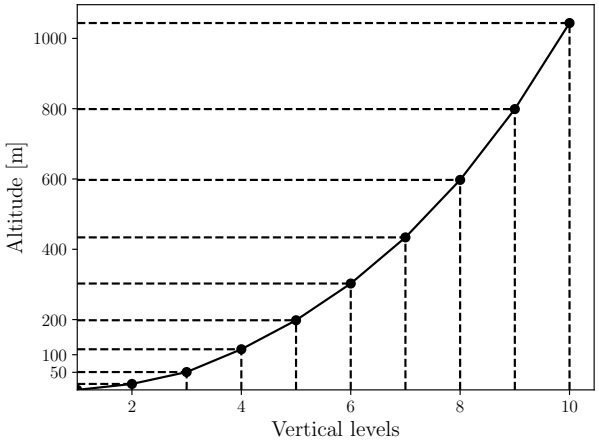

Figure 2.4: Mean elevation of the first 10 levels

2.2.2.3 | Temporal Resolution

As described in Section 1.3.3, ARW uses two time steps for integration, a low-frequency time step Δt and an acoustic time step $\Delta \tau$. The choice Δt is limited by the value of

the advective Courant number $u\Delta t/\Delta x$. Depending on the time scheme and the order of spatial discretizations, there exist theoretical values $C_{r_{theory}}$ above which stability is not guaranteed, displayed in Table 2.3.

Table 2.3: Maximum stable Courant numbers for one-dimensional linear advection. Source: Wicker and Skamarock (2002).

Time Scheme	Spatial Order			
	$3rd$	$4th$	$5th$	$6th$
Leapfrog	Unstable	0.72	Unstable	0.62
RK2	0.88	Unstable	0.30	Unstable
RK3	1.61	1.26	1.42	1.08

We are using the RK3 method with a $5th$ spatial order, as it offers a high order of discretization without imposing overly restrictive stability constraints. For 3D advection, the Courant number is $1/\sqrt{3}$ times the 1D Courant number. To stay in numerical stability conditions, the following inequality should be satisfied:

$$\Delta t_{max} < \frac{C_{r_{theory}}}{\sqrt{3}} \cdot \frac{\Delta x}{u_{max}} \tag{2.2}$$

In this work, the ARW solver's adaptive time step is utilized. The equations for updating the integration time step are detailed in Section 3.3.3 of Skamarock et al. (2008). A target maximum Courant number $C_{r_{target}}$ must be specified as input. A horizontal Courant target of 0.84 and a vertical Courant target of 1.2 was imposed on the $3rd$ domain. When dealing with real-data, u_{max} can reach 100 $m.s^{-1}$ horizontally for jet streams. With a horizontal resolution of $\Delta x = 800$ m, stability conditions impose $\Delta t_{max} < 6.72$ s. In practice, the adaptive time step stabilizes at $\Delta t_{D03} = 4.8$ s for domain 3, $\Delta t_{D02} = 14.5$ s for domain 2, and $\Delta t_{D01} = 58.0$ s for the first domain.

2.2.2.4 | CPU Parallelization

WRF-Chem was compiled to support parallelized computation using OpenMPI. To optimize computation time, we want to use as many cores as possible. This maximum number depends on the size of the smallest domain, in this case D03. During computation, the domain grid is broken down into tiles, each of which is allocated a processor. For a single core, a minimum number of 10 cells must be processed in both directions. Simulations are carried out using high performance computing at the Mesocentre of the University of Aix-Marseille. The maximum number of cores per node on the partition allocated to the project is 32. For a domain of 102×102 cells, the maximum number

of cores allowed is 64, shared between 2 nodes. This is the configuration adopted in the simulations presented in this manuscript.

2.2.3 | Parameterization

The present section gives a brief description of the different physics parameterization used as a set-up. Note that some choices were limited for a matter of compatibility with MOZART and aerosol schemes.

2.2.3.1 | Land Use

For land use (LU) data, IGBP MODIS Noah (Moderate Resolution Imaging Spectroradiometer) from 2010 was used. It identifies 21 land categories based on satellite imagery. Details of LU categories are shown in appendix A. The dataset has a spatial resolution of 30" ($\approx 800\ m$).

2.2.3.2 | Radiation

Radiation models estimate the atmospheric temperature profile and various radiative fluxes.

Shortwave radiation, primarily in the ultraviolet (UV) range, originates mainly from the Sun, while longwave radiation, in the infrared (IR) range, is emitted by the Earth's surface. These two types of waves are absorbed differently by the components of the atmosphere. Consequently, shortwaves and longwaves are treated separately. The Rapid Radiative Transfer Model (RRTMG) (Iacono et al., 2008) is used for both. The wavelength spectrum is subdivided into several bands. RRTMG is a 16-band model. The absorption of gases present in the atmosphere depends strongly on the frequency. Water vapor, CO_2, and ozone are the main gases absorbing IR radiation. Calculating the absorption coefficients for each species at every frequency is costly. RRTMG reduces the number of radiative transfer operations by subdividing the bands into intervals with strong absorptions by certain species, ensuring precise calculations for major absorbing species and less precision for minor ones.

Clouds have a significant effect on IR radiation, as they absorb most of it. Their influence is accounted for by RRTMG, which interacts with the Microphysics and Cumulus models. The presence of a cloud in a grid cell is considered binary for this model. Surface emissivity is defined in the Land Use tables based on soil type. For instance, urban areas have a lower emissivity compared to oceans and seas. The albedo strongly

depends on the soil type, with water having a lower albedo compared to snow. The shortwave radiation feedback option is turned on, which means that upward transfers are computed, and aerosols direct effects are considered.

2.2.3.3 | Surface

The schemes discussed here concern the atmospheric surface layer, that usually extends from surface to 100m, where turbulent fluxes and stress can be considered constant with height.

Surface parameterization is a two step process. First, a surface layer scheme computes exchange coefficients and friction velocities. These values are then used by land-surface models (LSMs) to compute surface heat and moisture fluxes. These fluxes then serve as the lower boundary condition for vertical transport in the PBL schemes.

The chosen surface layer scheme, by Janić (2001), is based on similarity theory. Vertical wind speed profile in the surface layer is determined by:

$$V(z) = \frac{u_*}{k}\left[\ln\left(\frac{z}{z_0}\right) + \phi(z, z_0, L)\right] \tag{2.3}$$

where z_0 is the roughness length derived from land use parameterization, V is the horizontal wind speed, u_* is the friction velocity, $k = 0.4$ is the Von Karman constant, and ϕ is the stability function that depends on L, the Monin-Obukhov length. ϕ cancels out under stable conditions of the surface layer. The friction velocity is constant in the surface layer and depends on V_1, the wind speed at the first vertical level:

$$u_* = \frac{kV_1}{\ln(z_1/z_0) - \phi(z_1, z_0, L)} \tag{2.4}$$

Analogous variables are defined for temperature and humidity:

$$\theta_* = \frac{k\Delta\theta}{\ln(z_1/z_0) - \phi(z_1, z_0, L)} \tag{2.5}$$

$$q_* = \frac{k\Delta q}{\ln(z_1/z_0) - \phi(z_1, z_0, L)} \tag{2.6}$$

Δ refers to the difference between the value at the first level and at the surface. We can then determine the surface fluxes of momentum, heat, and moisture with the respective equations:

$$\tau = \rho u_* u_*; \quad H = \rho c_p u_* \theta_*; \quad E = \rho u_* q_*$$

The Noah scheme is used as a LSM. It models the soil in 4 layers and computes the

temperature and humidity of these four layers based on the fluxes calculated by the surface layer model and the radiative balances at the ground performed by the radiation model.

2.2.3.4 | PBL and Turbulence

The planetary boundary layer lies between the surface layer and the stable atmospheric layer. PBL schemes calculate vertical fluxes for heat, moisture, and momentum. The MYNN scheme, selected for this set-up, determines exchange coefficients by solving the turbulent kinetic energy (TKE) equation. Vertical exchange coefficients are given by $K_v = q \times l \times S$, where $q^2/2$ is the TKE, l is the master length scale and S is a stability function. Details on the definition of l and S are provided in Nakanishi and Niino (2009). For horizontal diffusion, we use the Smagorinsky model in two dimensions at first order, which is a LES (Large Eddy Simulation) turbulence model. Its principle is to numerically resolve large eddies while modeling the smaller, unresolved eddies. The length scale that differentiates these two types of eddies is defined in this model as $l_s = C_s\Delta$, where Δ is the square root of the product of the two grid dimensions, and C_s is the Smagorinsky constant with a typical value $C_s = 0.25$.

2.2.3.5 | Cumulus

Cumulus models represent the vertical convection effects responsible for cloud formation. They determine temperature and humidity profiles in the upper part of the planetary boundary layer and interact with the surface model during precipitation events. The chosen model is Grell-3 (Grell and Dévényi, 2002), particularly suited for mid-latitudes. It activates when a grid column contains a convective cloud. The model sets a threshold to identify a convective column and estimates the velocity of these movements.

2.2.3.6 | Microphysics

The role of these schemes is to accurately model water vapor, cloud formation, and precipitation processes. The Morrison scheme is used for the set-up (Morrison et al., 2005). It accounts for the number and mass mixing ratios of cloud ice, rain, snow, and graupel/hail, as well as the mixing ratios of cloud droplets and water vapor. By utilizing a two-moment prediction approach (number and mass mixing ratios), this scheme provides a more precise representation of particle size distributions, which is crucial for effectively simulating the evolution of clouds and precipitation. Moreover, the Morrison scheme is particularly relevant for taking in consideration the impact of aerosols on droplets.

2.2.3.7 | Urban Canopy

WRF provides an option for a precise representation of surface exchanges in urban areas with the Urban Canopy Model (UCM) (Kusaka et al., 2001). It models the urban surface as a single layer, with each surface unit being an urban canyon, meaning a street bordered by two side walls representing buildings. The UCM is two-dimensional and assumes the canyon to be infinite. It estimates momentum fluxes, radiative fluxes, and sensible heat fluxes. The solar angle and building heights are considered to incorporate shading effects into the simulation.

2.2.4 | Initial and Boundary Conditions

2.2.4.1 | MEGAN

In order to better represent biogenic emissions, the preprocessing tool Model of Emission Gases and Aerosols from Nature (MEGAN) is added to the set-up. These emissions are primarily biogenic VOCs (BVOCs), with isoprene being the most prevalent species. Past studies have highlighted that about 90% of global VOCs emissions originate from Nature. These species play an important role in atmospheric chemistry, particularly for photochemical reactions. Therefore, BVOCs could have a significant impact on the formation of ozone and SOA and ultimately air quality. MEGAN allows WRF-Chem to include biogenic isoprene emissions with a spatial resolution of $1{\times}1\ km^2$. The net emission flux of isoprene and other aerosols or gases released into the atmosphere from ecosystems at a given location and time is calculated as follows:

$$E = \epsilon \cdot \gamma \cdot \rho \tag{2.7}$$

where ϵ is the emission estimate for standard conditions in the canopy. γ is the emission activity factor accounting for variations from standard conditions. ρ represents the gains and losses occurring within the canopy. The standard conditions are described in Guenther et al. (2006). Emission factors are calculated by plant functional type (PFT) and then summed according to the local PFT distribution. These emission factors are estimated from laboratory measurements and in situ observations. This information is combined with land cover data derived from satellite images, ground inventories, and ecoregion descriptions. Physical and biological factors are initialized using monthly data.

2.2.4.2 | MOZBC

The MOZBC preprocessing tool is utilized to set the initial and boundary conditions for gas-phase species and aerosols. The dataset for this process is derived from the output of the global chemical model WACCM. Boundary conditions are applied exclusively to the outermost domain, D01, while initial conditions are mapped on all three domains.

2.2.4.3 | Nudging

Nudging, or four-dimensional data assimilation (FDDA) is a method available in the ARW solver in order to keep simulations close from observation or, in our case, analysis data. This is done by adding a relaxation term in the dynamic equations. Considering a scalar variable θ, nudging is given by:

$$\frac{\partial \theta}{\partial t} = F(\theta) + G_\theta W_\theta (\hat{\theta}_0 - \theta) \tag{2.8}$$

with $F(\theta)$ the regular term without nudging, G_θ is a time-scale controlling the nudging strength, and W_θ is a weight in time or space to control the nudging, while $\hat{\theta}_0$ is the interpolated analysis field value towards which the nudging relaxes the solution. The analysis data used for nudging are the same as the ones processed by WPS, i.e. GDAS/FNL 0.25°. Only horizontal wind components u and v, potential temperature θ and water vapor mixing ratio q_v are nudged. In order to avoid a prevalence of the nudging term over advection and physics mechanisms represented by $F(\theta)$, the time-scale parameter was set to $G_\theta = 0.0003\ s^{-1}$, which is the recommended value when using nudging. The W_θ matrix was configured to restrict nudging to levels above the 10th level. In this study, when nudging is applied, it remains active for the entire duration of the simulation.

2.3 | Marine Traffic in Marseille

This section describes the marine traffic data provided by the GPMM for 2021. AtmoSud is acknowledged for sharing this dataset. Note that for this period, COVID-19 may have impacted the regular traffic of the port. The dataset includes all arrival and departure calls registered in both basins between January 1, 2021, and January 1, 2022. For each call, ship characteristics are provided, including GT and type, as well as the shipping route used to enter the port, and the berthing site. Table 2.4 provides an overview of the number of calls per ship type, as well as the average docking time and GT. Original ship labels have been retained for reliability, and translations from French to English are provided in appendix B. Boats with highest GT are cruise ships (PAQUEBOT).

Table 2.4: General call statistics registered at the port of Marseille-Fos in 2021.
Docking Time (in hour) and GT are the average for each ship type.

Ship Type	#	Docking Time	GT
AUTRES VRACS	495	117	7539
BARGE OCEA.VERTI	99	208	1
BARGE OCEAN.RORO	2277	41	458
CAR CARRIER	7755	18	35959
CAR-FERRIE	11583	27	34679
CARGO CONVENT.	43263	65	4067
CARGO REFRIGE.	66	122	12401
CARGO SPECIALIS.	759	65	5585
GAZ LIQUEFIES	22374	45	22846
MINERALIER	1914	98	41032
NAV. CIT. VRAC ALIM	297	56	3569
NAV. DE PLAISANCE	1248	9	1002
NAV. NON CLASSE	99	81	334
NAV. PROD. CHIM	100485	26	5687
NAV. TRAV. MARITIME	594	139	601
NAVIRE DE GUERRE	66	58	1801
NAVIRE DE PECHE	33	116	160
NAVIRE RO-RO	47751	19	28048
NAVIRE SCIENTIFI	429	39	634
P. CONT. CELLULAIRE	43065	26	41147
P.CONTE.POLYVALE.	1782	32	26863
P.CONTENEURS ROUL	1254	25	49332
PAQUEBOT	14289	154	72358
PETROLIER	62865	34	21397
PORTE-BARGES	99	14	12642
REMOR.RAVITAI	825	35	727
ROPAX	7755	20	28448
VRAQUIER SOLIDE	14190	74	21503
ALL	387711	39	21042

The average docking time of 154h for a cruise ship is particularly long. In fact, during COVID-19, many cruise ships remained at berth until the pandemic was over and cruising activities resumed. Chemical tankers (NAV. PROD. CHIM) account for the largest share of calls, representing over a quarter of the total number. This highlights the significant activity of chemical plants in Fos. Most calls in Fos are for small boats (i.e., low GT),

as shown in Fig.2.5. In Marseille, there is a noticeable peak in the distribution at GT ≈ 28000, corresponding to the significant activity of the NAVIRE RO-RO category in the port of Marseille.

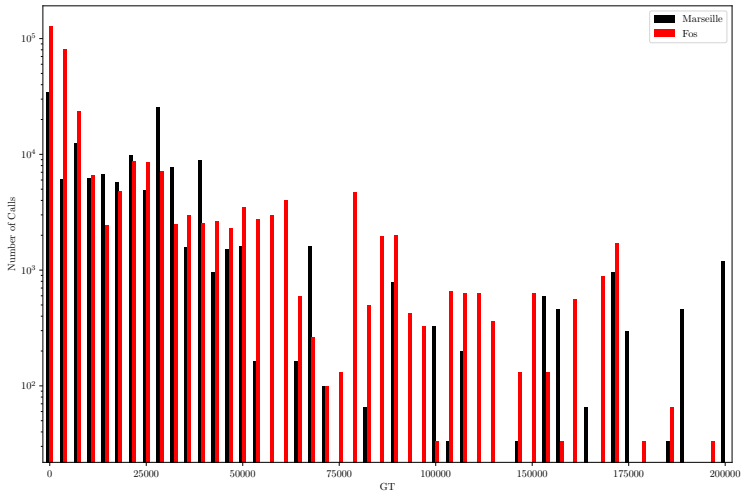

Figure 2.5: GT distribution in Marseille (in black) and Fos (in red). Y-axis is in log scale.

Fig. 2.6 shows the average number of calls per hour for both arrivals and departures. In Fos, activity is higher during the daytime, although there is still important activity at night. In Marseille, there is a significant peak for arrivals at 7-8h and for departures at 18-19h, with very little activity during the night. This pattern aligns with cruise and ferry operations, which typically navigate during the night and remain berthed during the day.

2.4 | Emissions

This Section presents the pre-processing of emission data for the WRF-Chem model. Emissions are incorporated into the model using `wrfchemi_D0n_YYYY-mm-dd_HH:MM:SS.nc` files, which only allow for surface emissions. In this study, one file is used per hour of prediction. Ship emissions are distinguished from other emissions, which are considered here as background emissions.

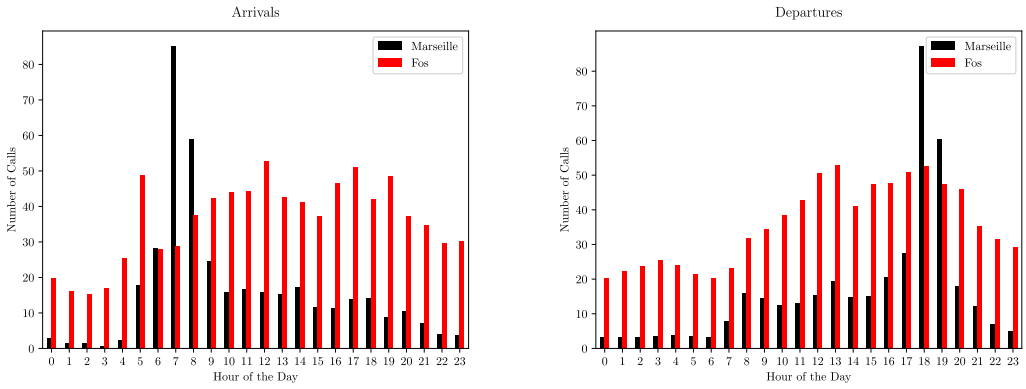

Figure 2.6: Average number of calls per hour.

2.4.1 | Ship Emissions

Ship emissions are estimated from the GPMM marine traffic dataset described in Section 2.3. This estimation was done by AtmoSud, on the basis of a standardized methodology, known as the PCIT (Pôle de Coordination National sur les Inventaires d'Émission Territoriaux) emission guide (Kauffmann et al., 2012), which is applied by all AASQAs in France. The methodology for ship emissions in the PCIT guide utilizes Eq.2.1, with most ship characteristic values sourced from EMEP/EEA tables. Estimated species are CO, NO, NO_2, SO_2, NH_3, BC, $PM_{2.5}$ and PM_{10}.

2.4.1.1 | Methodology

Three phases are considered when estimating ship emissions: the hotelling phase, when the ship is at berth; the manoeuvring phase, when the ship approaches or leaves the port area; and the cruising phase, when the ship is navigating. Emissions for each phase were calculated over its entire duration. To integrate emissions from t_h to t_{h+1}, which corresponds to a single input file, the proportion of phase time within this interval is calculated and applied to total phase emission values.
Hotelling emissions are considered as point sources. Mapping hotelling emissions onto the WRF grid simply involves dividing the emissions by the surface area of a grid cell.
For the manoeuvring phase, emissions are computed along the boat's trajectory. To map these emissions onto the grid, the trajectory's intersection with each grid cell is calculated. The emission allocated to a single cell corresponds to the percentage of this intersecting portion relative to the entire trajectory length, applied to the total emissions

Table 2.5: Assumptions for the ship stack height. Source: AtmoSud

Ship Type	Stack Height m	WRF Level
Tanker	29	2
Container	44	3
Cargo	34	3
Ferry/Roro	23	2
Cruise	47	3
Pleasure Craft	30	2
Other	29	2

for the phase. As maneuvering time averages 35 minutes, emissions are calculated over the entire length of the trajectory at once, rather than taking into account the boat's real-time speed and position along its path.
However, the cruising phase involves longer distances and times. Ships's real time position is computed and associated emissions are integrated hourly. Assumptions for the average cruise speed by ship type were taken from Table 3-14 of EMEP/EEA.
Emissions can be inferred at different staggered vertical levels, which allows for the consideration of various ship stack height, depending on ship type. Table 2.5 provides the assumptions on ship stack height and the corresponding vertical levels at which emissions were input. Only the 2nd and 3rd levels were used, as their respective average altitudes are 17 m and 50 m, with the 1st level being ground level.

Fig.2.7 depicts an emission field of $PM_{2.5}$ used as an input for WRF-Chem. All three phases are represented.

2.4.1.2 | Ship Emission Statistics

Ship emissions were calculated for the year 2021. This statistical overview focuses solely on the Hotelling and Manoeuvring phases, as cruise emissions typically have less direct impact on the population. Fig.2.8 compares average hourly emissions from Marseille and Fos for each month. Globally, there is no significant difference among the months, although emissions in spring and summer are slightly higher compared to fall and winter. Note that low emissions in January might be in part due to data extraction limitations, which excluded emissions starting before January 1, 2021, but potentially emitting afterward.

Fig.2.9 details emissions per ship category. No significant difference between days of

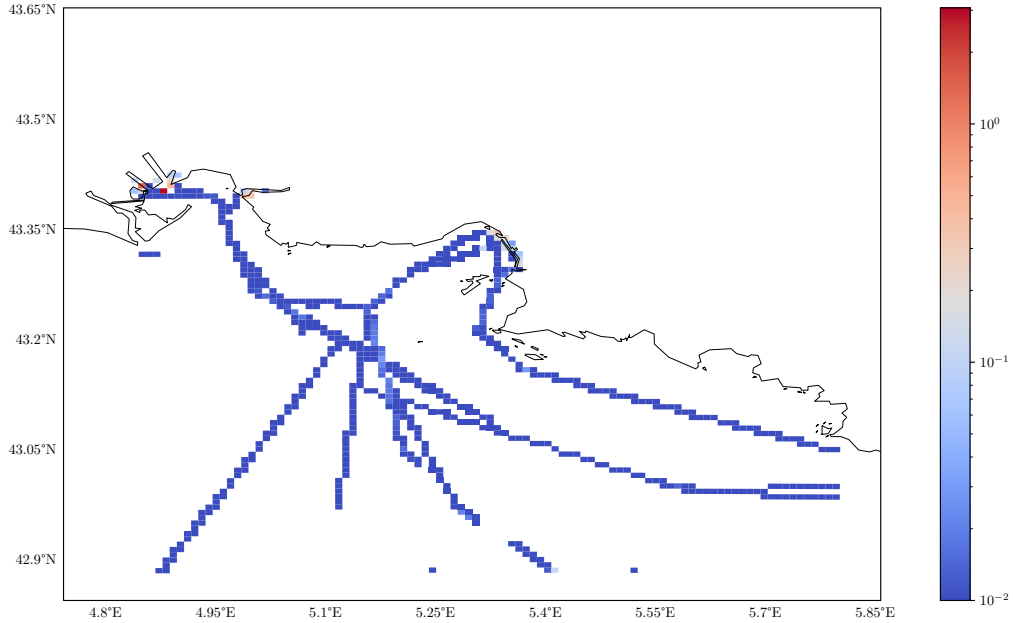

Figure 2.7: 24-hour averaged $PM_{2.5}$ estimated ship emissions from 2nd and 3rd levels for 2021-07-01 in $\mu g \cdot m^{-2} \cdot s^{-1}$ (log scale).

the week is noticed, except for cruises, corresponding to an intensified tourism activity during the weekends. Hourly emissions are higher during daytime, from 5 am to 6 pm, driven by the activity of specific ship types such as cruises, Ro-Ro, and ferries. Hence, this trend is more pronounced in Marseille than in Fos.

2.4.2 | Background Emissions

Background emissions refer to all emissions excluding those from boats. It is very difficult to accurately represent all emission sources over a mesoscale domain. During this PhD project, the low-resolution EDGAR-HTAP inventory was initially used, and then replaced by the high-resolution CIGALE inventory.

2.4.2.1 | Methodology

AtmoSud has produced a detailed emission inventory in the PACA region (CIGALE, 2019). Temporal profiles were applied to the yearly averaged emissions to capture yearly,

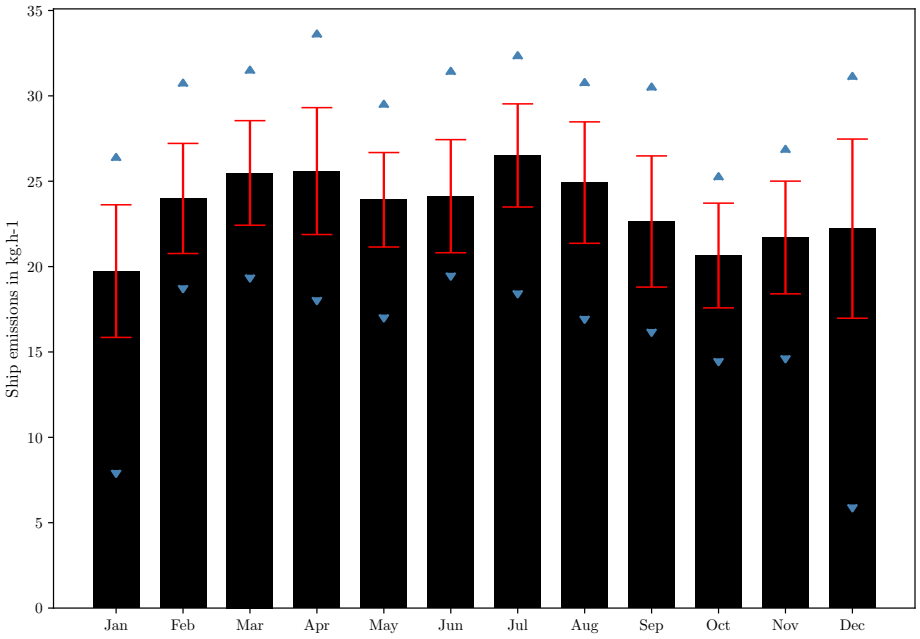

Figure 2.8: Ship emission per month. Min and Max daily average values in blue. Standard deviation in red.

monthly, daily and hourly fluctuations. Only CO, NO, NO_2, SO_2, NH_3, BC, $PM_{2.5}$ and PM_{10} were used from this inventory. The dataset provided by AtmoSud gives emissions per sector on a regular grid with a spatial resolution of $1{\times}1$ km^2. Emission fields are projected on the WRF grids of the three nested domains. A 2D spatial interpolation needs to be computed between CIGALE grid and WRF grids. The implemented code is based on the method of area intersection illustrated in Fig.2.10. To compute the concentration on the WRF grid, the following formula is used:

$$c = \frac{\sum A_i \times c_i}{\sum A_i} \tag{2.9}$$

With c_i the concentrations of all intersecting cells and A_i the areas of the intersections.

Fig.2.11 represents background emissions projected onto the DOI. Due to a lack of information concerning background emission stack heights, we assumed all sources to be at ground level. Note that ship emissions were excluded from the original CIGALE inventory dataset for the purposes of this study, as we aim to replace them with hourly ship emissions derived from GPMM traffic data.

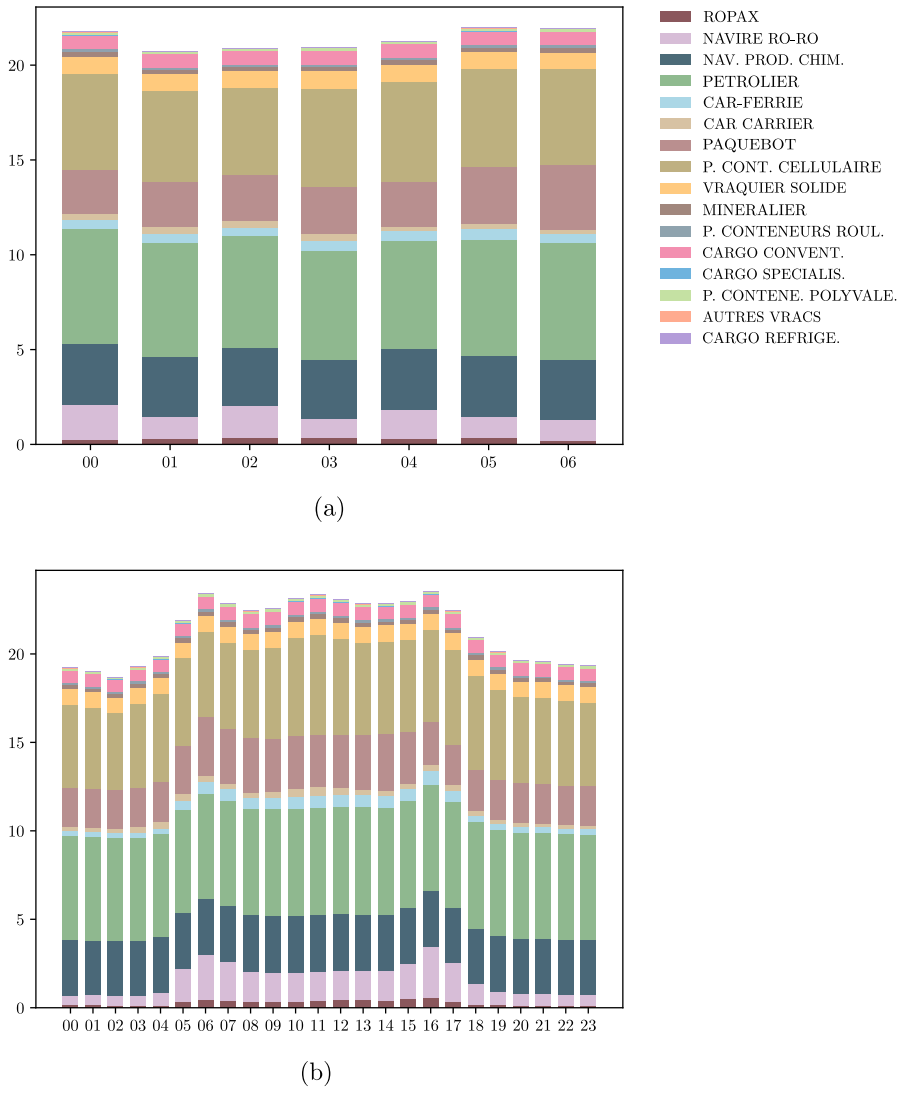

Figure 2.9: Ship emission by category. (a) represents weekday average and (b) represents hourly average.

2.4.2.2 | Back Ground Emission Statistics

Table 2.6 shows the annual emissions by sector. Background emissions are taken from the CIGALE 2019 inventory, while marine emissions are calculated on the basis of the GPMM dataset. The share of maritime emissions is significantly higher for NOx compared to $PM_{2.5}$, PM_{10}, and SO_2 across the entire domain (32%, 5.4%, 4.0%, and 2.5%

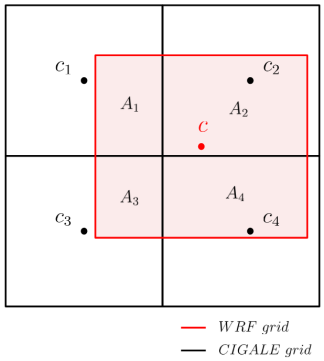

Figure 2.10: 2D interpolation method using area intersection.

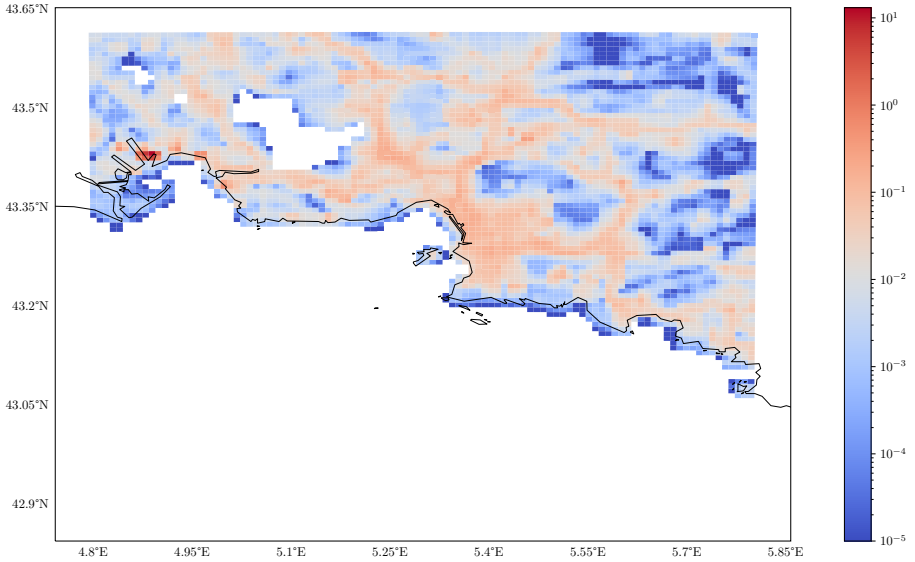

Figure 2.11: 24-hour averaged $PM_{2.5}$ background emissions for 2021-04-22 in $\mu g \cdot m^{-2} \cdot s^{-1}$ (log scale).
High emissions in the western port correspond to the refineries and chemical plants of Fos-sur-Mer.

respectively), as shown in Fig.2.12. This is particularly notable in Fos, where NOx emissions total 8890 tons, representing nearly half of the total NOx emissions in W3. For all pollutants, industrial sources contribute significantly in W3, primarily due to the high activity of chemical, oil, and gas plants in Fos. For SOx, the combined contributions from the Energy and Industry sectors exceed 97%.

Fig.2.12 identifies three major contributors to PM emissions:

- Industrial sources in Fos and Marseille.
- The residential, commercial, and institutional sector, primarily due to fireplace emissions.
- Non-maritime transport, largely represented by road traffic.

Table 2.6: Yearly emissions by activity sector. D03 was divided into two areas by a longitudinal line at 5.10°E. W3 and E3 zones are respectively to the west and east of the demarcation line.

Zone of Interest	Activity Sector	$PM_{2.5}$ *Ton*	PM_{10} *Ton*	SOx *Ton*	NOx *Ton*
D3	Energy	142.52	196.40	5537.68	3477.99
	Industry	1210.17	2086.43	5690.03	9209.16
	Res./Com./Inst.	1380.52	1419.73	188.61	1240.69
	Non-maritime Transport	819.76	1332.09	37.24	10969.76
	Maritime Transport	205.45	216.86	289.78	11910.54
	Agriculture	40.77	86.25	2.30	328.27
	Waste Management	3.83	4.73	23.34	48.09
	Biogenic	24.59	30.06	1.31	72.47
E3	Energy	2.46	3.58	108.17	242.02
(5.10°E to 5.80°E)	Industry	409.48	946.39	1728.16	3173.03
	Res./Com./Inst.	1214.79	1249.50	169.37	1132.04
	Non-maritime Transport	742.45	1201.35	33.95	10040.03
	Maritime Transport	52.49	55.40	74.23	3020.42
	Agriculture	20.46	55.42	1.23	238.34
	Waste Management	3.30	3.76	6.92	24.84
	Biogenic	20.97	25.63	1.11	54.78
W3	Energy	140.06	192.82	5429.51	3235.97
(4.80°E to 5.10°E)	Industry	800.69	1140.04	3961.87	6036.13
	Res./Com./Inst.	165.73	170.23	19.24	108.65
	Non-maritime Transport	77.30	130.74	3.29	929.72
	Maritime Transport	152.96	161.46	215.55	8890.12
	Agriculture	20.31	30.83	1.07	89.93
	Waste Management	0.54	0.98	16.43	23.26
	Biogenic	3.62	4.43	0.19	17.69

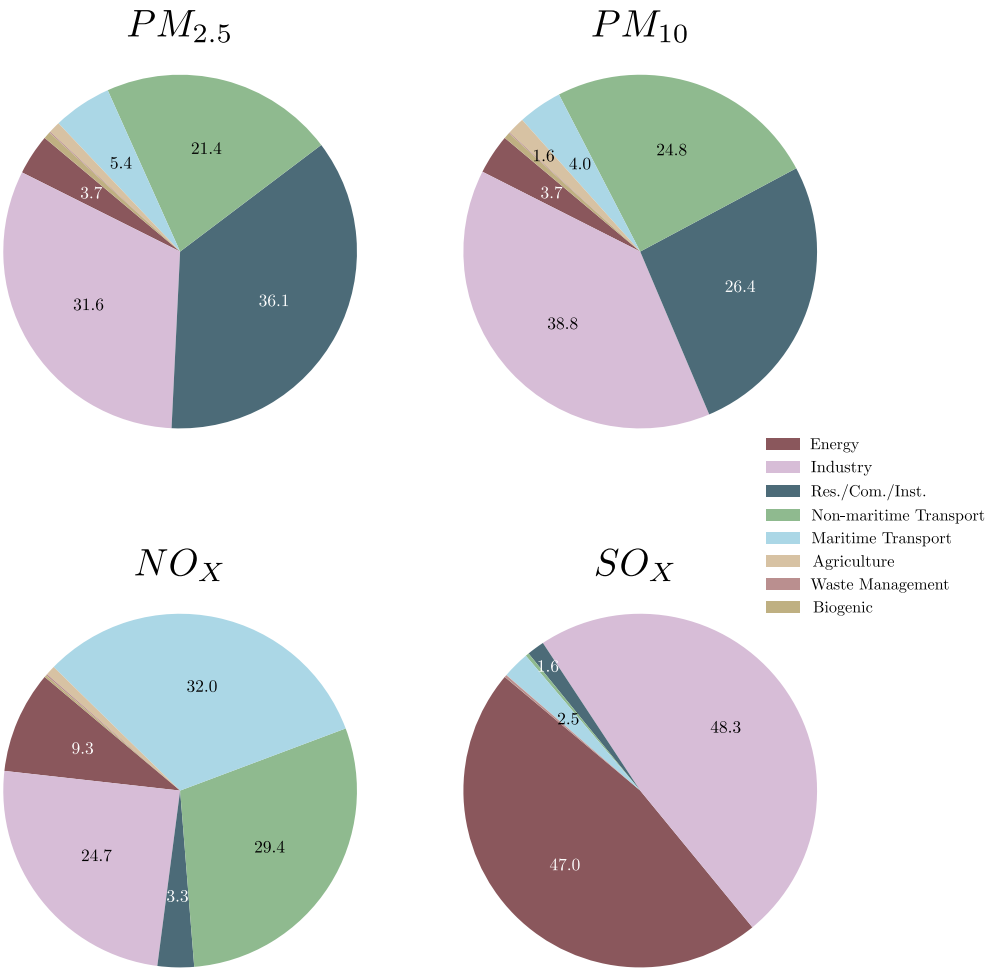

Figure 2.12: Share of emission sectors in the DOI

2.5 | Model Validation

For validation purpose, predictions were compared with observations using different metrics:

- Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.10}$$

where y_i = Observed value; $\hat{y}_i$ = Predicted value; n = Number of observations.

- Root Mean Squared Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.11}$$

- Pearson Correlation Coefficient (r):

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}} \tag{2.12}$$

where $\bar{y}$ = Mean of observed values; $\bar{\hat{y}}$ = Mean of predicted values.

- Normalized Mean Bias (NMB):

$$\text{NMB} = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)}{\sum_{i=1}^{n} y_i} \times 100\% \tag{2.13}$$

- Normalized Mean Error (NME):

$$\text{NME} = \frac{\sum_{i=1}^{n}|\hat{y}_i - y_i|}{\sum_{i=1}^{n} y_i} \times 100\% \tag{2.14}$$

2.5.1 | Validation on Meteorology

The meteorological part of the model has been compared with results from the operational forecast model Arome developed by Meteo France. Arome provides hourly data with a resolution of $0.01° \times 0.01°$ ($1.3\ km^2$). Although the study focuses on the year 2021, the validation process uses data from 2022 due to restricted access to certain meteorological data. The period of comparison extends from 2022-09-05 to 2022-09-18, with a forecast of 23 hours every day. WRF outputs were interpolated on Arome's horizontal grid using the cubic method of scipy.interpolate library on Python. The variables under comparison are: temperature (T2) and relative humidity (RH) at 2 m, pressure at sea level (PS) and the two horizontal components of wind speed U and V at 10 m. Table 2.7 presents the differences between WRF and Arome. They are calculated according to the following three metrics: Mean Bias Error (MBE), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The values in the table are the time-averaged metrics over 14 days of simulation. The differences on the domain are relatively small since the spatial average absolute temperature difference is 0.95 K. However, larger differences, approaching 7 K,

are observed over some areas as shown in Fig.2.13, which is a snapshot of the T2 Absolute Difference field from 2022-09-10 at 7 PM. These zones correspond to a high elevation of ground. The mismatch between the topology of the two models induces temperature differences. This also explains the high local differences on the surface pressure fields of approximately 24 hPa, although the RMSE over the entire domain remains low at 2.01 hPa. The models seem to differ more in terms of relative humidity, with a high RMSE of 9.70% and a low MBE of 0.74%, indicating high variability.

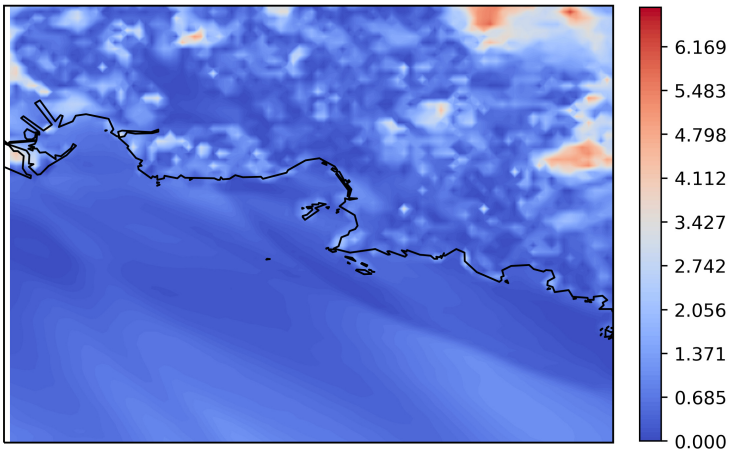

Figure 2.13: T2 absolute difference between WRF and Arome for 2022-09-10 at 7 PM (UTC+2).

Table 2.7: Comparison between WRF and Arome from 2022-09-05_00:00:00 to 2022-09-18_23:00:00

Variable	T2 K	U $m \cdot s^{-1}$	V $m \cdot s^{-1}$	RH %	PS hPa
MAE	0.95	1.67	1.99	7.65	1.12
MBE	0.04	-0.37	0.40	0.74	-0.02
RMSE	1.26	2.14	2.54	9.70	2.01

The models were then compared to observations from the meteorological station Marignane over the same period, located at (43.44°N, 5.23°E). The station provides data on the same variables as seen above. The trends of the predicted variables over the 24-hour simulation period are found to be close to the station measurements, according to the time series on Fig.2.14. Analysis on the 14-day period is summarized in Table

2.8. Temperature is slightly overestimated by both WRF and Arome, with average errors of 0.22 and 0.35 K respectively. WRF performs better for wind speed, although both models underestimate wind speed by approximately 4 $m \cdot s^{-1}$. Arome has better results regarding relative humidity in Marignane. Both pressure surfaces RMSE are smaller than 1 hPa, which we believe to be accurate. Figures from both Table 2.7 and Table 2.8 confirm that WRF and Arome provide forecasts with similar errors over the DOI. Therefore, the results obtained in this section allow us to consider that the meteorological predictions of the WRF model are sufficiently consistent and reliable for this case study.

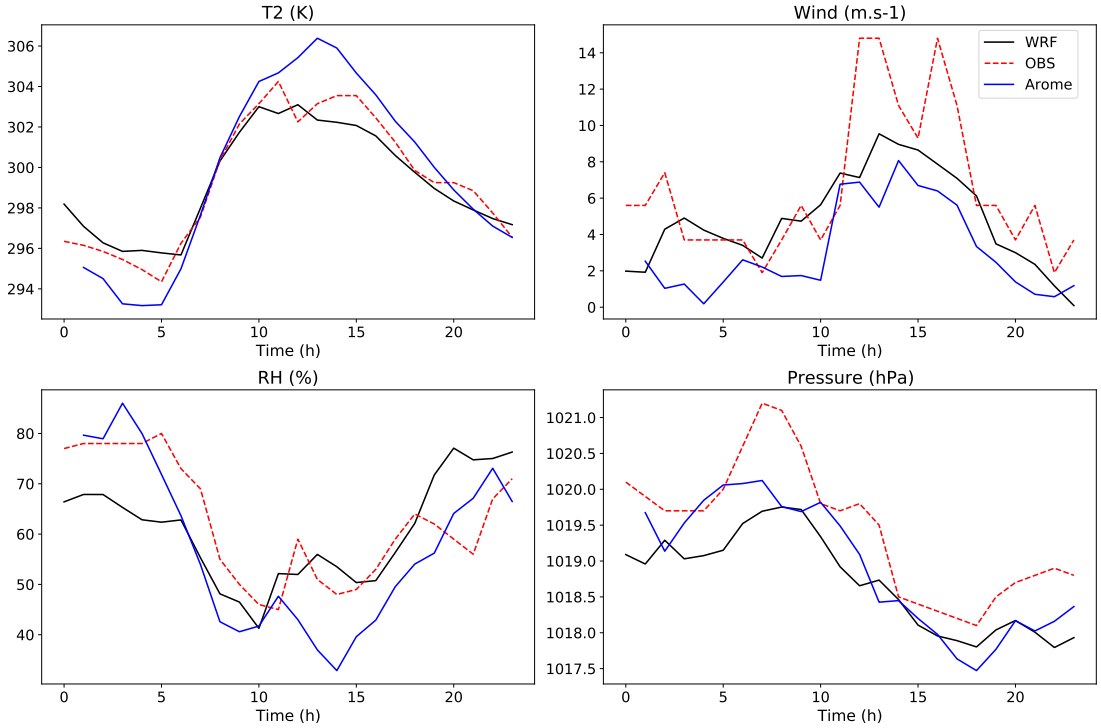

Figure 2.14: Time series of WRF, Arome and Observations from Marignane for 2022-09-05. ($Wind = \sqrt{U^2 + V^2}$)

Table 2.8: Errors of WRF and Arome with respect to observations from Marignane
from 2022-09-05 to 2022-09-18

Variable		T2	Wind Speed	RH	PS
Unit		K	$m \cdot s^{-1}$	%	hPa
WRF	MBE	0.22	-3.96	-5.38	-0.58
	RMSE	1.35	5.15	9.73	0.79
AROME	MBE	0.35	-4.91	-2.37	-0.38
	RMSE	1.01	5.80	7.82	0.62

2.5.2 | Validation on Pollution

The chemical predictions from both setups (MOZCART and MOZAIC) were compared with ground measurements from 27 monitoring stations in the AtmoSud network, represented in Fig.2.15. The validation period, from 2021-05-31 to 2021-07-31, coincides with the simulation period for the Real Case scenario and the Scipper campaign, during which temporary sensors were deployed along the East basin of the port.[1]

Table 2.9 and Table 2.10 present the metrics for the MOZCART and MOZAIC models, respectively. The MOZAIC model shows better agreement with observations for $PM_{2.5}$, PM_{10}, and O_3 compared to MOZCART. The difference in performance for aerosols may be attributed to the varying complexity between the bulk and sectional approaches. For O_3, the improved results in the MOZAIC setup could be due to the inclusion of chemical boundary conditions, which were not considered in the MOZCART setup. For SO_2 and NO_2, the latter performs slightly better, though both predictions exhibit poor accuracy. SO_2 is consistently overestimated by both models, suggesting a possible overestimation of SO_2 in the emission inventory. NO_2 is underestimated by MOZCART but overestimated by MOZAIC. A short run of MOZCART with chemical boundary conditions also showed an overestimation of NO_2, which may suggest that the NO_2 concentration in the input conditions provided by WACCM is overestimated in the region of Marseille.

Both models perform better in Marseille than in the full domain for all species. Both MOZAIC and MOZCART models show their best performance at the Longchamp station. This station, located in a park, is unique in that it is not directly influenced by nearby pollution sources, providing a baseline for background pollution levels. Additionally, Longchamp is the only station providing ground measurements for all five species considered for validation, making it a critical reference point for assessing the models'

[1] Although the simulation period was aligned with that of the Scipper campaign, sensor data were not accessible in due time for the manuscript submission. Consequently, validation relies solely on data from the AtmoSud monitoring network.

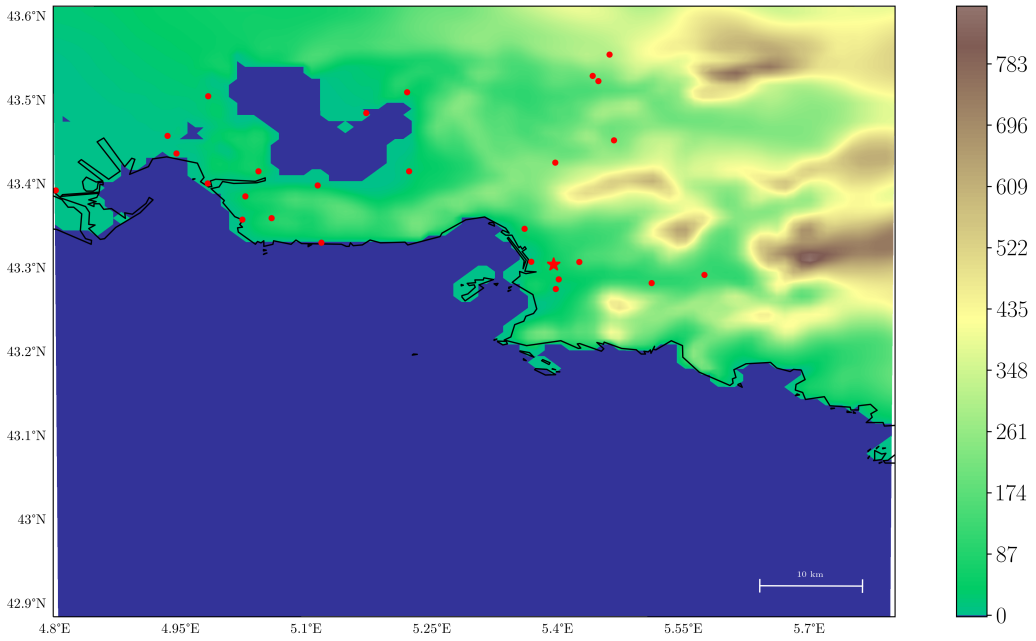

Figure 2.15: Locations of the 27 AtmoSud monitoring stations (red dots) with terrain height in m represented by the background color. The red star indicates the Longchamp station.

accuracy.

Time series at Longchamp station are shown in Fig.2.16. The PM time series reveals that MOZAIC successfully captures day-to-day fluctuations and extreme events, such as the one on 2021-06-20, which was not captured at all by MOZCART. O_3 trends are well represented by MOZAIC, with a slight negative bias, while MOZCART never reaches realistic values. This highlights the crucial importance of accurate ozone initial and boundary conditions when using WRF-Chem. Both models provide similar predictions for SO_2, even though MOZCART did not use chemical initial and boundary conditions. This similarity reinforces the hypothesis that there may be an overestimation of SO_2 in the emission inventory. Scatter plots in Fig.2.17 reveal high accuracy for $PM_{2.5}$, with extreme values remaining close to the 1:1 line. The scatter plot for SO_2 indicates that MOZAIC predicted numerous unrealistic extreme values.

The MOZAIC model generally shows better performance than MOZCART, accurately capturing day-to-day variations and extreme events for $PM_{2.5}$ and PM_{10}. However, MOZAIC struggles with unrealistic predictions of extreme SO_2 events. MOZCART, which lacks chemical boundary conditions, performs similarly for SO_2, suggesting poten-

Table 2.9: MOZCART Evaluation from 2021-05-31 to 2021-07-01 using AtmoSud network. # Total and # Marseille rows gives the number of stations available for the period and the pollutant considered within the domain D03 and the sub-domain of Marseille respectively. MAE and RMSE are in $\mu g.m^{-3}$.

Pollutants	$PM_{2.5}$	PM_{10}	SO_2	O_3	NO_2
# Total	12	13	11	9	11
MAE	8.35	16.25	8.11	75.80	17.07
RMSE	12.33	27.22	20.69	79.92	22.15
Correlation	0.26	0.06	0.09	0.40	0.22
NMB	55.04%	2.25%	258.07%	-92.19%	-38.25%
NME	93.25%	72.68%	335.83%	92.23%	74.08%
# Marseille	4	5	2	1	6
MAE	7.02	15.45	3.06	68.86	21.99
RMSE	10.00	21.86	5.57	73.46	28.28
Correlation	0.25	0.04	0.02	0.36	0.19
NMB	33.44%	-18.75%	171.71%	-89.58%	-60.74%
NME	75.45%	63.60%	233.24%	89.58%	72.25%

Table 2.10: MOZAIC Evaluation from 05-31-2021 to 07-01-2021 using AtmoSud network. # Total and # Marseille rows gives the number of stations available for the period and the pollutant considered within the domain D03 and the sub-domain of Marseille respectively. MAE and RMSE are in $\mu g.m^{-3}$.

Pollutants	$PM_{2.5}$	PM_{10}	SO_2	O_3	NO_2
# Total	12	13	11	9	11
MAE	4.61	12.97	13.29	37.37	29.50
RMSE	6.31	21.05	28.43	47.76	38.05
Correlation	0.58	0.37	0.07	0.35	0.16
NMB	12.75%	-1.72%	479.46%	-39.81%	136.31%
NME	51.39%	56.83%	541.25%	45.44%	173.80%
# Marseille	4	5	2	1	6
MAE	4.28	13.86	3.57	29.65	24.84
RMSE	5.81	20.18	6.12	39.35	33.16
Correlation	0.62	0.31	0.02	0.37	0.11
NMB	3.64%	-15.71%	232.73%	-25.25%	38.64%
NME	45.68%	55.67%	276.29%	38.58%	93.24%

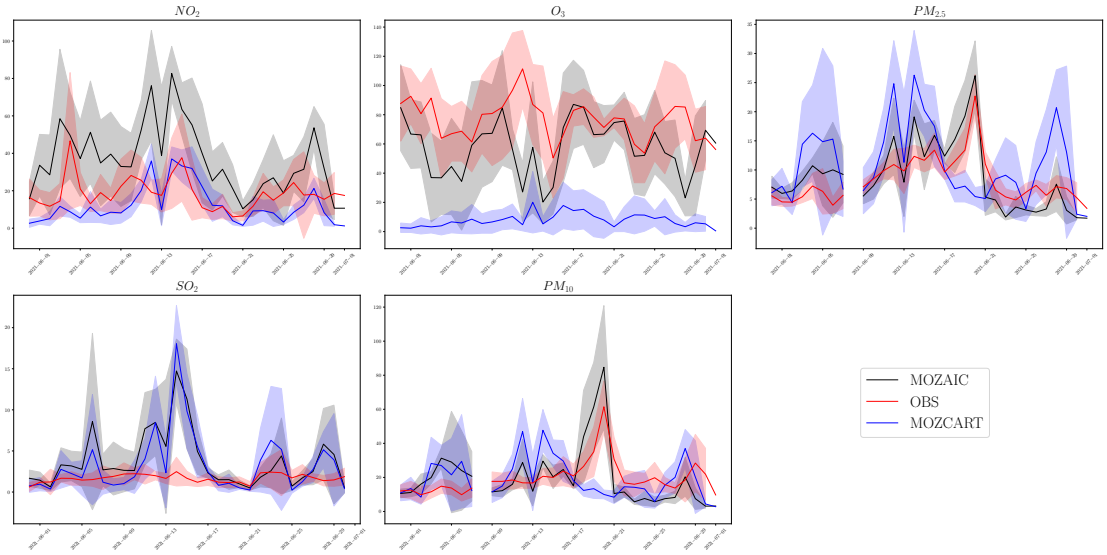

Figure 2.16: Time series of concentrations at Longchamp station for MOZAIC (in back), MOZCART (in blue), and Obs (in red) in $\mu g.m^{-3}$ from 2021-05-31 to 2021-07-01. Shaded areas represent the standard deviation of daily concentration values.

tial issues with the emission inventory. Overall, MOZAIC's more detailed configuration offers improved performance, but still requires further refinement to address specific limitations regarding NO_2 and SO_2.

2.6 | Real Case Study

This section presents a real case study of air pollution in Marseille and its surroundings. The aim is to quantify the impact of maritime traffic in Marseille using real-time ship traffic data. The analysis period matches that of the validation, spanning from 2021-05-31 to 2021-07-01. The MOZAIC setup, detailed in Table 2.1, was chosen for its superior accuracy compared to the MOZCART setup.

2.6.1 | Case Study

Table 2.11 presents detailed statistics on ship traffic during the period of interest. June 2021 experienced average ship traffic overall, with a slight increase in the presence of

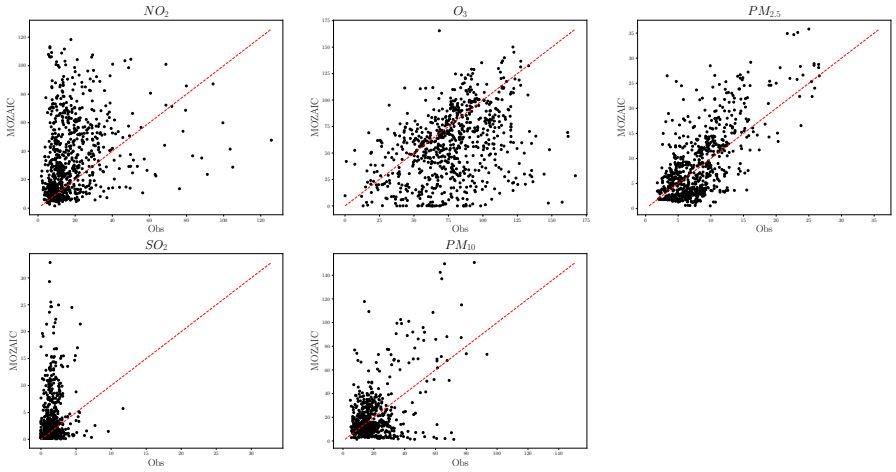

Figure 2.17: Scatter plots of MOZAIC predictions vs. Observations at Longchamp station from 2021-05-31 to 2021-07-01. The red dashed line marks the 1:1 line.

Chemical and Oil tankers (NAV.PROD.CHIM and PETROLIER). Additionally, ROPAX and CAR-FERRIEs showed increased activity as tourism began to rise in early summer.

Fig.2.18 shows average wind speed, complemented by local wind roses. Over the sea, wind speeds are higher, with a predominant wind direction associated with the mistral. A secondary, frequent southern wind pattern is also observed. Inland, wind speeds exhibit greater variability. In the Marseille area, the prevailing wind direction is from the southwest, whereas further inland, northern and eastern winds dominate.

Table 2.11: Call statistics registered at the port of Marseille-Fos from 2021-05-31 to 2021-07-01.
Docking Time (in hour) and GT are the average for each ship type.
+/- column refers to the percentage point difference with monthly average (i.e. 8.33% of total calls for 2021).

Ship Type	#	+/-	Docking Time	GT
AUTRES VRACS	33	-1.6%	225	2090
BARGE OCEA.VERTI	33	+25%	240	1
BARGE OCEAN.RORO	66	-5.4%	153	250.5
CAR CARRIER	627	-0.3%	12.8	36487
CAR-FERRIE	1287	+2.8%	33.8	34620
CARGO CONVENT.	3795	+0.4%	74.5	3785
CARGO SPECIALIS.	66	+0.4%	48.7	4997
GAZ LIQUEFIES	1782	-0.4%	56.2	22518
MINERALIER	264	+5.5%	139	55444
NAV. DE PLAISANCE	32	-5.8%	146	1
NAV. PROD. CHIM	9471	+1.1%	27.6	5371
NAV. TRAV. MARITIME	66	+2.8%	240	36.5
NAVIRE RO-RO	4257	+0.6%	22.8	28071
NAVIRE SCIENTIFI	33	-0.6%	14.2	300
P. CONT. CELLULAIRE	3894	+0.7%	25.6	40451
P.CONTE.POLYVALE.	198	+2.8%	41.2	30829
P.CONTENEURS ROUL	99	-0.4%	21.5	47232
PAQUEBOT	1254	+0.4%	419	36630
PETROLIER	6402	+1.9%	38.1	19711
REMOR.RAVITAI	99	+3.7%	57.0	1249
ROPAX	759	+1.5%	23.2	27622
VRAQUIER SOLIDE	1023	-1.13%	85.3	13398

2.6.2 | Results

The contribution of ships is estimated using the "zero-out" method, which involves running two simulations, one with ship emissions included and one without. The difference between the resulting outputs provides an estimate of the ships' contribution to the total concentration for a given species. While ship emissions were introduced at the 2nd and 3rd levels (corresponding to 17 m and 50 m, respectively), the ship contributions were calculated at the first mass point level, approximately 8.5 m above ground level.
Table 2.12 presents the contribution of ships to the main pollutants, shown both in $\mu g \cdot m^{-3}$ and in %. The higher contributions within the Marseille sub-domain compared

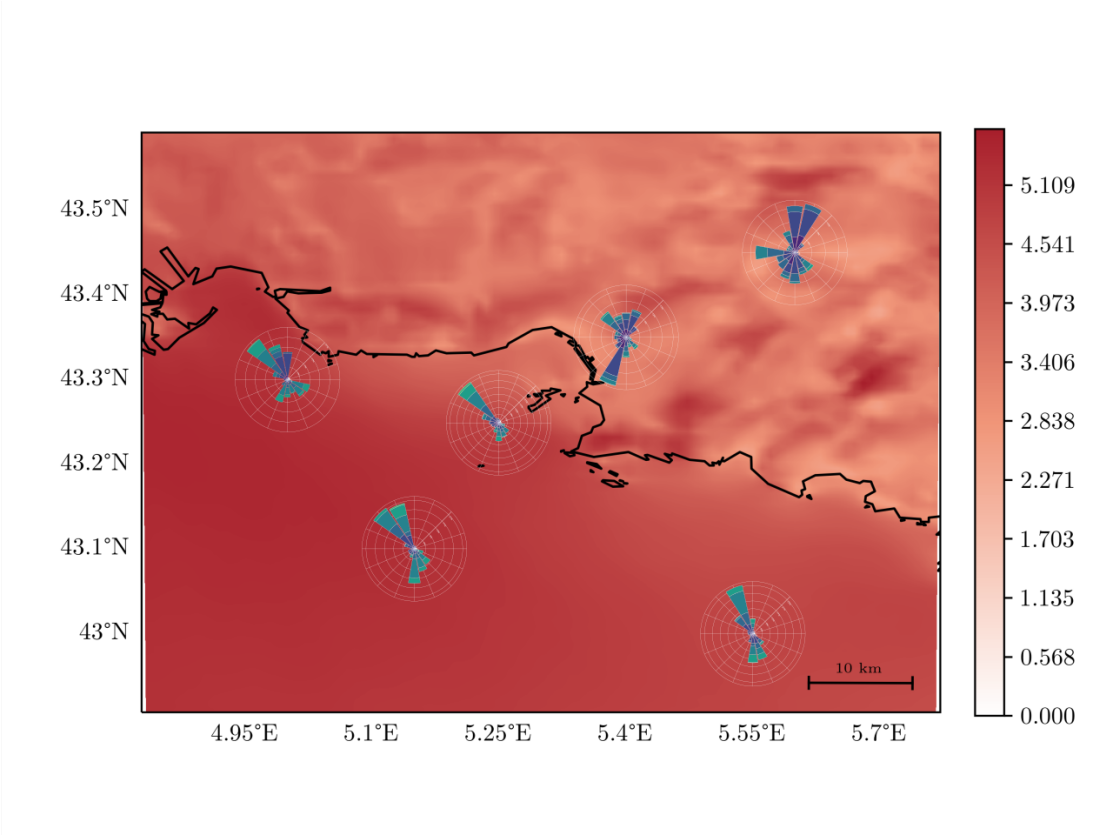

Figure 2.18: Average wind speed (in $m \cdot s^{-1}$) and wind roses at 6 different locations on D03 at first vertical level from 2021-05-31 to 2021-07-01.

to the overall domain of interest, are consistent with the presence of ship emissions from the EAST Basin within this sub-domain. $PM_{2.5}$ contribution is higher than that of PM_{10}, with respective values of 6.49% and 4.00%. The ship-emitted $PM_{2.5}$ concentration values may appear relatively low, with an average of less than 1 $\mu g \cdot m^{-3}$ in Marseille. However, these values are averaged over a month. When compared to the guidelines referenced in Table 1.1, during the month of June 2021, ships account for approximately 15% of the $PM_{2.5}$ annual limit. SO_2 contribution values are low, which is in line with the emission share shown in Fig.2.12. However, SO_2 results might be biased, as the model tends to highly overestimate SO_2 concentrations (see Section 2.5.2). If this overestimation is due to an overestimation of background emissions as suggested, the contribution of ships to SO_2 may be underestimated. NO_2 contribution is relatively high compared to the other pollutants, reaching 17.86% over Marseille. The comparison with WHO guidelines for NO_2 shows that ships account for 50% of the 10 $\mu g \cdot m^{-3}$ annual limit over Marseille. In

Fos, where the Max value occurred, the ship-related concentration averaged over June surpasses the daily limit of 25 $\mu g \cdot m^{-3}$. The spatial distribution of the contribution

Table 2.12: Contribution of ships to pollutant concentrations from 2021-05-31 to 2021-07-01.
Max and %_Max are respectively the maximum value and contribution of the time-averaged ship-emitted concentrations.
D03 and %_D03 are respectively the spatio-temporal ship-emitted concentration and contribution within the DOI.
Mrs and %_Mrs are respectively the spatio-temporal ship-emitted concentration and contribution within the sub-domain of Marseille, which is defined in 2.1.

	$PM_{2.5}$ $\mu g \cdot m^{-3}$	PM_{10} $\mu g \cdot m^{-3}$	SO_2 $\mu g \cdot m^{-3}$	O_3 $\mu g \cdot m^{-3}$	NO_2 $\mu g \cdot m^{-3}$
Max	1.14	1.85	1.34	8.17	28.25
D03	0.58	0.7	-0.01	1.20	2.74
Mrs	0.71	0.94	0.1	-0.69	5.57
%_Max	9.6%	8.1%	11.6%	12.1%	47.1%
%_D03	5.12%	2.95%	-0.08%	2.12%	10.02%
%_Mrs	6.49%	4.00%	1.94%	-1.21%	17.86%

reveals distinct patterns depending on the pollutant. The contribution of SO_2 is relatively low overall, with higher values observed in Marseille, particularly near the port area, as shown in Fig.2.19. In contrast, the contribution of NO_2 is higher, particularly near ship sources at docking sites in both basins of the port. For both species, the ship contributions progressively decrease as the distance from the shore increases. However, the O_3 contribution exhibits the opposite trend, with higher values further inland. The negative regions in Fig.2.20 correspond to elevated NO_2 contributions. This is primarily due to the complex chemical process involved in ozone formation in the atmosphere. The emission of nitrogen oxide (NO) by ships enhances a specific chemical reaction, resulting in a decrease in ozone concentration. The reaction can be represented as:

$$NO + O_3 \rightarrow NO_2 + O_2 \tag{2.15}$$

$PM_{2.5}$ and PM_{10} contributions, displayed on Fig.2.21, show stronger contribution inland. Note that the color scale only spans from 0 to 10% instead of 0 to 50% in previous figures. A comprehensive analysis of this phenomenon requires to understand how the different particle size is handled by the model.

MOSAIC model categorizes particle size into 4 bins: bin_1 (0.039–0.156 μm), bin_2 (0.156–0.625 μm), bin_3 (0.625–2.5 μm), and bin_4 (2.5–10 μm). Fig.2.22 illustrates how

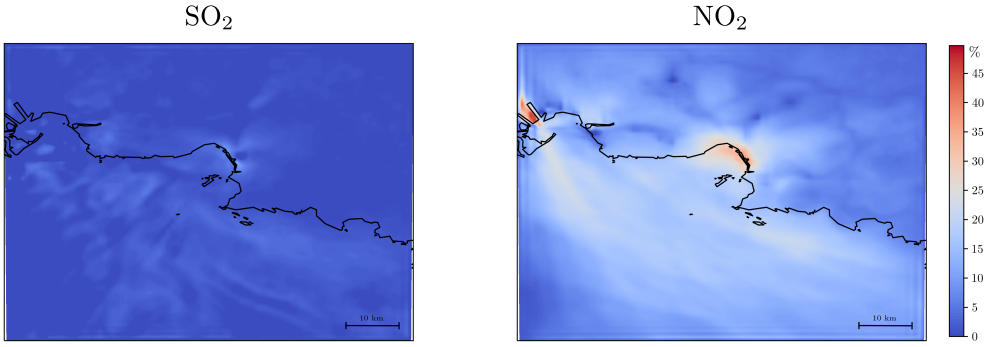

Figure 2.19: Ship contribution to SO_2 and NO_2 from 2021-05-31 to 2021-07-01.

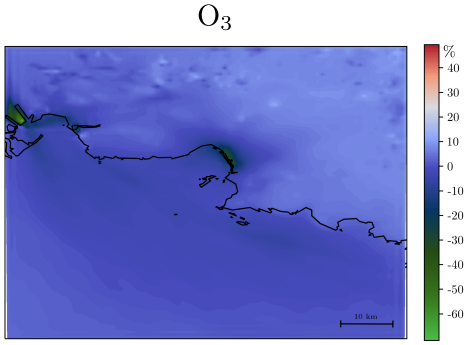

Figure 2.20: Ship contribution to O_3 from 2021-05-31 to 2021-07-01.

ship-emitted particles of various bin sizes disperse over D03. Note that to calculate the total mass and number of aerosol species from WRF-Chem outputs, water was excluded from the sum, as only dry particles are considered. Bin_4 exhibits a similar pattern to PM_{10}, with higher contribution in the vicinity of the two basins. This is consistent with the fact that PM_{10} is composed of the sum of all bins, with a major contribution from the large sizes in these regions. The heavier particles seem to be less advected during the time of their fall from 2nd and 3rd levels, where they were emitted. On the contrary bin_2, and to a lesser extent bin_3, contribute to more inland regions, particularly in the northeast of D03, which indicates a stronger influence of advection on smaller particles. Even though mass contribution of bin_1 ship-emitted particles is negligible, contribution is very large in terms of number, as it is the dominant bin in the particle size distribution

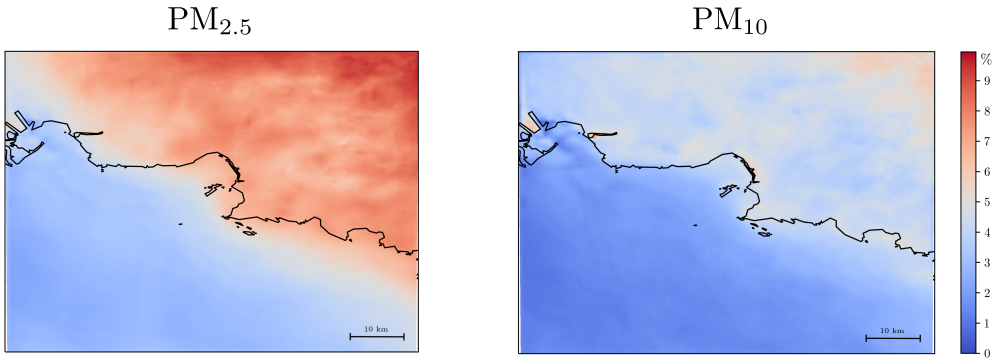

Figure 2.21: Ship contribution to $PM_{2.5}$ and PM_{10} from 2021-05-31 to 2021-07-01.

of boat emission (Karl et al., 2023). In addition, such particles are found up to 1000 m above ground. Therefore, they are likely to play a significant role in SOA formation. The prevailing southwest wind direction in Marseille during this period (see Fig.2.18) aligns with the dispersion of $PM_{2.5}$ in the northeast region. The Stokes number describes the ability of particles to adopt the fluid velocity. It is calculated as the ratio of the particle stop distance to the characteristic length of the flow, following the formulation of Binkowski and Shankar (1995):

$$St = \frac{u_*^2 V_g}{g\nu} \tag{2.16}$$

where u_* is the friction velocity defined in Eq.2.4, ν the kinematic viscosity of air, and V_g is the gravitational settling velocity, defined as:

$$V_g = \frac{\rho_p d_p^2 g C_c}{18\mu} \tag{2.17}$$

where ρ_p and d_p are respectively the density and the diameter of the particle, and $C_c = 1.246$ is the Cunningham correction factor. Although each bin contains different densities for the various aerosol species represented in the model, a global density of 2.165 g.cm^{-3} is used for simplicity. The median diameter is used for each bin. u_* is a diagnostic variable from WRF-Chem outputs. The Stokes number was calculated for each bin, as represented in Fig.2.23. As expected, it is higher over land than over the sea, primarily due to variations in friction velocity. Smaller particles in bin_1 have a Stokes number on the order of 10^{-3}, indicating a regime where gravitational forces are negligible and particle movement is dominated entirely by advection within the flow and Brownian motion. For

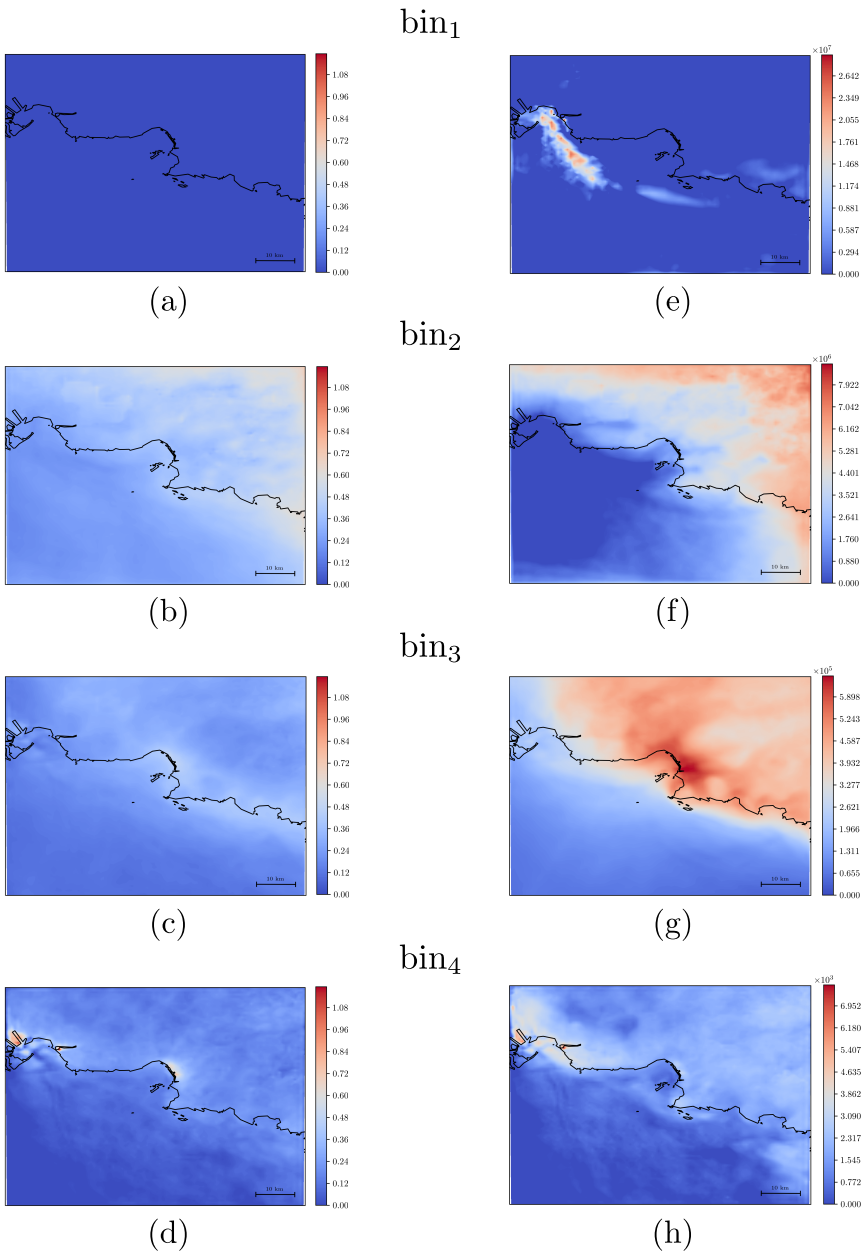

Figure 2.22: Ship-emitted particles concentration (left column) and number (right column) for the 4 bins from 2021-05-31 to 2021-07-01. (a), (b), (c) and (d) share the same color scale in $\mu g.m^{-3}$. (e), (f), (g) and (h) have their own color scale in $\mathrm{kg}^{-1}_{\text{dry air}}$.

bin_2 and bin_3, the Stokes numbers are on the order of 10^{-2} and 10^{-1}, respectively. In this regime, the influence of gravitational settling on the particles' trajectories cannot be ignored and coexists with that of the flow. Coarser particles in bin_4 have a Stokes number on the order of 10^1, where gravitational forces predominate, though the behavior is still not purely "ballistic".

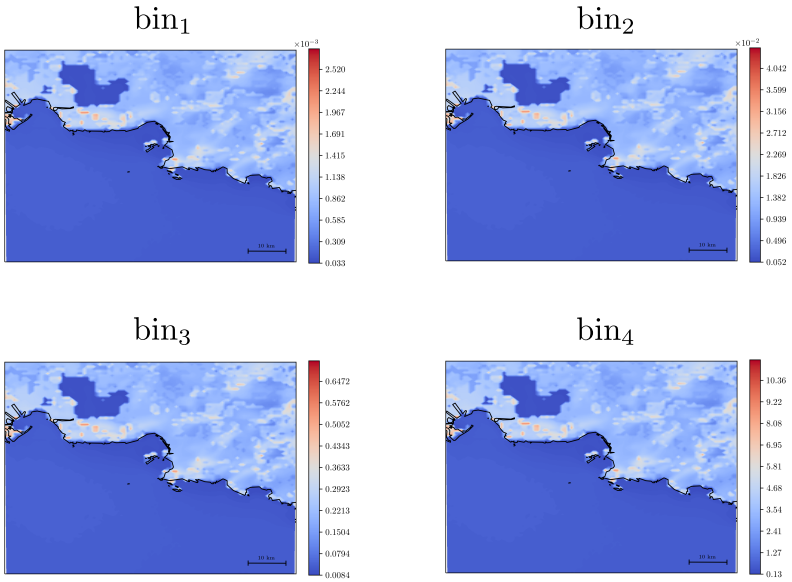

Figure 2.23: Stokes number for the 4 size bins over D03.

2.6.3 | Conclusion on real case study

Atmospheric pollution in Marseille was simulated for the entire month of June 2021 using two versions of the WRF-Chem model—MOZCART and MOZAIC—and these simulations were compared against actual measurements. Ship emissions were estimated using real traffic data from GPMM authorities, while background emissions were sourced from the CIGALE inventory. Both datasets were processed to provide hourly inputs for the WRF-Chem models. The impact of ship emissions on overall pollution was assessed using the zero-out method, revealing that ships contributed an average of 6.49% to $PM_{2.5}$ levels in Marseille. Although contributions to SO_2 might be underestimated, ships were found to be a significant source of NO_2, contributing an average of 17.86% to NO_2 levels in the city. The spatial distribution of ship-emitted particles shows that near the port areas, coarse particles represent the primary mass contribution. In contrast, finer particles tend

to be transported towards the northern regions of the simulation domain, as gravitational settling becomes secondary to the influence of the wind flow on the particles.

2.7 | Influence of Meteorological Conditions

The study presented in this section was subject to a publication in the Journal of Atmospheric Environment (Chevet et al., 2024). The aim is to quantify the contribution of ship traffic to air pollution in Marseille and its vicinity under 8 typical seasonal meteorological conditions, utilizing the fully coupled chemistry model WRF-Chem, considering a single high ship emission scenario. This should allow us to better understand the contribution of ships in the worst-case scenarios, which cannot be captured when using annual emission inventories. The main primary and secondary pollutants are analyzed, including SO_2, NO_2, $PM_{2.5}$, PM_{10} and O_3. Special emphasis is given to $PM_{2.5}$, acknowledged as the main pollutant in Europe, resulting in 238000 premature deaths in 2020, as reported by the EEA (European Environment Agency, 2022).

2.7.1 | Model Set-Up

Model set-up presents some slight differences from the MOZAIC setup presented in Section 2.2. The grid only counts 50 vertical levels (instead of 51 before). The 2nd level is then at 50 m. All levels above are common in both setups. The last difference lies in the absence of nudging and chemical initial and boundary conditions for this specific study.

2.7.2 | Ship Traffic and Emission Scenario

The traffic scenario includes 45 boats per day with a GT ranging from 2158 to 235 600. 28 incoming and 25 outgoing calls per day are modeled. This scenario, defined from observations using AIS data, represents a high occupancy of the port. Note that even if the number of calls is low compared to real data (see Fig.2.6), traffic is considered high in terms of GT, as only large boats ($GT > 10000$) are represented in this scenario. Details of the average gross tonnage per ship category are shown in Table 2.13.

Tables used for emission calculation might differ slightly from the ones used by AtmoSud. As shown in Eq.2.1, the total emission from a boat is the sum of the emissions from the Main (Propulsion), Auxiliary and Boiler engines. For propulsion engines, the nominal power is derived from EMEP Table 3-12 functions, using the Mediterranean Sea fleet (2006). Table 2-4 from EPA gives auxiliary to propulsion ratios for the nominal

Table 2.13: Ship traffic scenario

	Passenger	Cargo	Tanker	Container	Bulk	Tot.
No. of boats in WEST basin	0	1	16	6	1	24
No. of boats in EAST basin	15[1]	4	0	2	0	21
Average GT	65917	23272	11160	71038	31000	41844
Average Hotelling time	10h30	19h00	15h30	10h00	10h45	13h08

[1] 6 Cruises and 9 Ferries

power.
The load factor depends on the engine type but also on the phase. LFs for Main and Auxiliary engines come from EMEP Table 3-15 and EPA Table 2-7 respectively. For boilers, Table 2-17 from EPA provides boilers power for each phase, which already includes the LF.
Emission Factors for Main Engines are taken from EPA Table 2-9 combined with EMEP Table 3-7 to take into account the different Propulsion engine sub-types (SSD, MSD, GT, ST). For Auxiliary Engines, Table 2-16 from EPA provides EFs according to the fuel type. The EPA methodology assumes that Auxiliary Engines are using 29% of MGO and 71% of RO. However, regarding the regulations imposed by IMO in 2020, we assume that Auxiliary Engines use 100% of MGO with a sulfur mass content of 0.5% or less. EFs for boilers are taken from the ST (Steam Turbine) row of EPA Table 2-9. Emissions are then converted to $\mu g \cdot m^{-2} \cdot s^{-1}$ for particles and $mol \cdot km^{-2} \cdot h^{-1}$ for gas-phase species. Cruising speeds depend on the ship type. They were taken from EPA Table 2-6.

The height of ship emissions is defined at the second vertical level, i.e. 50 m above the ground. Due to the lack of information on maneuvering time and road layout, only the hotelling and cruising phases are taken into account for a matter of simplicity. Fig.2.24 depicts the emission field of $PM_{2.5}$ based on the ship traffic scenario introduced here. In the two basins, the isolated red dots represent emissions from boats during the hotelling phase. Additionally, the trails illustrate emissions from cruising boats.

2.7.3 | Time-slot Selection

In order to capture the behavior of boat pollution under various meteorological conditions, two 72-hour statistically significant time slots per season were considered. The selection of these time slots was based on a statistical study of observations from the meteorological station of Marignane for the full year 2021 with a 3-hour time interval.

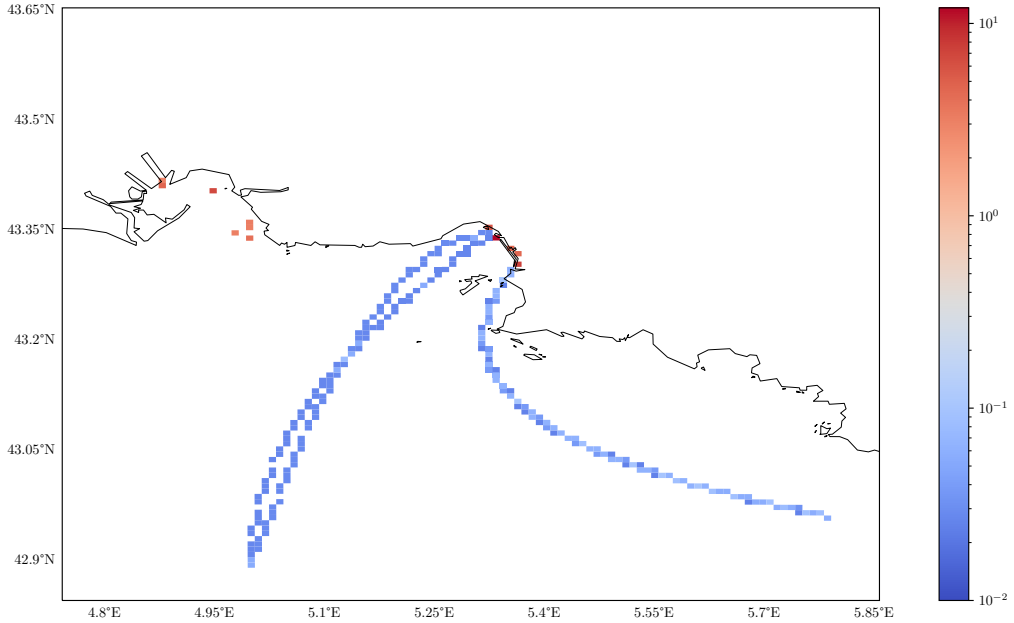

Figure 2.24: 24-hour averaged $PM_{2.5}$ estimated ship emissions from ship traffic scenario in $\mu g \cdot m^{-2} \cdot s^{-1}$ (log scale).

Observations are first filtered and split in two categories of wind speed, defined from the Beaufort scale (bf). The wind is considered weak between 0.3 and 3.4 $m \cdot s^{-1}$ (1-2 bf) and moderate from 3.4 to 7.9 $m \cdot s^{-1}$ (3-4 bf). For each season, one time slot with weak wind and one time slot with moderate wind are selected. Seasonal averages of the 2m temperature and nebulosity were calculated and used as criteria for selecting the time slots. Wind has a strong advective effect on the chemical species present in the atmosphere. The location of the areas most impacted by pollution is therefore strongly dependent on wind direction. As discussed in Ma et al. (2022), a sea breeze can significantly enhance the transport of pollutants from shore to land.

Fig.2.25 shows the annual distribution of wind direction measured at the station of Marignane during 2021. Three main directions have been identified: North-West, West, and South-East, which align with the local winds known as the Mistral, the Tramontane, and the Marin, respectively. The selection, presented in Table 2.14, was made so that each one of these three winds are dominant during at least one of the time slots.

For each time slot, two simulations, with and without ship emissions considered, are

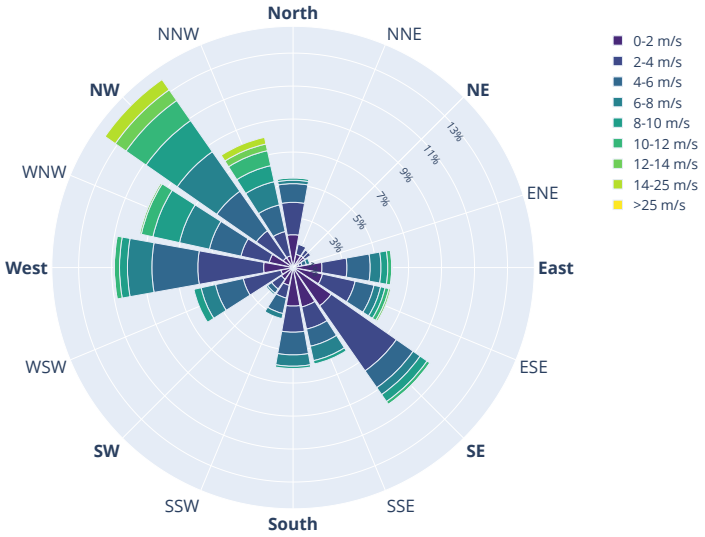

Figure 2.25: Windrose of 2021 from Marignane meteo station

Table 2.14: Weather conditions for each time slot over 2021

	Start/end dates	Wind speed $m \cdot s^{-1}$	Wind direction	Temperature K	Nebulosity $Octas$
Winter	03 – 02/03 – 05 (**W1**)	1.48	SE	282.5	4.04
	01 – 23/01 – 26 (**W2**)	5.68	NW	281.4	3.42
Spring	04 – 22/04 – 25 (**Sp1**)	2.71	S	288.1	1.16
	05 – 03/05 – 06 (**Sp2**)	4.35	W	288.8	1.21
Summer	07 – 19/07 – 22 (**Su1**)	2.68	NW	300.0	0.91
	08 – 28/08 – 31 (**Su2**)	5.40	NW	295.0	0.83
Fall	11 – 19/11 – 22 (**F1**)	1.19	SE	283.5	1.79
	11 – 06/11 – 09 (**F2**)	3.98	N	283.8	2.00

performed with the same conditions otherwise, in order to apply the "zero-out" method explained in Section 2.6.2.

2.7.4 | Results

Fig.2.26 and Fig.2.27 show time-averaged contribution of boats to $PM_{2.5}$ for time slots corresponding to weak and moderate wind speeds respectively.

Comparing the two figures, one can notice that under weak wind conditions, the contribution of boats is more spread out and lower than for moderate wind conditions. High

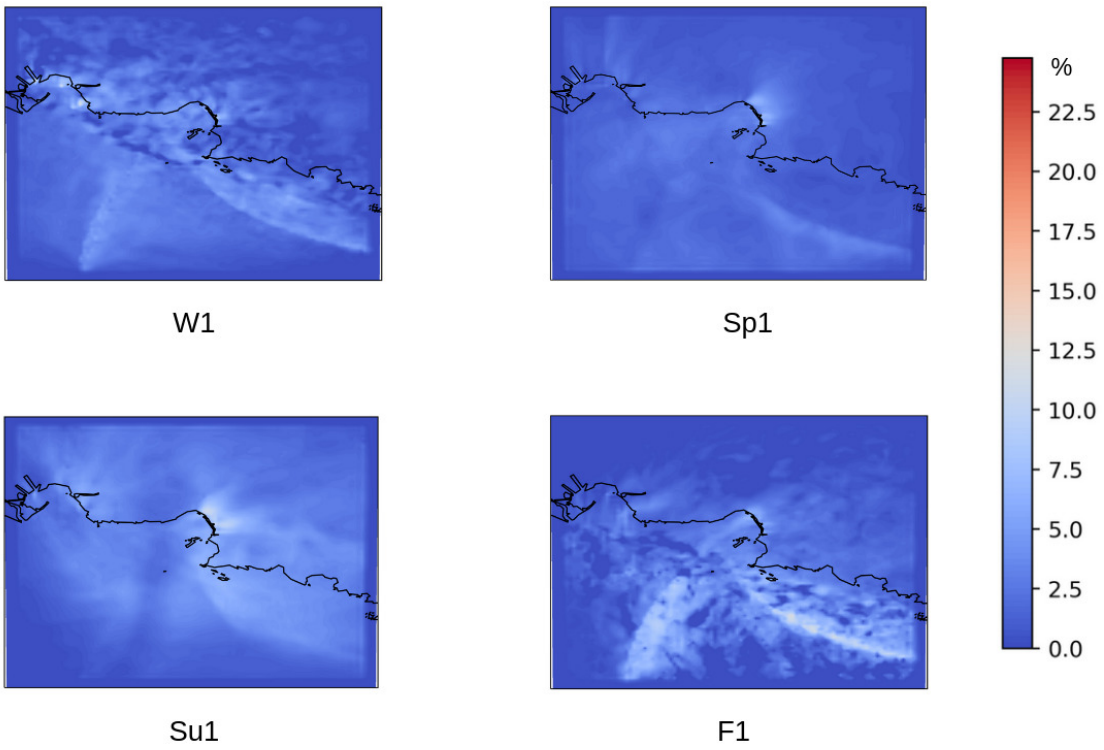

Figure 2.26: 72-hour averaged contribution of ships to $PM_{2.5}$ concentration in %
for weak winds.

contribution values are observable as red spots on Fig.2.27. The contribution of boats in these areas ranges between 20 and 25%. The location of these hot spots corresponds to the positions of the docked vessels. Plumes oriented in the direction of the prevailing winds are clearly observed, especially for time slots **Su2** and **F2** where the wind direction was less fluctuating. Contributions to NO_2 and SO_2 are provided in Appendix C.

Table 2.15 presents the contribution for each pollutant. %_Max is the spatial maximum observed over the time-averaged contribution field whereas %_Avg_Mrs is the mean contribution over the rectangular sub-domain focused on Marseille represented in Fig.2.1. Calculations are done on the first simulated vertical level only. Ships appear to have a larger impact on gas-phase species concentration than on aerosol concentration. Boats contribute to approximately 1.48 to 7.6% of the particulate matter concentration, while their contribution to SO_2 and NO_2 concentrations is higher, accounting for over 30%. $PM_{2.5}$ and PM_{10} ship contribution levels are consistent with the results summarized in Viana et al. (2014) (1-14% and 1-7% respectively). However, NO_2 levels are

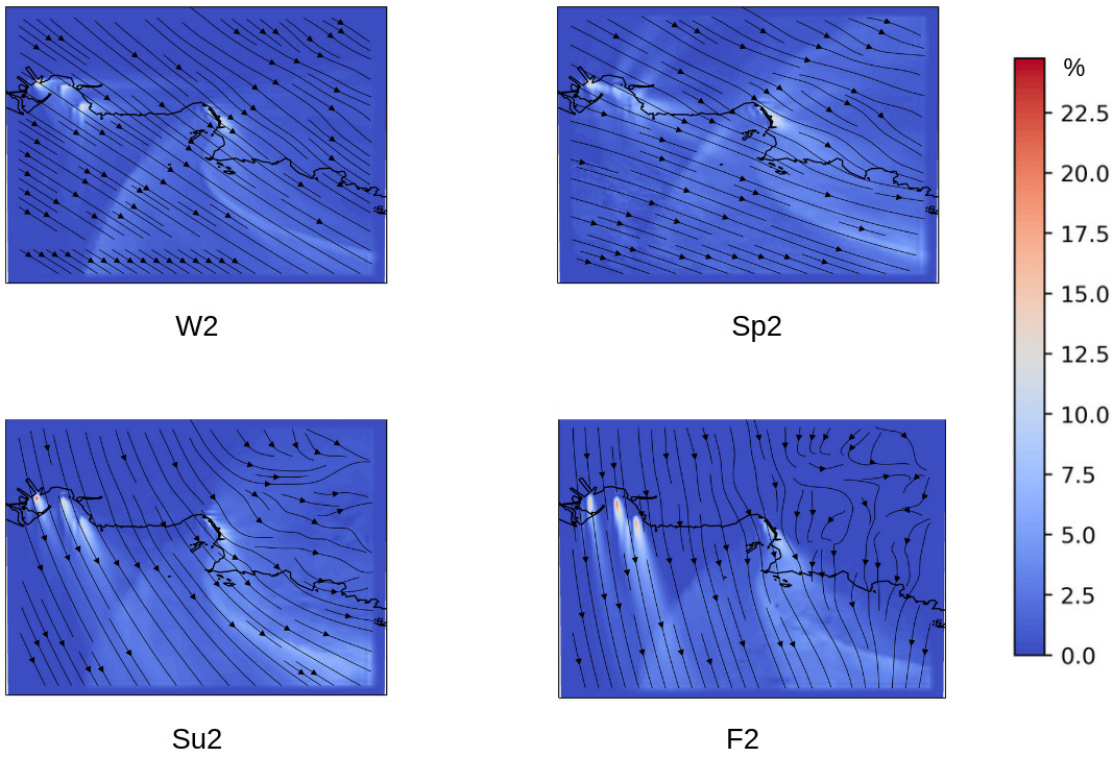

Figure 2.27: 72-hour averaged contribution of ships to $PM_{2.5}$ concentration in %
for moderate winds.

higher (∼70%) than those presented in the study (7-24%). "The Scipper Project," a recent campaign aimed at evaluating ship pollution, conducted assessments in various port cities, including Marseille. The findings indicate that boats in Marseille contribute an average of 31% to NO_2 levels, while in Gothenburg and Kiel, the contributions reach 73% and 75% respectively (Oppo et al., 2023). The elevated levels of NO_2 contribution observed from the WRF-Chem simulations can be attributed to the specific conditions considered in this study, which include a high occupancy scenario at the port and weak or moderate wind conditions. $PM_{2.5}$ and SO_2 contribution levels are in line with the Scipper Project values for Marseille (i.e. 3% and 34% respectively). For all time slots and for every species, we note that the mean contribution of boats is higher over Marseille than over the whole domain. For instance, as it is for PM_{10}, pollution from boats stands for 0.47 $\mu g \cdot m^{-3}$ on the domain against 1.56 $\mu g \cdot m^{-3}$ over Marseille on average for time slot **Su1**. It is observed that in most cases, the contribution of ships to ozone is negative. The reason for this is the same as explained in Section 2.6.2, with the primary

Table 2.15: Contribution of ships to pollutants concentration

Time slots		$PM_{2.5}$	PM_{10}	SO_2	O_3	NO_2
W1	%_Max	13.7%	15.3%	69.6%	11.1%	70.0%
	%_Avg_Mrs	2.25%	2.91%	39.7%	-9.86%	68.0%
W2	%_Max	16.2%	15.7%	88.0%	9.79%	93.7%
	%_Avg_Mrs	2.34%	2.74%	34.2%	4.14%	68.9%
Sp1	%_Max	9.47%	12.1%	57.4%	3.58%	57.9%
	%_Avg_Mrs	3.21%	4.41%	43.2%	-19.7%	64.7%
Sp2	%_Max	17.1%	16.3%	71.9%	12.0%	81.8%
	%_Avg_Mrs	3.13%	3.68%	33.1%	-5.83%	75.0%
Su1	%_Max	11.2%	14.1%	60.8%	3.31%	76.3%
	%_Avg_Mrs	5.69%	7.60%	39.3%	-31.5%	79.1%
Su2	%_Max	23.8%	24.9%	80.5%	1.13%	79.6%
	%_Avg_Mrs	2.33%	2.88%	33.9%	-11.6%	49.9%
F1	%_Max	12.4%	14.6%	54.6%	3.26%	57.0%
	%_Avg_Mrs	3.71%	5.23%	39.4%	-55.2%	50.7%
F2	%_Max	20.4%	25.5%	61.7%	5.88%	87.4%
	%_Avg_Mrs	1.48%	2.06%	25.5%	-14.5%	36.4%

cause being the photo-chemical reaction described in Eq.2.15.

The results discussed here are obtained without considering the maneuvering phase, as indicated earlier. However, emission factors associated with this phase are typically greater than those for the hotelling phase. This suggests that the values presented here might underestimate the actual contributions and concentrations. Considering all phases would provide a more comprehensive representation of maritime pollution.

Only the relative contribution of ships to total pollution has been discussed so far. These graphs do not provide information on the meteorological conditions that lead to the highest concentration values. Fig.2.28 aims to compare the time slots in terms of concentration of ship-emitted $PM_{2.5}$. It was obtained by substracting the time-averaged concentration values of **W2** (corresponding to the lowest $PM_{2.5}$ values) from all time slots. As expected, weak winds (first row) lead to higher differences than moderate winds (second row).

Highest values of concentration in $PM_{2.5}$ are mostly located over Marseille. For the same traffic scenario, the time-averaged $PM_{2.5}$ concentration can vary by 3 $\mu g \cdot m^{-3}$ locally from one meteorological condition to another. In every case, the difference is stronger above Marseille. For slot **F1**, the average difference is 0.22 $\mu g \cdot m^{-3}$ for the entire domain and reaches 2.55 $\mu g \cdot m^{-3}$ over the city.

Total pollution was analyzed along shorelines defined at a certain distance from the

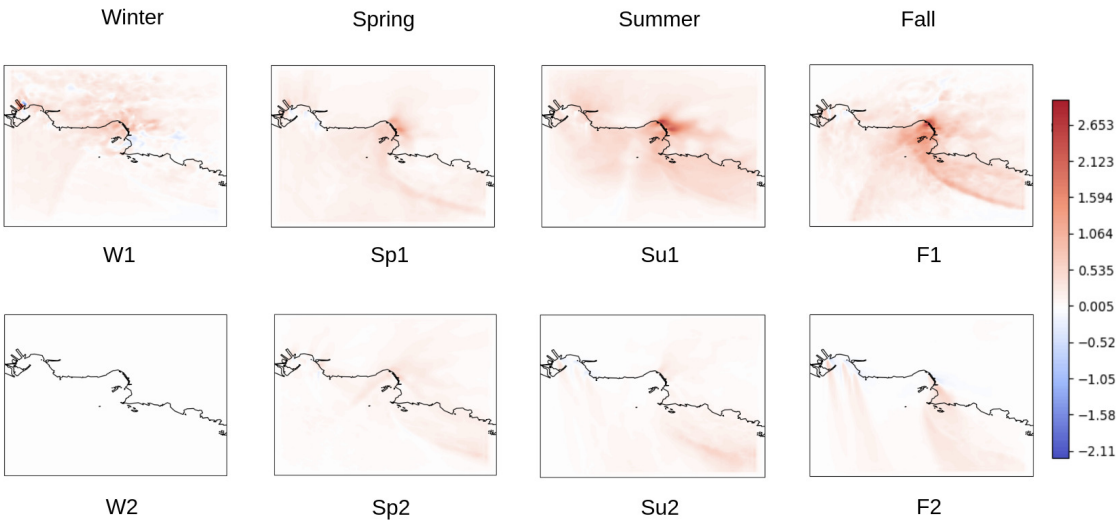

Figure 2.28: Differences between all time slots and **W2** in ship-emitted $PM_{2.5}$ concentration in $\mu g \cdot m^{-3}$. First and second rows correspond to weak and moderate winds respectively.

coast, as represented in the upper half of Fig.2.29.

Table 2.16: $PM_{2.5}$ Concentration levels along distances from shore for time-slot **Su1**.
Tot. is for total domain. Marseille and Fos are splitting shorelines in longitude according to the borders represented in Fig.1.7. Outside refers to the remaining longitudes of the domain.

	Marseille	Fos	Outside	Tot.
1 km	14.06	22.15	10.84	13.91
3 km	12.50	14.41	9.15	11.04
5 km	11.31	11.28	8.22	9.61
10 km	9.95	8.78	8.69	9.04
20 km	8.19	6.52	5.26	6.19

Given the higher contribution from ships for **Su1**, our focus for this analysis will be on the corresponding time slot. Outside of Marseille and Fos, $PM_{2.5}$ concentration does not seem to depend on the distance from the coast. However, for longitudes crossing Marseille, the shorter the distance, the higher the concentration. Indeed, Table 2.16 emphasizes a stronger gradient in Marseille than outside its borders. This was observed for all time slots, meaning that the meteorological conditions do not explain this correlation

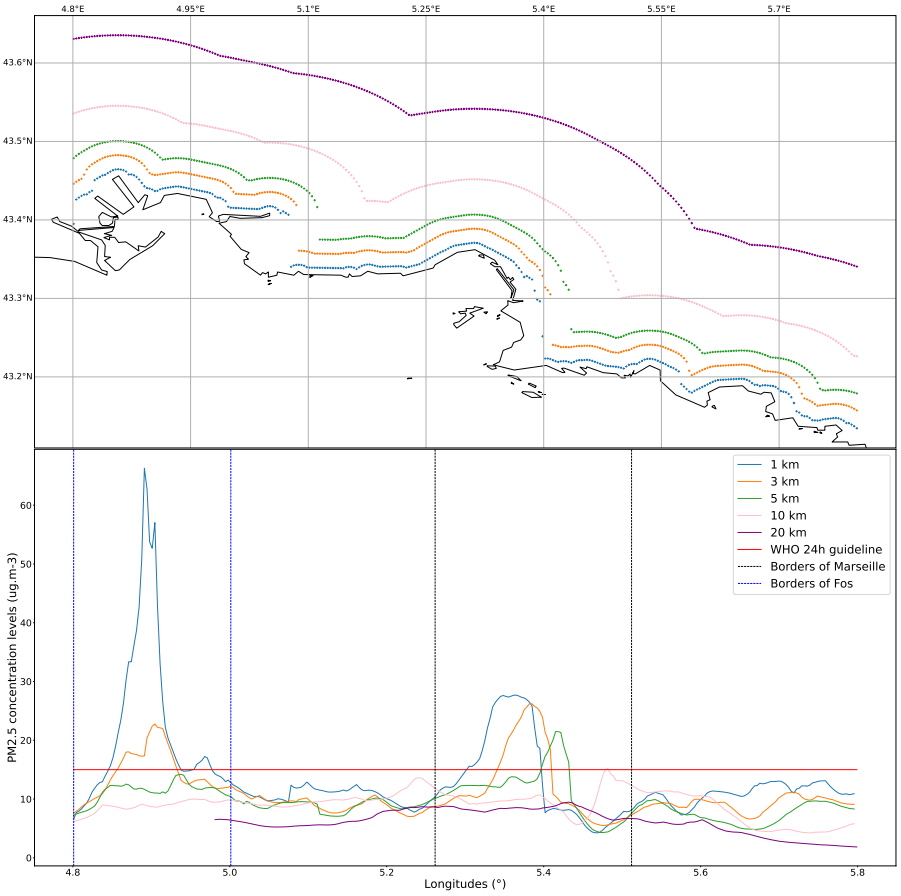

Figure 2.29: Top: map of the 5 shorelines; Bottom: 72-hour averaged $PM_{2.5}$ concentration levels along the shorelines for time-slot **Su1**.
Borders of Fos: 4.80°E to 5.00°E ; Borders of Marseille: 5.26°E to 5.51°E

between distance from the shore and concentration. Since boats are the main source of pollution on the coast of Marseille, it seems legitimate to believe that the main cause of this concentration gradient is the presence of hotelling boats. The pollution in the Fos area is mainly caused by the presence of chemical plants near the shore. The influence of ships in this area is therefore difficult to identify from this figure. We note that the WHO guideline for $PM_{2.5}$ defined as 15 $\mu g \cdot m^{-3}$ 24-hour average is far exceeded in Fos-sur-Mer

and Marseille close to the shoreline, where concentration reaches 66.3 $\mu g \cdot m^{-3}$ and 27.7 $\mu g \cdot m^{-3}$ respectively.
When the wind is weak, the sea-land breeze dominates, establishing an onshore wind during the morning that persists throughout the day, reffered to as the diurnal breeze. In the evening, an inversion occurs, leading to the observation of a nocturnal offshore wind. These inversions occur when the land warms up (in the morning) or cools down (in the evening). Moreover, the sea breeze is expected to be more intense in spring and summer when temperatures reach higher values during the day, as opposed to the land breeze, which are more intense in Winter and Fall (Federico et al., 2010). To analyse this phenomenon, we introduce the breeze vector, defined as a unit vector oriented inland along the normal to the Marseilles coastline. We then calculate the scalar product between the breeze vector and each of the normalized wind vectors over the Marseilles sub-domain. Thus, the scalar product equals 1 when the wind is onshore and perfectly aligned with the breeze vector. To better understand the effects of this phenomenon on maritime pollution, Fig.2.30 shows the time series of the average scalar product superimposed on the average $PM_{2.5}$ concentration attributed to ship emissions over Marseille for each time slot of weak wind. Generally, an onshore wind is observed during the day and offshore at night. Morning inversions take place between 8 and 11 AM, while evening inversions occur between 6 and 9 PM. As the time slots are chosen according to each season, it is coherent to note that the inversion takes place earlier in the morning and later in the evening during summer, effectively prolonging the onshore period.

Consequently, maritime pollution is carried towards the city for a longer period compared to other seasons. Additionally, it is observed that for Su1, a south wind neutralizes the offshore regime, and no predominant wind direction is observed at night. This seems to be even more conducive to the accumulation of $PM_{2.5}$ above the city, as it is no longer swept by the land breeze. For these two reasons, the configuration appears highly unfavorable in terms of maritime pollution. Indeed, Su1 is the period during which the maritime contribution in $PM_{2.5}$, PM_{10}, and NO_2 is the strongest in Marseille (Table 2.15). The fact that the maritime emission scenario remains the same for each simulated day reinforces our interpretation, as inter-seasonal variations cannot be explained by emission variations. The inversion of the wind regime appears to have a significant impact on the concentration of $PM_{2.5}$. In most cases, a concentration peak is observed during wind regime transition phases. This phenomenon seems clearer for Sp1, which can be explained by a greater amplitude of wind direction oscillation.

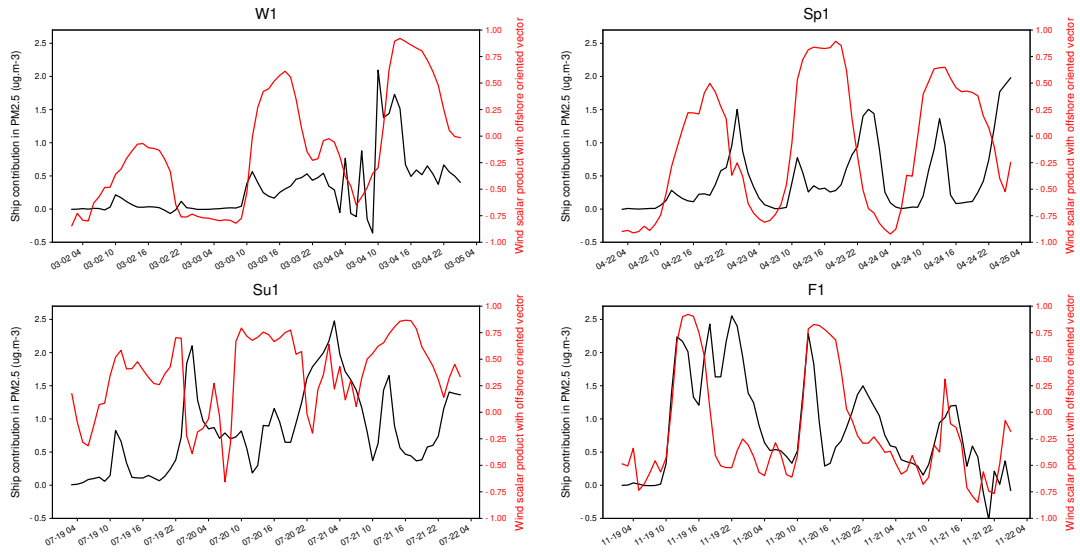

Figure 2.30: In black: Time series of $PM_{2.5}$ ship-emitted average concentration over Marseille; In red: Time series of average scalar product between the sea-breeze vector and wind direction over Marseille

2.7.5 | Conclusion on the influence of Meteorological Conditions

Air pollution from ships is still a major problem in Mediterranean coastal cities, since maritime traffic is constantly increasing and restrictions on the fuel used in this area are not sufficient. While studies on the impact of ships are becoming more and more present worldwide, few have been conducted in the city of Marseille. In this work, we developed a city-scale model using WRF-Chem that allows us to estimate the contribution of boats to air pollution for the major pollutants. The contribution of boats was estimated at over 5% and 7% for $PM_{2.5}$ and PM_{10} respectively in the worst-case scenarios. For SO_2 and NO_2, contributions reach higher values (40% and 80%). The O_3 contribution is negative in most cases. The analysis of typical seasonal meteorological conditions for the same ship traffic scenario enabled us to examine the impact of wind on ship pollution. Comparing different time slots revealed that the dispersion of ship-induced pollution is significantly influenced by wind speed, displaying a more pronounced effect during periods of low wind speed. Additionally, there appears to be a significant influence of the sea breeze, as indicated by the occurrence of peaks in maritime pollution during sea-land breeze inversions.

As the study is not based on a real case scenario, but on the same shipping traffic scenario under different weather conditions, the contribution values for ships are not intended to

be representative of a situation that might have occurred in the past.

2.8 | Discussions

WRF-Chem stands out as a versatile CTM, offering considerable flexibility in configuring physical, chemical, and aerosol processes. Its online coupling of meteorological and pollution processes enables a more in-depth representation of the interactions between gas-phase and aerosol species and their environment. The open-source nature of WRF-Chem facilitates continuous enhancements and fosters collaboration among a broad research community, providing it with a distinct edge over proprietary CTMs. The pre-processing of GPMM ship traffic and the CIGALE emission inventory, developed during this PhD project, is compatible with all chemical and aerosol schemes available in WRF-Chem due to their shared input format for emissions. While this manuscript focuses on the MOZCART and MOZAIC setups, exploring additional schemes could offer valuable insights into the observed overestimation of SO_2 during the validation process. Such an investigation might help verify whether this overestimation is due to a positively biased emission inventory from CIGALE or other factors. The significant improvement observed between the bulk (GOCART) and sectional (MOSAIC) approaches underscores the importance of considering particle size and number, rather than mass only. Although MOSAIC was used with 4 bin sizes in this study, an 8-bin version is also available and might improve accuracy of the predictions of fine particles. Other schemes such as MADE/SORGAM (Ackermann et al., 1998; Schell et al., 2001) use a modal approach, by considering three log-normally distributed modes (nuclei, accumulation and coarse mode).
In the pre-processing method implemented in this study, particulate matter emissions are introduced into the model only as $PM_{2.5}$ and PM_{10}. MOSAIC then redistributes these emissions across different size bins according to a predefined distribution. However, if information on the size distribution of ship-emitted particles is known, this predefined distribution could be bypassed. Jonsson et al. (2011) showed measured size distribution per number for individual ships, which could be integrated in the model.
Few studies involving atmospheric pollution modeling include validation for UFPs due to the scarcity of measurements in this size range. However, the Lonchamp station in Marseille provides highly refined measurements, which makes UFP validation feasible and is included in the perspectives of the study.
Another perspective is to evaluate reduced emission scenarios. The municipality of Marseille is currently carrying a project to assess the health benefits of various restrictions or

changes at the city scale. The EQIS-PA framework (Évaluation Quantitative d'Impact sur la Santé de la Pollution de l'Air) incorporates scenarios for reducing ship emissions, facilitating more informed decision-making. This allows for the evaluation of various emission reduction strategies, such as the number of installed CENAQ systems, fuel types, or cruise ship quotas. The potential health benefits of these hypothetical emission reductions can then be estimated by comparing them to the real case study.

3

Modeling of Atmospheric Pollution with Machine Learning

3.1 | Introduction

3.1.1 | From Numeric to Machine Learning

In air quality forecasting, a distinction is made between research models, used to conduct studies for a comprehensive understanding of mechanisms, and operational models, utilized daily in air quality monitoring as part of a preventive approach. Typically, operational forecasts are updated every six hours and rarely extend beyond 72 hours. Although specific criteria may not be precisely defined, we will consider a model as operational if it can generate a 72-hour forecast within a computation time of less than six hours.
The WRF-Chem model is capable of generating 72-hour predictions of aerosol concentration fields and gaseous species. However, in the initial project phase, the configurations required about 24 hours of computation over 72 hours using 64 cores on two nodes.
The aim of the second part of this PhD project is to develop a machine learning model capable of predicting 2D concentration fields with accuracy comparable to that of the WRF-Chem model, while achieving computational times suitable for operational use.

3.1.2 | Machine Learning Forecasting

Machine Learning (ML) encompasses techniques enabling computers to learn from data and make predictions or decisions. In the context of this work, ML was leveraged for its forecasting capabilities. ML techniques have become increasingly popular for time series prediction due to their ability to model complex patterns and relationships in data. In a forecasting context, the model's algorithm must account for the time dependence between observations, as the original dataset usually consists of time series data. The simplest ML predictive models are regression-based models. A common approach is to use auto-

regression, which takes into account the relationship between an observation and a certain number of lagged observations (previous time steps). A p-order auto-regression is given by:

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_t \tag{3.1}$$

where X_t is the current value, c is a constant, ϕ_i are the coefficients and ϵ_t is white noise. The most popular auto-regression model is called ARIMA for Auto-Regressive Integrated Moving Average. Such models are commonly used for air quality prediction (Gourav et al., 2020; Kumar and Jain, 2010; Samal et al., 2019). However, air pollution forecasting usually involves multiple monitoring sites or grid cells where pollution is evaluated. Spatio-temporal relations in these different evaluation sites is crucial for better accuracy, as stated by multiple studies (Abirami and Chitra, 2021; Li et al., 2016). Auto-regression-based models struggle to capture this spatiotemporal dependence.
Another approach is using Artificial Neural Networks (ANN), which are algorithms inspired by biological neural networks. ANNs consist of interconnected units called neurons or nodes. The concept of a simple neuron is illustrated on the left of Fig.3.1, and the associated equation is:

$$Y = \sigma(W_X^\top \cdot X + b) \tag{3.2}$$

X represents the input to the neuron, W_X denotes the weight vector associated with the input X, b is the bias term, and σ represents the activation function, introducing non-linearity to the model. When combined, these neurons form a complex network through which information propagates. Together, these computations create a powerful non-linear function that maps input data to output predictions. The principle of an ANN involves adjusting its weights during each epoch, which is a complete iteration over the ensemble of observations in the training dataset. This adjustment is guided by a gradient descent algorithm, which minimizes a loss function. As a result, the network iteratively updates its weights to reduce prediction errors and improve its accuracy.
To extend ANNs for time-dependent data, recurrent neurons, depicted on the right of Fig.3.1, introduce temporal dependencies through an additional term:

$$Y_t = \sigma(W_X^\top \cdot X_t + W_Y^\top \cdot Y_{t-1} + b) \tag{3.3}$$

The output of current time step is used as input for the next step. In practice, Recurrent Neural Networks (RNNs) are designed to handle more intricate temporal dynamics beyond the basic recurrent neuron illustrated here.
Long-Short Term Memory (LSTM) networks represent a more sophisticated variant in

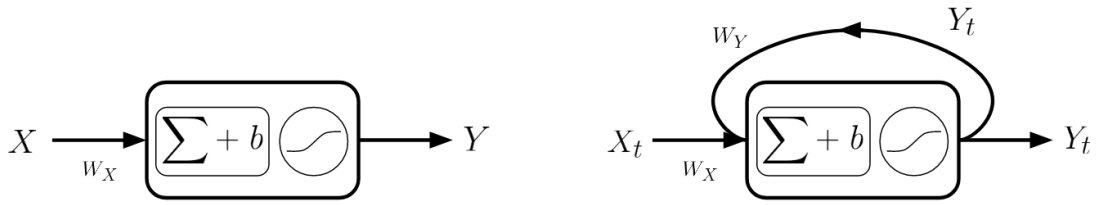

Figure 3.1: Basic concept of a neuron (left) and a recurrent neuron (right)

the field of RNNs. LSTMs are specifically designed to meet the challenge of capturing and learning long-term dependencies in sequential data. They achieve this by incorporating memory cells and gating units, which enable them to retain information over extended time intervals and selectively update or forget information as required. Fig.3.3 illustrates a single LSTM cell and the operations between the various components, governed by the following equations:

$$f_t = \sigma(W_f \cdot [h_{t-1}, X_t] + b_f) \tag{3.4}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, X_t] + b_i) \tag{3.5}$$

$$\tilde{C}_t = tanh(W_C \cdot [h_{t-1}, X_t] + b_C) \tag{3.6}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_{t-1} \tag{3.7}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, X_t] + b_o) \tag{3.8}$$

$$Y_t = h_t = o_t \odot tanh(C_t) \tag{3.9}$$

With σ the sigmoid activation function and tanh the hyperbolic tangent function. W_f, W_i, W_C, W_o are the weight matrices for the forget, input, candidate cell state, and output gates, respectively. b_f, b_i, b_C, b_o are the bias vectors. x_t, h_t, and C_t represent the input, hidden state, and cell state, respectively, at the current time step. The output of a LSTM cell Y_t is a simple copy of the hidden state. As illustrated in Fig.3.2, during the forward pass of information, the output is not necessarily yielded at every step.

In practice, a LSTM cell is composed of multiple units. Each unit corresponds to a single hidden neuron in the LSTM cell. The number of these units determines the size of the vectors used for the hidden state h_t, the cell state C_t and the output Y_t.

In this work, RNNs were used for time series prediction, as they were found more effective than ARIMA in the framework of time series prediction (Siami-Namini et al., 2018). In fact, LSTMs are extensively used for air quality prediction based on time series

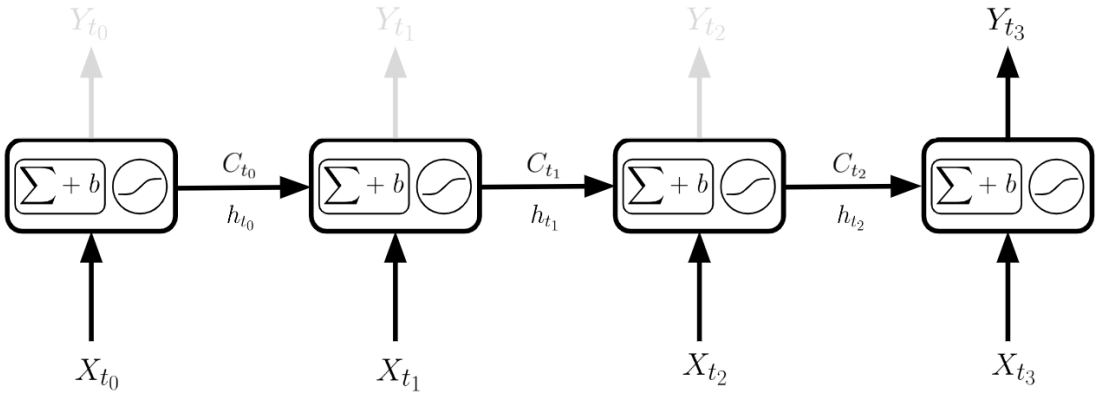

Figure 3.2: Illustration of a forward pass through a LSTM Cell. The input sequence is composed of 4 steps X_{t_i}, while only one output Y_{t_3} is produced, which typically corresponds to the predicted time step in the context of time series prediction.

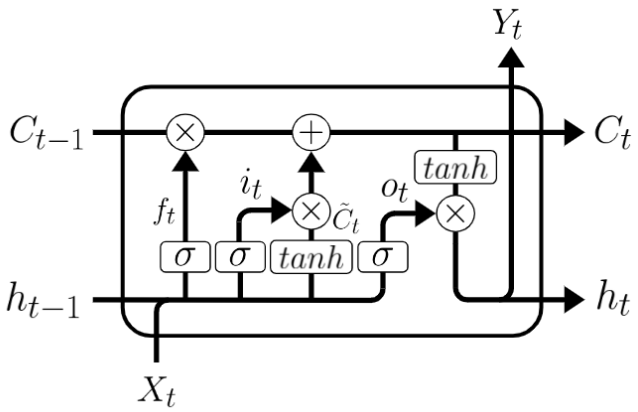

Figure 3.3: Concept of a LSTM Cell.

from measurements (Drewil and Al-Bahadili, 2022; Gilik et al., 2022; Rao et al., 2019; Seng et al., 2021). All of the studies cited above use multivariate LSTM models that includes meteorological data, and some of them use hybrid techniques, such as adding Convolutional Neural Network (CNN) layers before a LSTM, to account for the spatio-temporal relationship between the different measurement locations. The prediction of air quality based on measurements is the most common application of machine learning for this purpose. However, this leads to a scarcity of information across the domain, as measurement networks are typically limited to just a few sensors. Other models add satellite data to produce outputs with wider coverage and greater spatial detail (Tang et al., 2024). Polar-orbiting satellite sensors like MODIS, OMI, and TROPOMI offer

high-resolution global data but with low temporal frequency, typically only once or twice a day. In contrast, geostationary satellites such as GEMS, TEMPO, and the upcoming Sentinel-4 provide hourly information on aerosols and gases at regional scale. The combination of these complementary data types and ground-based measurements could create a powerful dataset for air pollution forecasting, though challenges like scale discrepancies, data gaps, sparse ground stations, and operational difficulties remain (Shin et al., 2020). Another approach involves using CTM outputs as input data for machine learning models. CTMs provide gridded data and can offer fine spatial and temporal resolution, depending on their setup. Geng et al. (2021) used WRF/CMAQ daily $PM_{2.5}$ concentration outputs over China with a spatial resolution of 36 x 36 km, along with ground measurements and satellite data. However, poor accuracy was raised for dust concentration estimation from the CTM, which might impact the overall ML model accuracy. Di et al. (2019) combined PM predictions from GEOS-Chem and CMAQ models, along with other data sources across the United States. The resulting model performs well, although CMAQ was identified as a potential source of inaccuracy in certain years. Fang et al. (2023) coupled the regional feature selection machine learning forecast (RFSML v1.0) to GEOS-Chem over China. However, the computational demands of the CTM simulation were highlighted as one of the main drawbacks of the method. From these examples, two major concerns emerge when using CTMs as input for a machine learning model: the multiple sources of uncertainty within the CTM, such as emission inventories, initial and boundary conditions, meteorological predictions, and physical and chemical parameterizations; and the significant computational cost of generating sufficient data from simulations.

This study employs the CTM-based approach, as the development of a validated model setup for air quality prediction, discussed in the first part of this manuscript, naturally supports this choice.

While the usual goal of such models is to provide accurate predictions by closely matching observed concentrations, the aim of this study differs slightly. First, the input data is restricted to WRF-Chem outputs, including both meteorological and chemical variables. Second, the model's performance is evaluated based on its distance from the WRF-Chem predictions. In our view, this approach offers two main advantages. First, predicting a 2D scalar field from initial conditions and explanatory variables is a problem that also arises in fluid mechanics, particularly in CFD. Thus, this choice helps preserve the generalizability of the developed model to other areas within fluid mechanics. Second, due to computational time constraints, the WRF-Chem setup employed is not the most accurate, as it utilizes the MOZCART model. The use of measurement data for validation

is not considered, given the inherent bias in the basic CTM model. Consequently, this work focuses more on developing a method and architecture tailored to the problem rather than aiming for immediate operational use.

3.1.3 | Reduced Order Modeling

3.1.3.1 | High Dimentionality Problem

As mentioned earlier, the aim of this section is to predict air pollution using only WRF-Chem outputs. The dataset is therefore composed of gridded data, at a resolution of 0.8 x 0.8 km. In a first approach, the domain size is reduced to 64 x 64 elements, which shares the same center as D03 described in Section 2.2. Only the first vertical level is considered, which introduces a limitation by excluding information from vertical processes in the input data and should be acknowledged as a potential source of error. WRF-Chem provides hourly concentration fields, and the simulations cover the year 2021, with the exception of February and March, when the simulations failed to converge due to incompatible initial conditions. In total, the dataset includes 7107 time steps. The ML model must handle 4096 dependent time series over this period, which correspond to each grid cell information. Moreover, additional information, such as meteorological and emission features, are later incorporated in the model. This high-dimensional problem is extremely complex, and the dataset in its raw form cannot be managed by a simple architecture composed solely of LSTMs or any of the model types mentioned earlier.
Reducing the dimension of the data emerges as a crucial step to avoid excessively complex architectures and costly computations. After all, the initial motivation for using machine learning was to minimize computational expense. Reduced-order modeling (ROM) encompasses methods for approximating high-dimensional problems through the use of low-dimensional representations of the original data, while retaining the essential information and dynamics of the system. Throughout this study, two methods were explored: Proper Orthogonal Decomposition (POD) and nonlinear manifold learning using autoencoders (AE).

3.1.3.2 | POD

Proper Orthogonal Decomposition, first introduced by Lumley et al. (1967) is widely used for ROM. When dealing with 2D fields, the snapshot method of POD is particularly effective (Sirovich, 1987). The core idea is to capture all the spatio-temporal information of the field in a single snapshot matrix. This involves flattening the spatial dimensions

into a vector, with each vector representing a snapshot of the field at a given time. POD is then applied to this matrix to identify an orthogonal basis, where the eigenvectors are ranked according to the amount of variability they capture in the data. Note that POD uses the same algorithm and is very similar to Singular Value Decomposition (SVD). The reduction process in POD is achieved by selecting a subset of modes, specifically the first few that capture the most significant information. These modes correspond to the largest eigenvalues and represent the primary patterns or structures in the data. By retaining only these dominant modes, the complexity of the system is significantly reduced while still preserving the essential features and dynamics of the original data. The mathematical description of the snapshot method is given in Section 3.2.1. POD is frequently employed in the context of ROM applied to Partial Differential Equations (PDEs), using the POD-Galerking projection to speed-up CFD simulations (Hijazi et al., 2020; Rapún and Vega, 2010; Rowley et al., 2004).
Recent studies have demonstrated promising results in the coupling of POD with RNNs. Maulik et al. (2020) employed a POD-LSTM reduced-order model to predict the dynamics of a flow governed by the Burgers' equations, successfully capturing the system's transient behavior. Mohan and Gaitonde (2018) employed the same POD-LSTM method on a 3D-DNS turbulence dataset, addressing the challenge of accuracy degradation over prediction time. Nazvanova et al. (2023) applied POD-LSTM to predict a flow past a cylinder.

3.1.3.3 | Autoencoders

While POD is a linear decomposition technique, autoencoders offer a nonlinear approach to ROM. AE are neural networks designed to compress data into a lower-dimensional manifold or latent space, effectively capturing complex, nonlinear relationships within the data. This makes AE particularly powerful for ROM in cases where the underlying system dynamics are not adequately represented by linear methods like POD. Recent studies have compared the performance of POD and autoencoders in the context of ROM. For instance, Nair and Goza (2020) developed a fully connected autoencoder-based ROM and demonstrated its superior performance compared to traditional POD-based ROM.
Similar to how LSTM has been combined with POD, recent studies have explored the use of autoencoders coupled with LSTM. This approach is predominantly applied to 1D multivariate time series prediction across various fields (Gensler et al., 2016; Jung and Choi, 2021; Sagheer and Kotb, 2019), including air pollution forecasting (Samal et al., 2020; Xayasouk et al., 2020; Xu and Yoneda, 2019). Few studies have implemented an AE-LSTM architecture for 2D flow prediction. Xu et al. (2023) compared the

POD-LSTM and CAE-LSTM (Convolutional Autoencoder) approaches for predicting flow past a cylinder. The CAE-LSTM demonstrated superior performance, particularly in long-term predictions. Eivazi et al. (2020) drew the same conclusion by comparing a fully-connected AE with POD-LSTM on unsteady flows. Gonzalez and Balajewicz (2018) developed a CAE-LSTM model named nMOR, which successfully predicted the trajectory of multiple Gaussian vortices. Both linear and nonlinear techniques will be used and compared in this work.

3.2 | POD-LSTM for Air Pollution Forecasting

3.2.1 | Methodology

Figure 3.4 illustrates the coupling principle between POD and LSTM modeling. Input sequences from the full order space are first reduced using the POD snapshot method, from which a basis of N_r vectors are extracted. They define the latent space. Time series from this latent space are then fed into the LSTM network, which predicts the subsequent time steps. These low-dimensional predictions are then reconstructed into the full order space through the inverse POD transform, yielding the predicted output sequences for the following time steps.

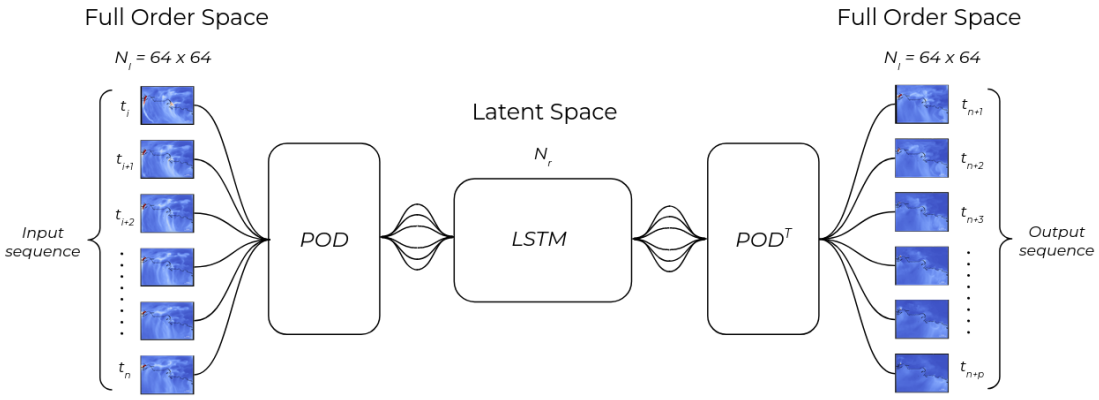

Figure 3.4: The POD-LSTM architecture.

3.2.1.1 | POD

In a first approach, only one feature is considered for the model. This feature is represented by the scalar field $\mathbf{u}(\mathbf{x}, t)$ such that $\mathbf{u} : \mathcal{T} \to \mathbb{R}^{N_{\text{lat}} \times N_{\text{long}}}$ and $\mathbf{x} \in \mathbb{R}^3$, where $\mathcal{T} = [0, T]$ and $N_{\text{lat}}, N_{\text{long}}$ are the dimensions of the vector corresponding to latitude and longitude respectively. We then apply the snapshot POD method, starting by building the snapshot matrix:

$$\mathbf{S} = [\hat{\mathbf{u}}^1 \mid \cdots \mid \hat{\mathbf{u}}^{N_t}] \in \mathbb{R}^{N_l \times N_t} \tag{3.10}$$

where $N_l = N_{\text{lat}} \times N_{\text{long}}$ is the number of spatial degrees of freedom and N_t is the number of snapshots, equivalent here to the number of time steps considered. Time-averaged field is subtracted:

$$\begin{aligned} \hat{\mathbf{u}}_h^i &= \mathbf{u}_h^i - \bar{\mathbf{u}}_h \\ \bar{\mathbf{u}}_h &= \frac{1}{N_t} \sum_{i=1}^{N_t} \mathbf{u}_h^i \end{aligned} \tag{3.11}$$

The snapshot matrix is then decomposed into simple elements in the form called "Economy SVD":

$$\mathbf{S} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\intercal \tag{3.12}$$

where $\mathbf{U} \in \mathbb{R}^{N_l \times N_t}$ is the matrix of spatial eigenvectors, $\mathbf{\Sigma} \in \mathbb{R}^{N_t \times N_t}$ is the diagonal matrix containing the eigenvalues $\{\lambda_1, \ldots, \lambda_{N_t}\}$ in decreasing order, and $\mathbf{V} \in \mathbb{R}^{N_t \times N_t}$ is the matrix of temporal eigenvectors. Note that $\mathbf{U}$ and $\mathbf{V}$ are orthonormal, therefore $\mathbf{U}^\intercal\mathbf{U} = \mathbf{V}^\intercal\mathbf{V} = \mathbf{I}_{N_t \times N_t}$.
Let $\mathbf{D} = \mathbf{S}^\intercal\mathbf{S} \in \mathbb{R}^{N_t \times N_t}$ be the spatial correlation matrix and $\mathbf{T} = \mathbf{S}\mathbf{S}^\intercal \in \mathbb{R}^{N_l \times N_l}$ the temporal correlation matrix, then the problem of simple element decomposition of $\mathbf{S}$ amounts to finding the eigenvectors and eigenvalues of $\mathbf{D}$ and $\mathbf{T}$, since we have:

$$\begin{aligned} \mathbf{D}\mathbf{V} &= \mathbf{V}\mathbf{\Sigma}^2 \\ \mathbf{T}\mathbf{U} &= \mathbf{U}\mathbf{\Sigma}^2 \end{aligned} \tag{3.13}$$

The principle of POD is based on the Eckart-Young theorem, which proves the existence of a solution to the minimization problem where the cost function measures the L_2 distance between a given matrix and an approximation matrix with reduced rank:

$$\min_{\text{rank}(\tilde{\mathbf{S}})=r} \left\| \mathbf{S} - \tilde{\mathbf{S}} \right\|_2^2 \tag{3.14}$$

The optimal solution to this problem for an approximation matrix of rank N_r is written as:

$$\tilde{\mathbf{S}} = [\tilde{\mathbf{u}}^1 \mid \cdots \mid \tilde{\mathbf{u}}^{N_t}] = \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r^\intercal = \mathbf{U}_r \mathbf{\Sigma}_r \left[\mathbf{v}_r^1 \mid \cdots \mid \mathbf{v}_r^{N_t}\right]^\intercal \in \mathbb{R}^{N_l \times N_t} \tag{3.15}$$

where $\mathbf{U}_r \in \mathbb{R}^{N_l \times N_r}$ consists only of the first N_r columns of $\mathbf{U}$, $\mathbf{\Sigma}_r \in \mathbb{R}^{N_r \times N_r}$ is the diagonal matrix containing the first N_r eigenvalues $\{\lambda_1, \ldots, \lambda_{N_r}\}$, and $\mathbf{V}_r^\intercal \in \mathbb{R}^{N_r \times N_t}$ contains only the first N_r rows of $\mathbf{V}^\intercal$. The eigenvalues are such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{N_r}$, meaning that the variance is mostly represented in the first modes. Note that the term "energy" is often used to refer to the amount of variance captured by these modes. A first source of error is then related to the POD reduction, and depends directly on the choice of N_r. The optimal projection of the POD method allows us to define the relative projection error as:

$$\frac{\sum_{i=N_r+1}^{N_t} \left\|\hat{\mathbf{u}}^i - \tilde{\mathbf{u}}^i\right\|_{\mathbb{R}^{N_l}}^2}{\sum_{i=1}^{N_t} \left\|\hat{\mathbf{u}}^i\right\|_{\mathbb{R}^{N_l}}^2} = \frac{\sum_{i=N_r+1}^{N_t} \lambda_i^2}{\sum_{i=1}^{N_t} \lambda_i^2} \tag{3.16}$$

3.2.1.2 | LSTM

The prediction is made on the temporal coefficients of the matrix $\mathbf{V}_r^\intercal$, where each row represents a time series of N_t time steps corresponding to a specific mode. Let N_p the number of predictions given by the LSTM model, the snapshot matrix can be extended with predicted values:

$$\hat{\mathbf{S}} = \mathbf{U}_r \mathbf{\Sigma}_r \left[\mathbf{v}_r^1 \mid \cdots \mid \mathbf{v}_r^{N_t} \mid \mathbf{v}_r^{N_t+1} \mid \cdots \mid \mathbf{v}_r^{N_t+N_p}\right]^\intercal \in \mathbb{R}^{N_r \times (N_t+N_p)} \tag{3.17}$$

Instead of passing N_l time series to the ML model in a full order framework, only N_r were used.

A sub-model is trained for each mode and all sub-models share the same LSTM architecture presented here.

The LSTM architecture used in this work is in fact a combination of multiple hidden layers, summarized in Table 3.1. The input sequences, of length N_{inp}, first pass through a 1D-Convolutional layer (Conv1D) with 128 kernels of size 3 and a stride of 1. For more details on the principles of CNNs, Wu (2017) provides a clear and comprehensive explanation. The output from the Conv1D layer is fed into the first LSTM layer, which contains 256 units. This is followed by a second LSTM layer with 128 units, generating one output element in the temporal space for each input sequence. A dropout layer is then applied, discarding 10% of the learned parameters to prevent overfitting. Finally, a Dense layer converts the 128 elements from the dropout layer into N_{out} elements, which

is the number of predicted steps for a single inference. N_{out} also corresponds to the number of steps in the target sequences used for the loss calculation. Multiple sequences are processed together in batches, which helps improving the model's generalization. The batch size used throughout this study is fixed at 16. N_{var} denotes the total number of input variables. In this manuscript, explanatory variables are called independent variables or features, while the predicted variable is referred to as the dependent variable. Consequently, the input variables include both the dependent and independent variables, such that ($N_{\text{var}} = N_{\text{feature}} + 1$). hyperbolic tangent is used as activation function for Conv1D and LSTM layers. Prior to inputting the time series into the model, the data is standardized for each mode.

Table 3.1: LSTM architecture.

Layer	Input shape	Output shape	# Parameters
Conv1D	(Batch_size, N_{inp}, N_{var})	(Batch_size, N_{inp}, 128)	$128 \times (N_{\text{inp}} \times N_{\text{var}} + 1)$
LSTM	(Batch_size, N_{inp}, 128)	(Batch_size, N_{inp}, 256)	$4 \times (256 \times (256 + 128 + 1))$
LSTM	(Batch_size, N_{inp}, 256)	(Batch_size, 128)	$4 \times (128 \times (256 + 128 + 1))$
Dropout	(Batch_size, 128)	(Batch_size, 128)	0
Dense	(Batch_size, 128)	(Batch_size, N_{out})	$N_{\text{out}} \times (128 + 1)$

3.2.1.3 | Dataset

The dataset was entirely built from the outputs of WRF-Chem simulations, using the MOZCART setup. While the initial goal was to cover the entire year of 2021, temporal discontinuities exist in the final dataset due to reasons discussed in Section 3.1.3.1. These gaps, a common challenge in machine learning, must be addressed to ensure proper model training. All variables exhibit the same missing time steps since they were generated from the same simulations. This uniformity significantly simplifies the management of these gaps, as they are consistent across all variables. Moreover, the number of gaps is limited, with only three occurring throughout the entire year of 2021. The longest gap period is 2 months, which cannot be filled with typical methods like interpolation of mean/median fill. The solution adopted to avoid discontinuities in the sequences extracted from the dataset during both training and inference processes, is to identify the four continuous "blocks" of the full dataset. Therefore, a sequence cannot be composed of values from two different blocks. The dataset is then divided into training and test sets with a ratio of 0.8. To ensure proper training and evaluation within each block, this 0.8 split ratio is

applied individually to all four blocks.
The same methodology was applied to all ML experiments presented in the following sections.

3.2.2 | Test on Potential Temperature

This section presents the preliminary study of the POD-LSTM applied to a WRF-Chem output variable. In the initial approach, the model is univariate and does not utilize any features. The predicted variable used for a first test is the perturbation potential temperature, defined as:

$$\theta' = \theta - \bar{\theta}(\bar{z}) \tag{3.18}$$

where $\bar{\theta}$ is the reference state temperature which depends only on mean height $\bar{z}$. Note that θ was introduced in Eq.1.6. The variable was taken at first level of the simulation, which is about 25 m above ground level.

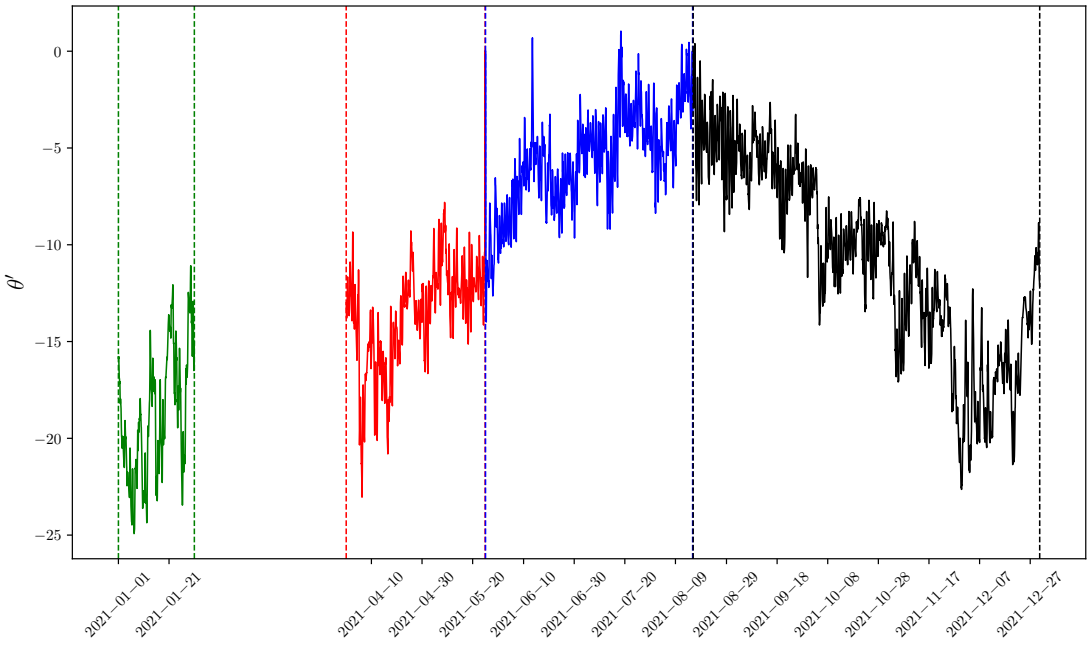

Figure 3.5: Time serie of the mean perturbation potential temperature in K. Each color corresponds to a time block.

Fig.3.5 provides an overview of the data's temporal distribution, with the four time blocks seperated by vertical dashed lines.

The POD transform is applied to the snapshot matrix built on the basis of θ'. Based on the eigenvalues of the POD decomposition, the cumulative mode energy at mode k can be calculated as:

$$\frac{\sum_{i=k+1}^{N_t} \lambda_i}{\sum_{i=1}^{N_t} \lambda_i} \tag{3.19}$$

The cumulative mode energy, as well as the relative projection error (RPE) from Eq.3.16 is shown in Fig.3.6. For this first test case, we chose $N_r = 32$, which represents already about 60% of the total energy. Note that the mean field was removed prior to the POD transform. The 60% figure then refers to the energy in the fluctuating component of the field. With only 32 modes, the RPE is already down to about 10^{-3}.

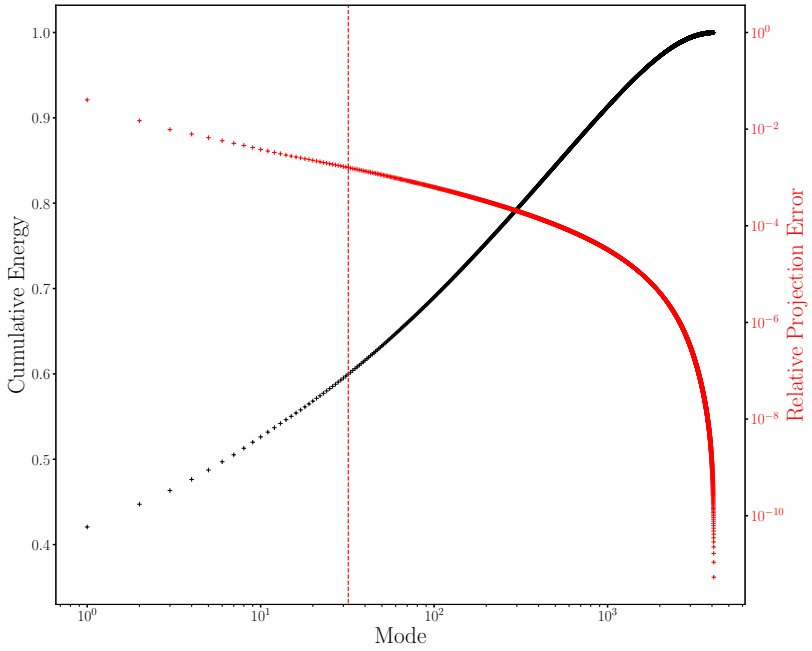

Figure 3.6: POD cumulative mode energy (in black) and relative projection error (in red) for θ'.
Vertical red dashed line represents mode 32.

The performance of the LSTM model varies across modes, with some modes being more periodic and predictable than others. Fig.3.7 shows MSE on the normalized temporal coefficients of the 32 modes. Here, the first mode is significantly better predicted compared to the others, which show relatively similar MSE.

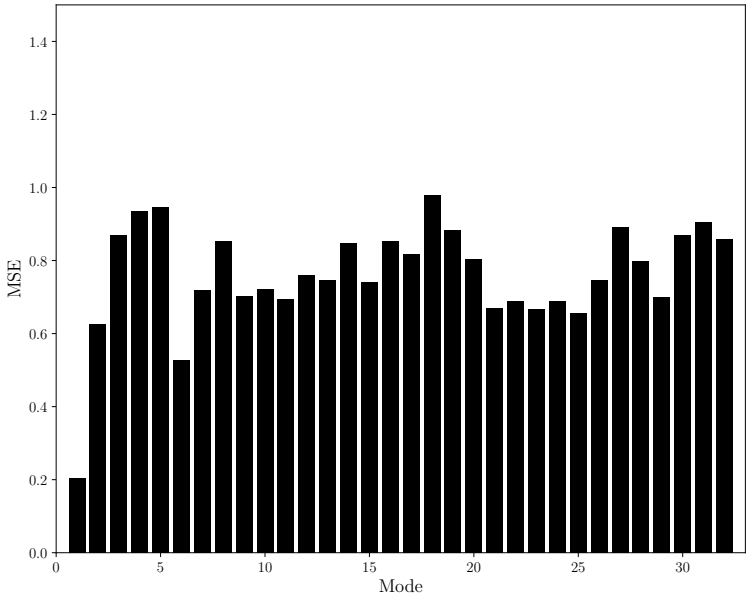

Figure 3.7: MSE on temporal coefficients of the 32 modes.

During the training process, the time series were divided into sequences with $N_{\text{inp}} = 18$ input steps and $N_{\text{out}} = 6$ output steps. After training, the model infers on sequences from the test dataset. The objective is to predict a 72-hour horizon, which is achieved iteratively: at each inference step, the model's output is used as input for the subsequent step, progressively extending the prediction until the full horizon N_p is reached. The relative prediction error is computed as:

$$E_r = \frac{\left\|\hat{\mathbf{S}} - \mathbf{S}\right\|_{\mathbb{R}^{N_l}}}{\|\mathbf{S}\|_{\mathbb{R}^{N_l}}} \tag{3.20}$$

Fig.3.8 gives the evolution of E_r with horizon time. The comparison with a naive prediction based on the mean field demonstrates that POD-LSTM predictions remain significantly more accurate, even after a 72-hour horizon. As expected, prediction errors accumulate during the inference process. For $N_p = 72$ with $N_{\text{out}} = 6$, a total of 12 inference steps are required for the complete prediction. The 95% confidence interval increases with time, as anticipated, though it remains relatively constrained.

Fig.3.9 presents snapshots of predictions at various forecast horizons (1h, 6h, 12h, 18h, 24h). The POD-LSTM model effectively captures spatial patterns. However, as the

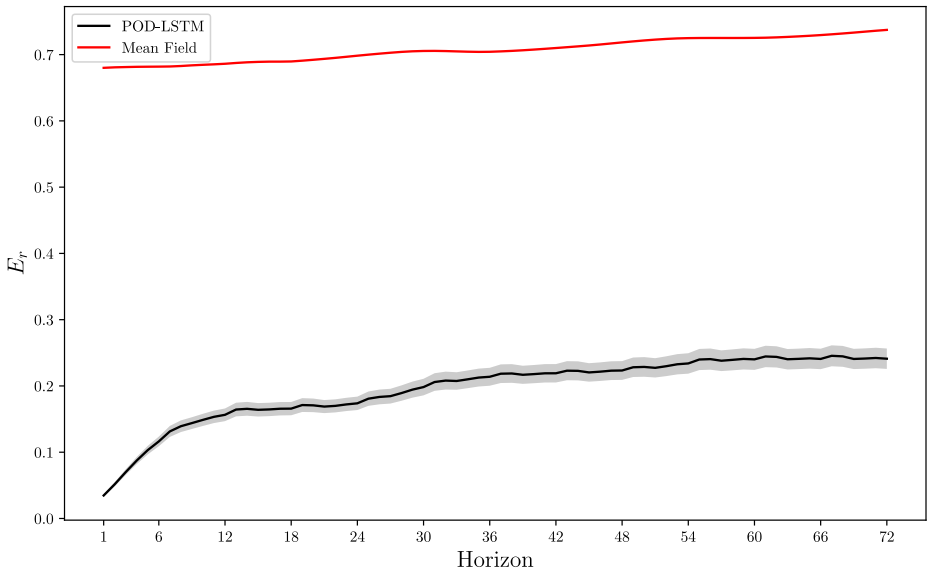

Figure 3.8: Relative prediction error for θ' (in black) and for mean field (in red). 95% confidence interval is represented in shaded grey.

forecast horizon extends, the predicted values show decreasing variance which may be attributed to the length of the output sequences N_{out}. During training, the LSTM model predicts temporal coefficients more effectively in the first time steps following the input sequence. However, as the forward pass progresses through time, it encounters increasing difficulty in predicting accurate values. The model is less penalized when predicting values closer to the mean, which accounts for the observed reduction in variability over time. This tendency toward mean values reduces the risk of large errors, thus aligning with the loss minimization strategy of the LSTM model. This effect is more pronounced with longer output sequence lengths. On the other hand, a shorter output sequence length would exacerbate the accumulation of errors, as more inference steps are required to reach the desired forecast horizon, increasing the overall prediction uncertainty. This first proof of concept has highlighted this trade off regarding N_{out}, and suggests to be careful in the choice of that hyper-parameter.

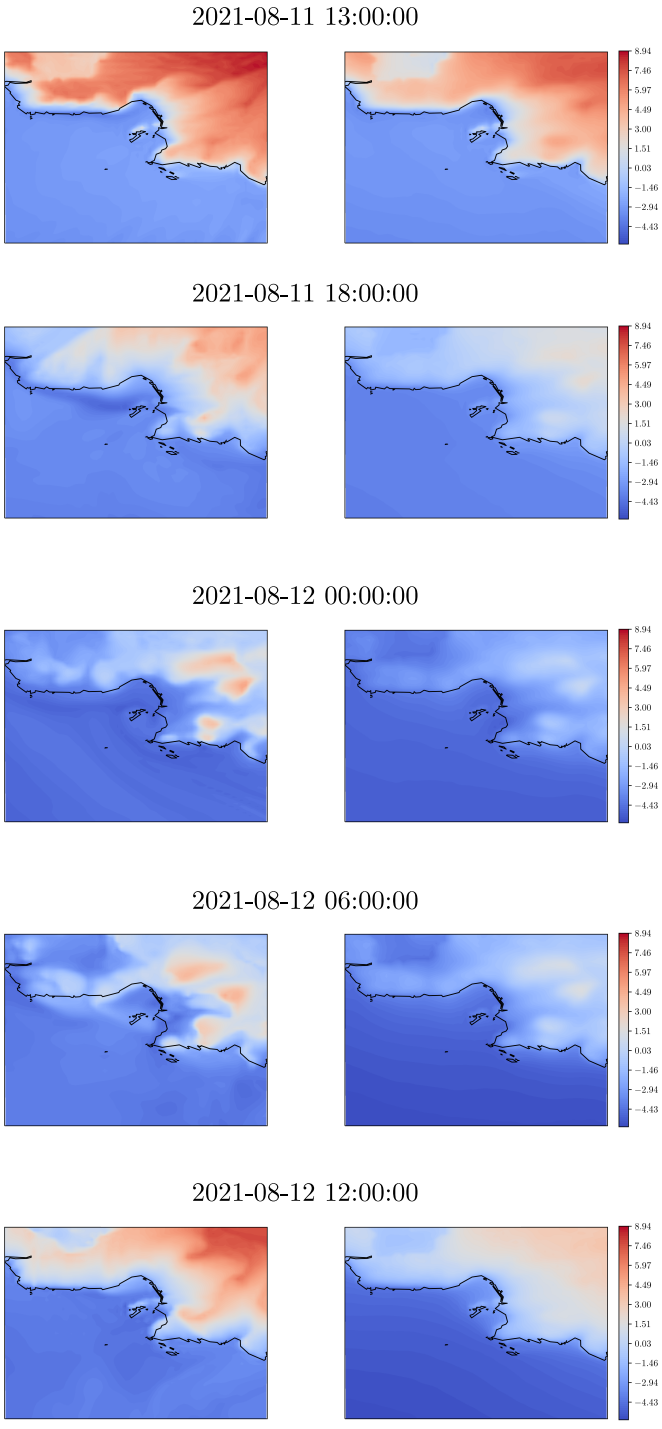

Figure 3.9: Snapshots of original field (left column) and predicted field (right column) in K starting from 2021-08-11 12:00:00.

3.2.3 | Application to $PM_{2.5}$

3.2.3.1 | Data Description

The same temporal partition is applied for $PM_{2.5}$ output variables as for θ'. Fig.3.10 shows the distribution of $PM_{2.5}$ concentration values at 25 m above ground within each block. Overall, no significant seasonal effect is observed, although some slight differences are noted across blocks, such as a higher median for block 3 and more outliers for block 2.

As for θ', both cumulative mode energy and RPE are computed from the POD transform

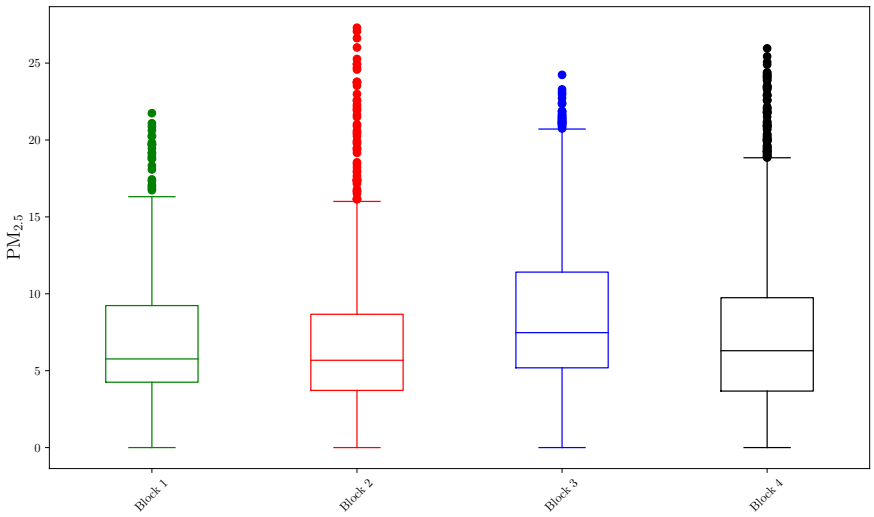

Figure 3.10: Box plots of $PM_{2.5}$ concentration in $\mu g.m^{-3}$ for each block.

of $PM_{2.5}$ snapshot matrix and displayed in Fig.3.11. Compared to the decomposition of θ', the first modes contribute significantly less, indicating that the energy of the field is more evenly distributed across the modes. 32 modes only captures about 30% of the total energy, while RPE is still about 0.1.

To give a comprehensive description of the POD modes, the spatial energy spectrum of each mode is computed. Eq.3.12 can be written in it's functional form as:

$$\mathbf{S} = \sum_{n} \lambda_n \phi_n \psi_n \tag{3.21}$$

where λ_n are the eigenvalues, ϕ_n the spatial modes and ψ_n the temporal coefficients. The Fourrier-transform is applied to each spatial mode:

$$\hat{\phi}_n(k_x, k_y) = \mathcal{F}\left[\phi_n(x, y)\right] \tag{3.22}$$

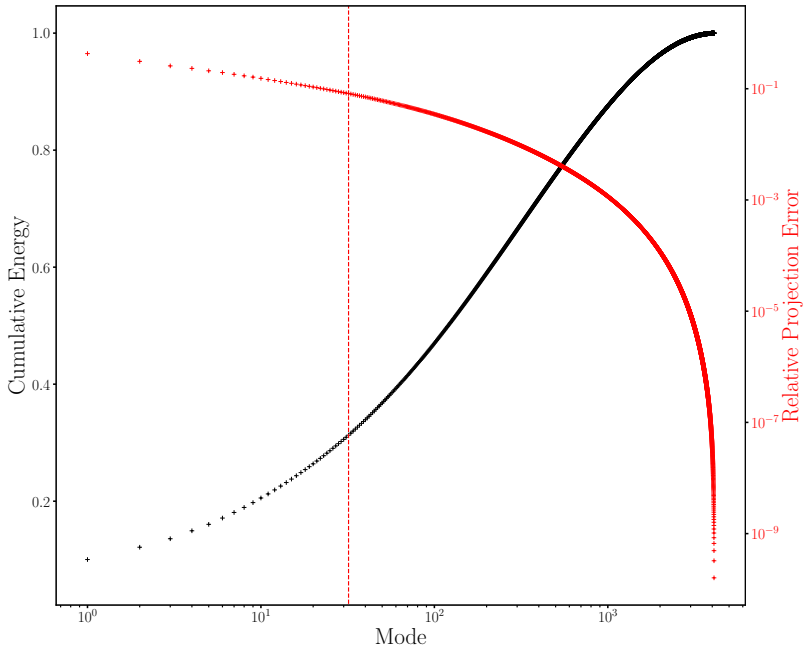

Figure 3.11: POD cumulative mode energy (in black) and relative projection error (in red) for $PM_{2.5}$ concentration.
Vertical red dashed line represents mode 32.

The corresponding energy spectrum is computed as:

$$E(k) = \left|\hat{\phi}_n(k_x, k_y)\right|^2 \tag{3.23}$$

with $k = \sqrt{k_x^2 + k_y^2}$ the wavenumber.
The energy spectrum of the POD spatial modes, represented in Fig.3.12, indicates that the low-order modes capture large scale structures, while the high-order modes represent small scale variations. The energy maxima of each mode progressively shifts from the domain scale to approximately the grid scale as the mode number increases. Since the modes are ordered according to their eigenvalues, this suggests that most of the $PM_{2.5}$ concentration variance happens in the large scales. This highlights an intrinsic problem when using POD as a ROM method, as the reduction of the space involves losing the small scale variations when most of the energy is carried on larger scales. The 32th mode has its energy peak at about 10 km, which is about a fifth of the domain size.

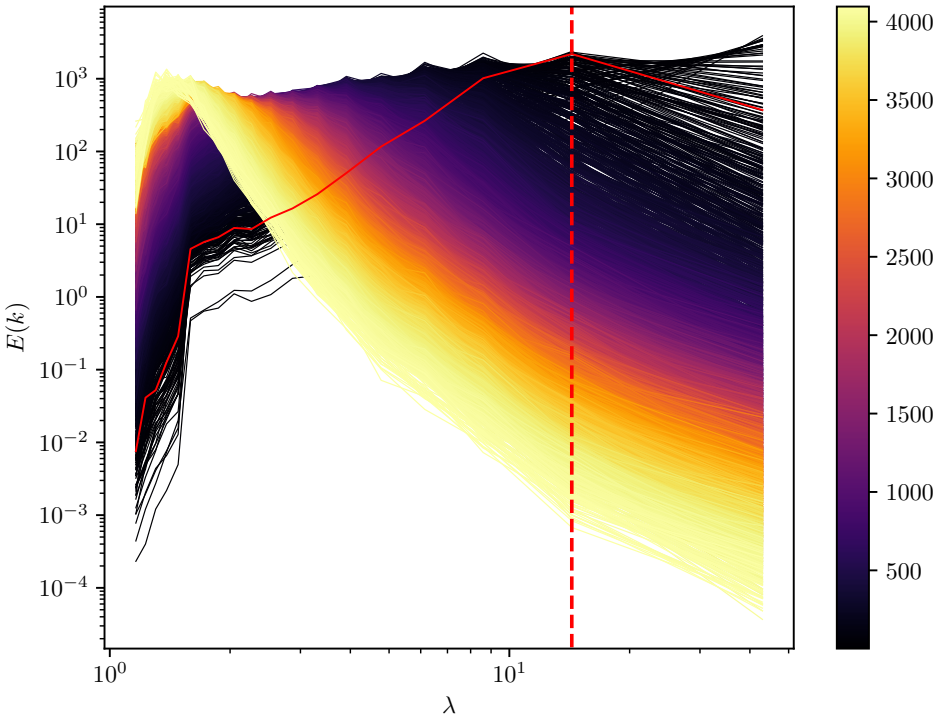

Figure 3.12: Energy spectrum of the POD modes with wavelength $\lambda = \frac{2\pi}{k}$ in km. Colors indicate the mode number. $E(k)$ for mode 32 is in red, while the vertical dashed line indicates its maxima.

3.2.3.2 | Results

The problem of input and output sequence lengths discussed in Section 3.2.2 was investigated here. A series of test simulations were conducted with various input and ouput lengths, in order to identify the best configuration. The MAE is computed on a 72-hour horizon time for each training session, as presented in Fig.3.13. The best configuration gives $N_{\text{inp}} = 18$ and $N_{\text{out}} = 3$. Input sequences that are too short may not fully leverage the LSTM cells' long-term memory, while excessively long sequences could introduce irrelevant information, as $PM_{2.5}$ typically remains in the simulation domain for about half a day, either being swept away by wind or removed through dry or wet deposition. As for the output sequence length, 3 hours appears to be a good compromise to limit both accumulation error and variance decay. Generally, smaller values of N_{inp} tend to result in higher MAE, although the relationship between sequence lengths and MAE is

not clearly defined.

Using this sequence length configuration with 32 POD modes, the relative error pre-

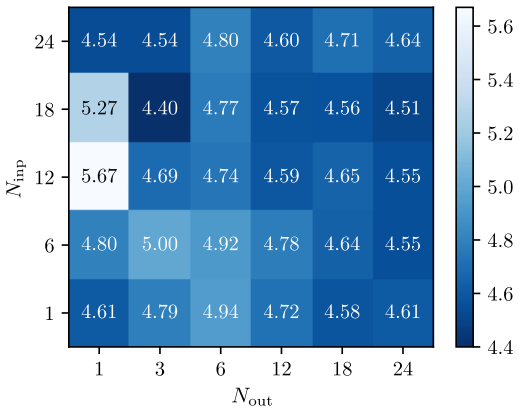

Figure 3.13: MAE in $\mu g \cdot m^{-3}$ for each pair $(N_{\text{inp}}, N_{\text{out}})$.

diction on $PM_{2.5}$ concentration is shown in Fig.3.14. During the first 6 hours, the error rises sharply from 0.2 to 0.4, while the mean field prediction error hovers around 0.7-0.8. After 30 hours, the POD-LSTM error stabilises on a plateau around 0.55, remaining under the mean field prediction error. Although the model captures the initial spatial distribution of $PM_{2.5}$, its predictions tend to converge towards a quasi-stationary solution as the concentration distribution evolves, likely as a strategy to minimize error in the absence of sufficient information. Such conclusion was expected, as air pollution is strongly influenced by meteorological features (see Section 1.3.2). Table 3.2 summarized the performances of the POD-LSTM model.

Table 3.2: Results of POD-LSTM with 32 POD modes. MAE and RMSE are in $\mu g \cdot m^{-3}$. E_r^{72} is the average relative prediction error over a 72h horizon.

	POD-LSTM
MAE	4.4
RMSE	7.28
E_r^{72}	0.54

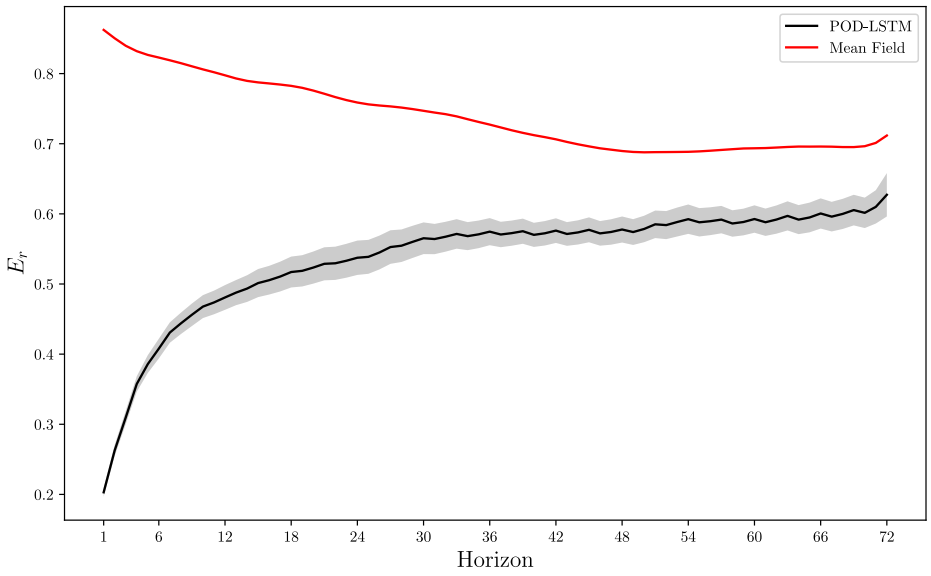

Figure 3.14: Relative prediction error for $PM_{2.5}$ concentration (in black) and for mean field (in red). 95% confidence interval is represented in shaded grey.

3.2.4 | Multivariate Model

Up to this point, only $PM_{2.5}$ concentration has been used as input to the POD-LSTM model. This section presents a multivariate approach of the POD-LSTM model, referred to as the MPOD-LSTM model. A correlation analysis was conducted between $PM_{2.5}$ concentration levels and meteorological features derived from WRF-Chem outputs. Details are provided in Appendix D. This analysis provides qualitative insights into the relationships between features and led us to focus on the following meteorological variables: U and V the two horizontal components of the wind, P the pressure and Q_{vapor} the water vapor mixing ratio. Additionally, V_d the dry deposition velocity, is assumed to be a meteorological variable, as it can be computed using only parametric information without requiring concentration data. A non-meteorological variable is added: $PM_{2.5}$ emissions, noted here as $E_{PM_{2.5}}$, which were used as input for the WRF-Chem model.

3.2.4.1 | Shared-basis methodology

To incorporate features, one straightforward approach would be to apply a POD transform to each feature individually and use the resulting temporal coefficients as inputs for each LSTM. However, because the POD transform is tailored to each variable, the spatial modes can vary significantly between features, potentially introducing irrelevant information for LSTM predictions. To address this issue, the features' snapshot matrices are projected in the POD modes of the predicted variable. This ensures mode-specific information from features, and simplifies the implementation of the MPOD-LSTM model. Let $\mathbf{S}_0 = \sum_n \lambda_n^0 \phi_n^0 \psi_n^0$ the POD transform of the dependent variable. Since $\{\phi_n^0\}$ forms an orthonormal basis of $\mathbb{R}^{N_l}$, the projection of an independent variable onto this reference basis can be expressed as:

$$\mathbf{S}_i = \sum_n \langle \mathbf{S}_i, \phi_n^0 \rangle \phi_n^0 \tag{3.24}$$

By identification, the equivalent temporal coefficients and eigenvalues of the projected POD transform for feature i are respectively:

$$\psi_n^i = \frac{\langle \mathbf{S}_i, \phi_n^0 \rangle}{\| \langle \mathbf{S}_i, \phi_n^0 \rangle \|_{\mathbb{R}^{N_t}}}; \quad \lambda_n^i = \| \langle \mathbf{S}_i, \phi_n^0 \rangle \|_{\mathbb{R}^{N_t}} \tag{3.25}$$

The energy carried by a mode can be computed for each feature from the λ_n^i of the projections as represented on Fig.3.15. Although the shared-basis reduction method is not optimal, the information captured from the projections onto the first 32 modes remains significant. It is unsurprising that P, Q_{vapor}, and the two horizontal wind components exhibit a higher proportion of their energy in the leading modes compared to $PM_{2.5}$ concentration. This is because these variables are associated with larger-scale phenomena than $PM_{2.5}$ concentration. On the contrary, $E_{PM_{2.5}}$ and to a lesser extent V_d are poorly represented in the first modes, which might cause a consequent loss of relevant information after reduction.

Using an output sequence length of $N_{\text{out}} = 1$ is more advantageous when incorporating features. The model predicts only the dependent variable, while the features are continuously sourced from the WRF-Chem dataset at each inference step. By setting $N_{\text{out}} = 1$ the model maximizes the frequency of feature updates per prediction. Algorithm 1 describes this inference process in details.

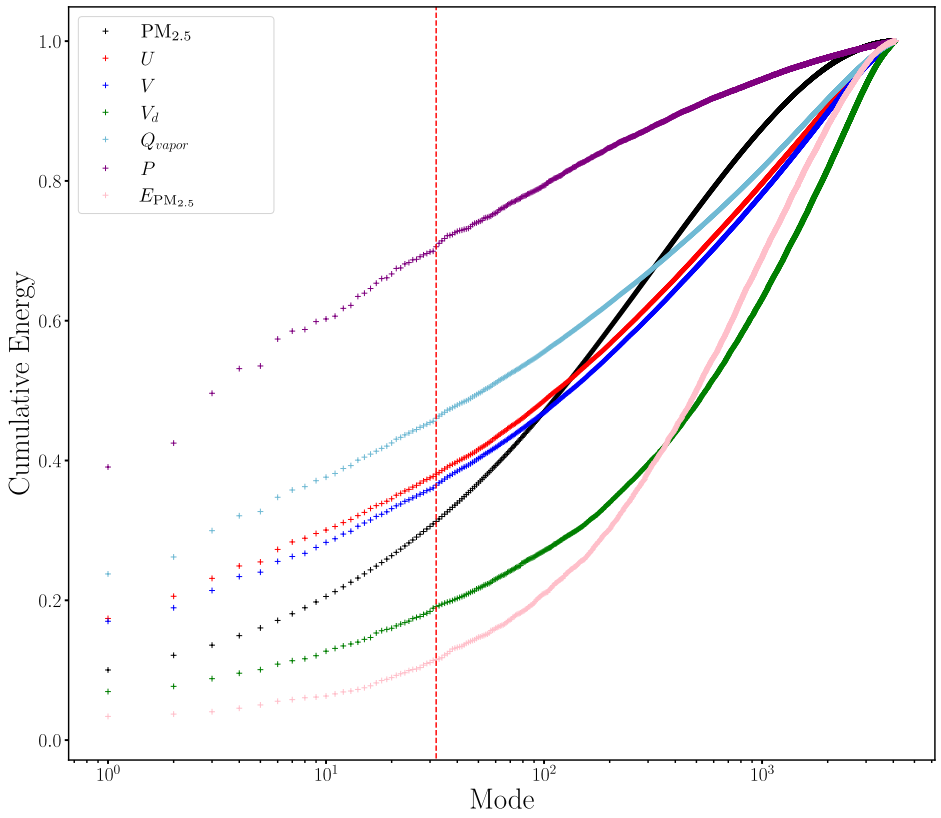

Figure 3.15: Cumulative mode energy of features from the projection on $PM_{2.5}$ POD basis.

3.2.4.2 | Results

Based on the pre-selected features outlined earlier, a series of tests were performed using various feature configurations as input. The optimal configuration was achieved when U, V, V_d and $E_{PM_{2.5}}$ were used together as features. Further testing on input sequence length showed that, for the multivariate model, the optimal results were obtained with $N_{inp} = 24$. Relative prediction error using 32 POD modes are presented in Fig.3.16. The sharp increase in the first 6 hours of prediction is similar to that of the univariate results. However, relative error stabilizes on a plateau around 0.4. This stabilization of the prediction error is explained by the update of features at each time steps during inference, providing valuable information on the long-term. Figure 3.17 presents snapshots of the

Algorithm 1: LSTM Inference Process for the multivariate POD-LSTM.

Inputs: Input list of independent variables:
$X = \left[\left[\psi_n^i(k) \right]_{i=1}^{N_{\text{feature}}} \mid k = 1, 2, \ldots, N_{\text{inp}} \right]$
Input list of dependent variable: $Y = [\psi_n^0(1), \psi_n^0(2), \ldots, \psi_n^0(N_{\text{inp}})]$
for $j = 1$ *to* N_p **do**
  Step 1: Feed current X and Y into the LSTM model: $[X, Y]$
  Step 2: The LSTM outputs the prediction for the dependent variable: $\hat{\psi}_n^0(N_{\text{inp}} + j)$
  Step 3: Update X and Y for the next time step:
  $X \leftarrow \left[\psi_n^i(1+j), \psi_n^i(2+j), \ldots, \psi_n^i(N_{\text{inp}} + j) \right]$
  $Y \leftarrow \left[\psi_n^0(1+j), \psi_n^0(2+j), \ldots, \psi_n^0(N_{\text{inp}}), \hat{\psi}_n^0(N_{\text{inp}}+1), \ldots, \hat{\psi}_n^0(N_{\text{inp}}+j) \right]$
end

original WRF-Chem outputs alongside the POD-LSTM predictions at various forecast horizons. The results demonstrate that even at a 24-hour horizon, the model effectively captures the major concentration variations. The predicted fields from the machine learning model exhibit smoother concentration gradients, as small-scale fluctuations are filtered out due to the dimensionality reduction in the POD basis. However, certain localized high-concentration areas near the boundaries, as in the snapshot from 2021-08-12 12:00:00, are not captured. This discrepancy is attributed to the cropping of the original 102 x 102 grid to 64 x 64 cells. Since $E_{\text{PM}_{2.5}}$ was also cropped, high emissions from the Fos region for example were excluded from the model's predictions. Table 3.3 summarized prediction performances of the MPOD-LSTM. The inclusion of features resulted in a reduction of about 25% on E_r^{72}.

Table 3.3: Results of MPOD-LSTM with 32 POD modes. MAE and RMSE are in $\mu g \cdot m^{-3}$. E_r^{72} is the average relative prediction error over a 72h horizon.

	MPOD-LSTM
MAE	3.65
RMSE	6.16
E_r^{72}	0.41

3.2.5 | Conclusion on POD-LSTM

A reduced-order model for the temporal prediction of a complex 2D scalar field was developed using a data-driven approach. The dataset consists exclusively of WRF-Chem

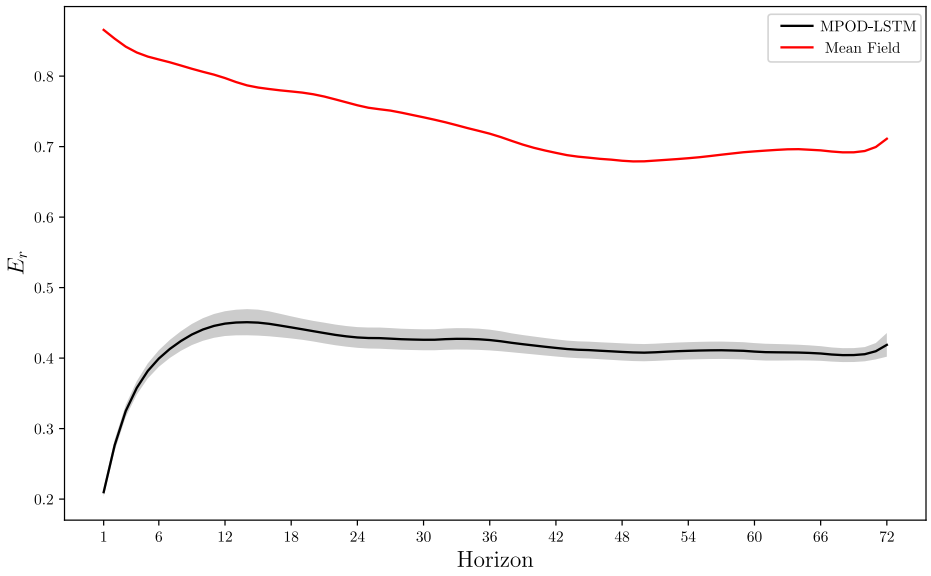

Figure 3.16: Relative prediction error for $PM_{2.5}$ concentration (in black) and for mean field (in red). 95% confidence interval is represented in shaded grey.

outputs, with simulations covering most of 2021. The snapshot POD method was employed as a ROM, while temporal predictions in the latent space were handled by an RNN architecture, primarily using LSTM. While the univariate POD-LSTM model demonstrated effective performance in predicting temperature over a 72-hour forecast horizon, its accuracy significantly deteriorated when applied to more complex dynamics, such as $PM_{2.5}$ concentration. The nonlinear dynamics of the concentration field are inadequately captured by the linear mode decomposition inherent in the POD approach, limiting the model's effectiveness. Autoencoders then appear as an efficient alternative to POD, providing a nonlinear reduction method.

The multivariate MPOD-LSTM model was developed by projecting independent variables onto the POD-reduced basis of the dependent variable. The results demonstrated that this multivariate approach significantly enhanced prediction accuracy, particularly for long-term forecasts.

While this method gives promising results, it does not effectively optimize the reduction of features information, which limits the model's overall performance and its ability to capture complex interrelationships between variables.

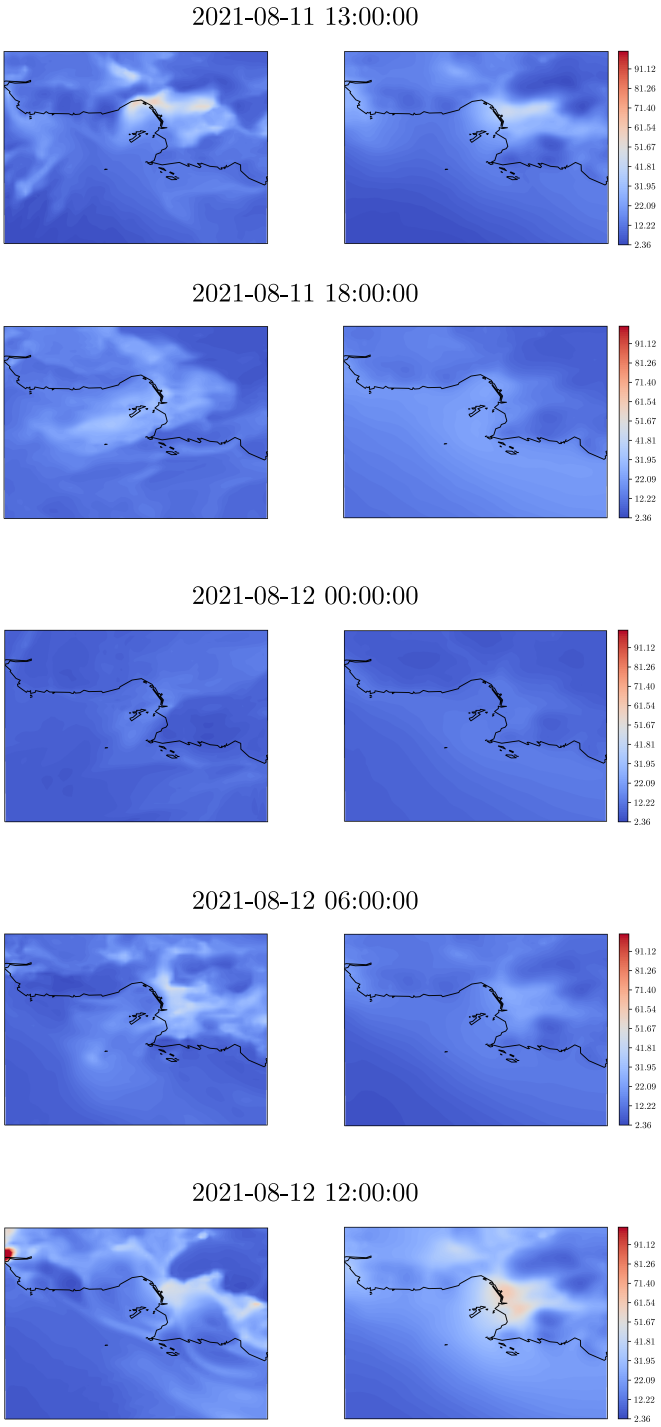

Figure 3.17: Snapshots of original field (left column) and predicted field (right column) in $\mu g \cdot m^{-3}$ starting from 2021-08-11 12:00:00.

3.3 | Autoencoder-LSTM

The core principle of an autoencoder is to generate a low-dimensional representation of the input vector and then reconstruct it from the coordinates in the latent space, as illustrated in Fig.3.18. A first part, named encoder, takes the input vector $\mathbf{u} \in \mathbb{R}^{N_l}$ and

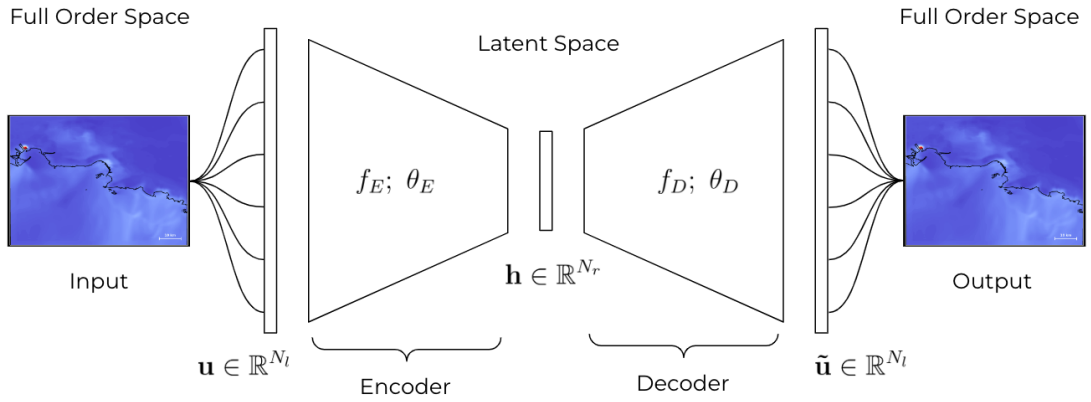

Figure 3.18: Principle of an Autoencoder.

outputs the low-order representation $\mathbf{h} \in \mathbb{R}^{N_r}$:

$$f_E(\mathbf{u}; \theta_E) = \mathbf{h} \tag{3.26}$$

where f_E is a complex nonlinear function and θ_E are the weights relative to the encoder. The decoder then give an approximation of the input vector from this low-order representation:

$$f_D(\mathbf{h}; \theta_D) = \tilde{\mathbf{u}} \in \mathbb{R}^{N_l} \tag{3.27}$$

The autoencoder is trained to minimize the reconstruction error:

$$\hat{\theta}_E; \hat{\theta}_D = \arg\min_{\theta_E; \theta_D} \mathcal{L}(\tilde{\mathbf{u}}, \mathbf{u}) \tag{3.28}$$

where $\mathcal{L}$ is the loss function that measures the discrepancy between $\mathbf{u}$ and $\tilde{\mathbf{u}}$.

3.3.1 | Autoencoder and POD

In essence, an autoencoder can be viewed as a nonlinear generalisation of the Proper Orthogonal Decomposition (POD). Considering a linear autoencoder, the encoding and

decoding operations can be written as:

$$f_E(\mathbf{u};\theta_E) = \mathbf{W}_E\mathbf{u} = \mathbf{h}; \quad f_D(\mathbf{u};\theta_D) = \mathbf{W}_D\mathbf{h} = \tilde{\mathbf{u}} \tag{3.29}$$

When using the L_2 distance as the loss function, we aim to minimize the reconstruction error:

$$\mathcal{L}(\tilde{\mathbf{u}}, \mathbf{u}) = \|\mathbf{u} - \mathbf{W}_D\mathbf{W}_E\mathbf{u}\|_2^2 \tag{3.30}$$

with $\mathbf{W}_E \in \mathbb{R}^{N_r \times N_l}$ and $\mathbf{W}_D \in \mathbb{R}^{N_l \times N_r}$. If we impose $\mathbf{W}_D = \mathbf{W}_E^\intercal$, the minimisation problem is the same as the one solve by POD, as expressed in Eq.3.14. The basis formed by the vectors of $\mathbf{W}_E$ spans the same subspace as the POD solution. However, the basis vectors are neither guaranteed to be orthonormal, nor are they necessarily ordered by significance, as they are in POD. In this way, the POD can be seen as a linear autoencoder where the dimensionality reduction is optimally performed by projecting onto a subspace spanned by the dominant modes of the data. In contrast, a nonlinear autoencoder is expected to capture more complex, higher-order structures in the data, beyond what POD can achieve with its linear subspace.

3.3.2 | Model Architecture

In the literature, there exist two major methods to build a low-dimensional nonlinear manifold with autoencoders.

- The fully-connected autoencoder uses a series of fully-connected layers. In the encoder, each layer outputs a vector of lower dimension than its input. In a symmetric way, the decoder layers have the opposite configuration. This approach showed promising results (Eivazi et al., 2020). However, it presents two main issues. First, the input snapshots are flatten in a single column vector, which makes spatial correlations more implicit than if the data was kept in its original form. The second issue is the computational cost of fully-connected layers, for which the required number of trained parameters increases significantly with the spatial dimension of the original data.
- The convolutional autoencoder uses convolutional layers instead of fully-connected layers as a way to reduce dimensionality. The convolution operations detailed hereafter exploits local spatial correlations of the field and has lower computational expenses of a fully-connected autoencoder (Gonzalez and Balajewicz, 2018).

For the reasons mentioned above, the CAE method is applied in this study. The nMOR model developed by Gonzalez and Balajewicz (2018) has been applied to the propagation

of multiple vortices in a 64 x 64 domain. This type of flow shares multiple characteristics with the problem of predicting concentration fields. Unlike the relatively simple unsteady flows present in ROM studies (flow past a cylinder, viscous Burgers equations), multiple vortex propagation exhibits less pseudo-periodic behavior and involves multiple directions of propagation. The nMOR model is directly applied to $PM_{2.5}$ concentration prediction. The dataset and its temporal partitioning is the same as in previous sections.
As the nMOR model is described in details in Gonzalez and Balajewicz (2018), only a brief overview of the model architecture is presented here.

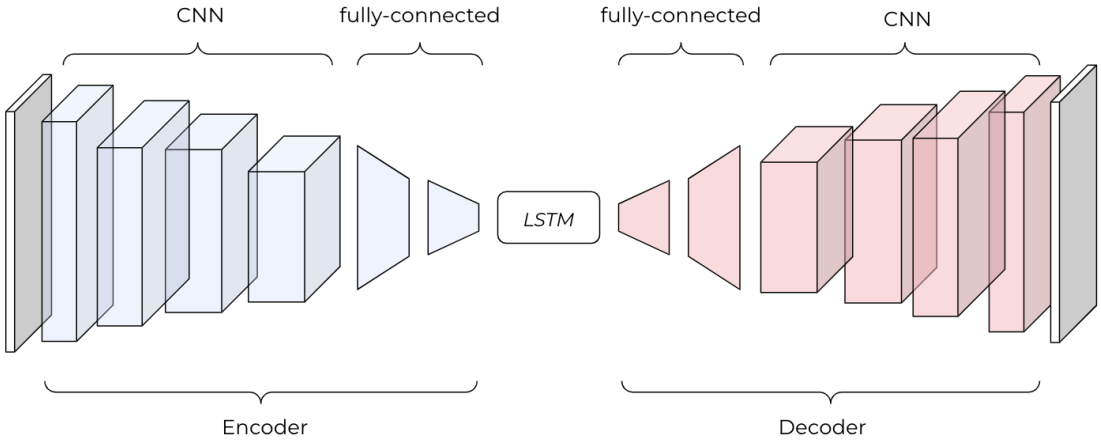

Figure 3.19: nMOR architecture.

Fig.3.19 illustrates the architecture of the nMOR model, composed of 4 CNN layers followed by 2 fully-connected layers for the encoder and a symmetric configuration for the decoder. Detailed information on CNN layers are provided in Table 3.4.

Table 3.4: Convolutional encoder (left) and decoder (right) filter sizes and strides. Source: Gonzalez and Balajewicz (2018).

Layer	filter size	filter	stride
1	5×5	4	2×2
2	5×5	8	2×2
3	5×5	16	2×2
4	5×5	32	2×2

Layer	filter size	filter	stride
9	5×5	16	2×2
10	5×5	8	2×2
11	5×5	4	2×2
12	5×5	1	2×2

The LSTM layer in between takes as input the low-dimensional representation $\mathbf{h}^0$ of input $\mathbf{u}^0$, the first element of a set of consecutive snapshots $\mathbf{U} = \{\mathbf{u}^0, ..., \mathbf{u}^{N_{\text{out}}}\}$. From this initial state, the LSTM evolves in time for N_{out} steps. The predictions $\hat{\mathbf{H}} = \{\hat{\mathbf{h}}^1, ..., \hat{\mathbf{h}}^{N_{\text{out}}}\}$ are then passed to the decoder, which outputs decoded predictions in the full-dimensional

space $\hat{\mathbf{U}} = \{\hat{\mathbf{u}}^1, ..., \hat{\mathbf{u}}^{N_{\text{out}}}\}$. During the training process, two loss functions are defined for both the autoencoder and the LSTM, expressed as follows:

$$\mathcal{L}_{AE} = \frac{1}{N_{\text{out}}+1} \sum_{i=0}^{N_{\text{out}}} \frac{\left\|\hat{\mathbf{u}}^i - \mathbf{u}^i\right\|_F^2}{\|\mathbf{u}^i\|_F^2 + \epsilon}; \quad \mathcal{L}_{LSTM} = \frac{1}{N_{\text{out}}} \sum_{i=1}^{N_{\text{out}}} \frac{\left\|\hat{\mathbf{h}}^i - \mathbf{h}^i\right\|_2^2}{\|\mathbf{h}^i\|_2^2 + \epsilon} \tag{3.31}$$

where ϵ is a small positive number and $\|\cdot\|_F$ is the Frobenius norm.
During back propagation, the gradients are computed on a joint global loss:

$$\mathcal{L} = \alpha\mathcal{L}_{AE} + \beta\mathcal{L}_{LSTM} \tag{3.32}$$

with $\alpha = \beta = 0.5$.

3.3.3 | Results

The model was trained with a batch size of 16 and input sequences of 48 hours. In order to compare results with that of the POD-LSTM method, the latent space dimension is set to 32, as the number of selected POD modes.
Fig.3.20 shows relative error prediction of the nMOR model. The performance doesn't vary much with horizon. This is related to the implementation of the LSTM, which is trained on long output sequences ($N_{\text{out}} = 48$) and only one snapshot as intitialization ($N_{\text{inp}} = 1$). The performance of the model is worse than the POD-LSTM in the first prediction steps, but nMOR outperforms in the long-term predictions. Overall, for a 72-hour horizon, nMOR is slightly better than POD-LSTM as Table 3.5 confirms, with all metrics lower than in Table 3.2.

Table 3.5: Results of the nMOR model. MAE and RMSE are in $\mu g \cdot m^{-3}$. E_r^{72} is the average relative prediction error over a 72h horizon.

	nMOR
MAE	4.47
RMSE	7.06
E_r^{72}	0.48

3.3.4 | Conclusion on Autoencoder-LSTM

The nMOR model developed by Gonzalez and Balajewicz (2018) was employed to predict $PM_{2.5}$ concentrations using outputs from WRF-Chem simulations. The results show that nMOR predictions are better than the POD-LSTM model in the long-term but less

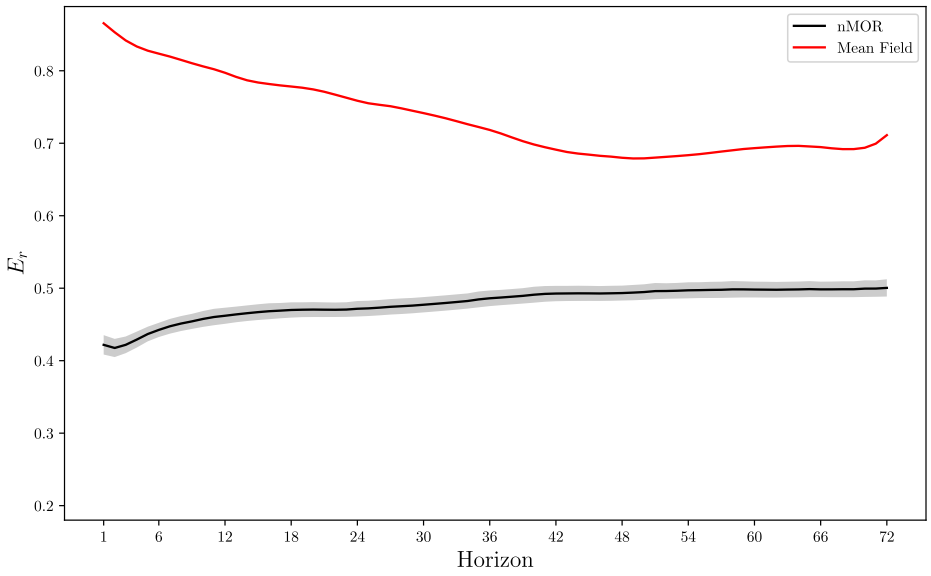

Figure 3.20: Relative prediction error for $PM_{2.5}$ concentration (in black) and for mean field (in red). 95% confidence interval is represented in shaded grey.

accurate in the first predicted time steps. The autoencoder's higher theoretical potential indicates that its current configuration may not be optimal. Allowing input sequences to span multiple time steps is thought to improve prediction accuracy.

3.4 | Discussions

A comprehensive dataset of weather and air pollution data was generated to train a hybrid deep learning model for predicting $PM_{2.5}$ concentrations with minimal computational costs. To address the high-dimensionality of the data, two ROM techniques were applied: POD and autoencoders. This study introduced an POD-LSTM model, along with a novel methodology for a multivariate POD-LSTM. The univariate POD-LSTM model is prone to error accumulation, as multiple inference iterations are often required when the prediction horizon exceeds the LSTM output length. The lengths of the input and output sequences were identified as key factors influencing this effect. The POD shared-basis approach led to a notable improvement in prediction performance, particularly for long-term forecasts, compared to the univariate model. However, the first POD

modes may not capture small-scale variations, which are lost during the dimensionality reduction process. This could affect the model's sensitivity to initial emission conditions. A sensitivity analysis of the model to emission disturbances could offer valuable insights. Since POD is limited to linear combinations of modes, a non-linear approach was also explored using a convolutional autoencoder-LSTM model. However, the nMOR model showed only minimal improvement over the POD-LSTM. The use of a single initial state for the LSTM to predict future steps may not be optimal for this problem. One potential improvement would be to use a longer input sequence, which would require adjustments to the implementation. A multivariate CAE-LSTM is also part of the future directions for this study, as we believe the CNN layers could efficiently capture spatio-temporal dependencies between the different features.

Table 3.6 presents a comparison of computational costs between the developed models and WRF-Chem simulations. The inference process, referred to as "online" computation, represents the use of the model in an operational context. Both machine learning models demonstrate a negligible computational cost when compared to the significantly higher demands of WRF-Chem simulations. Even during the training process, both models show lower computational costs, with the nMOR model being slightly more efficient than the POD-LSTM. It is worth noting that the MPOD-LSTM model exhibits similar training times to the POD-LSTM, as the number of features has minimal impact on the number of trainable parameters. However, the computational cost of the POD-LSTM/MPOD-LSTM model is proportional to the number of selected POD modes.

The MPOD-LSTM model requires meteorological features at each inference step, meaning that running NWP simulations remains necessary within this operational framework. However, since the ML model does not require chemical or aerosol data as inputs, it relies solely on the meteorological forecasts generated by WRF. The computation of a 72-hour weather forecast requires only 3h on 64 cores, instead of 24h when adding chemistry and aerosol processes, which aligns with the demands of operational functionality. Consequently, the operational system discussed here is not fully data-driven.

The goal of this study was to develop a machine learning reduced-order model capable of

Table 3.6: Computational costs required for training and inference processes for a 72-hour horizon time of the POD-LSTM and nMOR. Times are non-dimensional according to the time required for WRF-Chem simulations as $\frac{\tau_{\text{ML}}}{\tau_{\text{WRF-Chem}}}$.

	Training	Inference
POD-LSTM	0.14	1.1×10^{-3}
nMOR	0.12	2.3×10^{-5}

predicting complex 2D dynamical systems, such as $PM_{2.5}$ concentration. This framework was designed to be applicable to other complex dynamical systems beyond air pollution. However, for operational use and to ensure accurate concentration predictions, the input data should incorporate ground-based measurements. The loss function would then include a term that accounts for deviations from these measurement data.

4

Conclusion

In this concluding section, the main findings and contributions from this three-year doctoral journey are summarized. The relevance of these studies is discussed, highlighting their importance in the wider context of the field. This work has revealed a number of points that deserve further investigation, which are presented here as perspectives. By identifying these lines of research, we hope to encourage future researchers to build on the basis set out in this study.

4.1 | Summary

- **What is the ship contribution to air pollution in Marseille ?**

The ship contribution in Marseille was estimated using WRF-Chem, with ship emissions derived from real-time traffic data provided by GPMM and background emissions from the CIGALE inventory. Results for June 2021 indicated that, on average, ships contributed 6.49% to $PM_{2.5}$ and 4.0% to PM_{10}. Gas-phase species were also analyzed, showing a 1.94% ship contribution to SO_2 and 17.86% to NO_2, while ozone exhibited a negative contribution of -1.21%. The analysis per size of particles showed that ship-emitted fine particles ($\leq 0.625\mu m$) were advected from shore to inland regions, primarily in the north-east direction. Coarser particles remain close from the sources, which are mainly at dock sites in the two port basins, as gravitational settling predominated over the wind flow advection.

Ship pollution is a great concern for populations in the vicinity of the port in Marseille. Several civil associations have raised alarms about hazardous concentration levels, calling for stricter regulations and immediate action to mitigate the impact of ship emissions. The framework developed in this study provides scientific support for both civil

associations and regional policymakers in their efforts to address the issue.

- **How does weather influences ship pollution in Marseille ?**

The same high ship traffic scenario was analyzed under typical seasonal weather conditions. For each season, two 72-hour periods were selected: one with weak winds and one with moderate winds. The ship emissions were more dispersed and resulted in lower concentrations during weak wind conditions compared to moderate wind conditions. Sea-land breeze dynamics plays an important role in the dispersion of pollutant at low wind speeds. Inversions between the two breezes might induce a significant accumulation of ship-emitted pollutants over Marseille. When synoptic wind predominates, as in the moderate wind conditions, ship pollution is distributed as a plume in the prevailing wind direction.

A better understanding of the typical unfavorable conditions during high ship traffic can help prevent severe pollution events. Since ship traffic is often scheduled in advance, restricting port access to large vessels like cruise ships during such periods could significantly reduce ship pollution in Marseille, and overall air pollution levels.

- **Can Machine Learning predict $PM_{2.5}$ concentration ?**

WRF-Chem simulation outputs were used to train a machine learning model. Due to the high dimensionality of the data, Reduced Order Modeling (ROM) techniques were applied to reduce the amount of information fed into the Recurrent Neural Network (RNN), specifically an LSTM. Initially, a univariate POD-LSTM approach was tested, but it highlighted the need for additional information for accurate 72-hour predictions. A new multivariate POD-LSTM method was developed, projecting the independent variables onto the POD-reduced basis of the dependent variable. This approach led to a notable performance improvement, particularly for long-term predictions. However, the POD reduction method is constrained by its reliance on a linear combination of modes. Autoencoders, which generalizes POD to non-linear dynamics, were tested as a non-linear model order reduction (nMOR) technique on the same dataset. Despite this, the nMOR model provided only limited improvements.

Operational air quality models are important, specially in a large metropole like Marseille, to alarm inhabitants and take specific dispositions during such episodes. The

computational cost of a CTM such as WRF-Chem can be a great challenge to ensure operational use of the model. ML is then a powerful tool to greatly reduce computation expenses. This study sets the basis for the development of a data-driven operational air quality model in Marseille. The application of the codes implemented during this work can be extended to computational fluid dynamics in general.

4.2 | Perspectives

- **Gas-phase prediction accuracy**

The comparison between the MOZAIC setup and ground measurements showed relatively high accuracy for aerosols, but lower accuracy for NO_2 and SO_2. To improve gas-phase predictions, the sources of these high uncertainties must be identified. One key factor to investigate is the emission inventory. Comparing results with other WRF-Chem chemistry schemes could help determine whether the discrepancies with ground measurements stem from inaccuracies in the emission inputs or limitations within the model itself.

- **Emission Scenarios**

In the context of a study by the Municipality of Marseille to evaluate the health benefits of various regulations, the framework developed during this PhD could be used to assess specific scenarios involving the reduction of ship emissions. Potential scenarios to consider include the installation of onshore power supply, fuel regulations, quotas on cruise ships, and the installation of on-board PM filters. This study would assist in decision-making by quantifying the reduction in pollutant concentrations for each scenario, providing valuable data to support regulatory actions.

- **Sensitivity analysis of the ML model**

A key objective of this PhD project was to enable the model to incorporate real-time ship emissions rather than relying on annual averages, as most operational models do. Therefore, it is crucial that the machine learning model developed for operational use accurately responds to variations in ship emissions. A sensitivity analysis should be conducted to understand if ship emissions are properly captured by the model.

- **Autoencoder**

The results from the autoencoder method still require improvement. Theoretically, the autoencoder should outperform the POD-LSTM, as POD is a linear autoencoder, which significantly reduces the ensemble of solutions for the reduction. While the nMOR model was designed for the specific problem of the propagation of vortices, the implementation of the LSTM might not be optimal for the problem of $PM_{2.5}$ concentration. Indeed, the inference involves only a initial state of the flow and from this initial state the LSTM evolves in time. However, it has been seen that, in the case of POD-LSTM, the input sequence length has a great influence on the prediction performance. The prediction of $PM_{2.5}$ concentration requires past information, with an optimal length of around 18 hours in the case of POD-LSTM. Future work should focus on implementing a CAE-LSTM model that accommodates varying input sequence lengths. Similar to the POD-LSTM approach, a multivariate CAE-LSTM model should be developed. By using a 3D-CNN-based encoder, the full range of information from the independent variables can be incorporated, rather than relying on the filtered data obtained from the POD-reduced basis projection. This would allow for a more comprehensive and accurate representation of the system.

- **Operational air quality model**

The long term objective of the framework developed during this PhD is to propose an operational model for air quality in Marseille, based on real-time ship traffic. While AtmoSud provides daily air quality maps, it relies on annual averages for ship emissions, which may not capture high ship emission events, such as when the east basin hosts 5 or 6 cruises simultaneously. Incorporating real-time ship traffic data on a daily basis would necessitate either advance scheduling information from the GPMM authorities or the use of AIS data. However, utilizing AIS data would require additional processing to convert the raw data into emission estimates. The STEAM code developed by Jalkanen et al. (2009) could be used for that purpose. Ultimately, the development of a fully operational framework that integrates all elements, from WRF-Chem inputs to the ML model's outputs, would mark the completion of this work.

A

Land Use

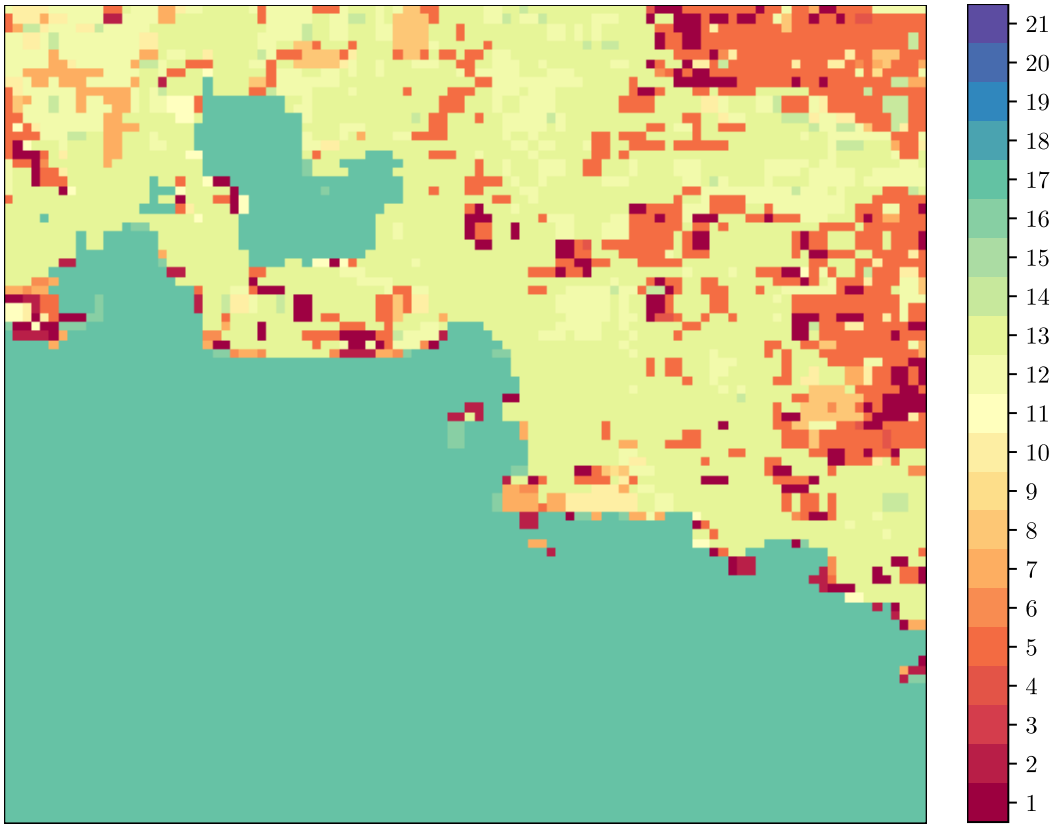

Figure A.1: 21-MODIS land use categories on domain D03.

ID	Category	Season	ALBD	SLMO	SFEM	SFZ0	THERIN	SCFX	SFHC
1	Evergreen Needleleaf Forest	Summer	12.0	0.30	0.95	50.0	4.0	3.33	29.2e5
		Winter	12.0	0.60	0.95	50.0	5.0	3.00	29.2e5
2	Evergreen Broadleaf Forest	Summer	12.0	0.50	0.95	50.0	5.0	1.67	29.2e5
		Winter	12.0	0.50	0.95	50.0	5.0	1.67	29.2e5
3	Deciduous Needleleaf Forest	Summer	14.0	0.30	0.94	50.0	4.0	2.86	25.0e5
		Winter	14.0	0.60	0.94	50.0	5.0	2.60	25.0e5
4	Deciduous Broadleaf Forest	Summer	16.0	0.30	0.93	50.0	4.0	2.63	25.0e5
		Winter	16.0	0.60	0.93	50.0	5.0	2.40	25.0e5
5	Mixed Forests	Summer	13.0	0.30	0.94	50.0	4.0	2.11	41.8e5
		Winter	13.0	0.60	0.94	50.0	6.0	1.12	41.8e5
6	Closed Shrublands	Summer	22.0	0.10	0.93	10.0	3.0	1.56	20.8e5
		Winter	22.0	0.20	0.93	10.0	4.0	1.30	20.8e5
7	Open Shrublands	Summer	20.0	0.15	0.88	10.0	3.0	2.14	20.8e5
		Winter	20.0	0.25	0.88	10.0	4.0	1.24	20.8e5
8	Woody Savannas	Summer	20.0	0.10	0.93	15.0	3.0	1.56	20.8e5
		Winter	20.0	0.20	0.93	15.0	4.0	1.30	20.8e5
9	Savannas	Summer	20.0	0.15	0.92	15.0	3.0	2.00	25.0e5
		Winter	20.0	0.15	0.92	15.0	3.0	2.00	25.0e5
10	Grasslands	Summer	19.0	0.15	0.92	7.5	3.0	2.37	20.8e5
		Winter	19.0	0.30	0.92	7.5	4.0	2.00	20.8e5
11	Permanent Wetlands	Summer	14.0	0.42	0.95	30.0	5.5	1.32	35.5e5
		Winter	14.0	0.725	0.95	30.0	6.0	1.32	35.5e5
12	Croplands	Summer	18.0	0.30	0.935	7.5	4.0	2.71	25.0e5
		Winter	18.0	0.60	0.935	7.5	4.0	2.00	25.0e5
13	Urban and Built-Up	Summer	18.0	0.10	0.88	50.0	3.0	1.67	18.9e5
		Winter	18.0	0.10	0.88	50.0	3.0	1.67	18.9e5
14	Cropland/Natural Mosaic	Summer	16.0	0.25	0.92	6.5	4.0	2.56	25.0e5
		Winter	16.0	0.40	0.92	6.5	4.0	2.00	25.0e5
15	Snow and Ice	Summer	55.0	0.95	0.98	1.0	5.0	0.00	9.0e25
		Winter	55.0	0.95	0.98	1.0	5.0	0.00	9.0e25
16	Barren or Sparsely Vegetated	Summer	25.0	0.02	0.85	6.5	2.0	0.81	12.0e5
		Winter	25.0	0.05	0.85	6.5	2.0	0.81	12.0e5
17	Water	Summer	8.0	1.00	0.98	0.01	6.0	0.00	9.0e25
		Winter	8.0	1.00	0.98	0.01	6.0	0.00	9.0e25
18	Wooded Tundra	Summer	15.0	0.50	0.93	15.0	5.0	2.67	9.0e25
		Winter	15.0	0.60	0.93	15.0	5.0	1.75	9.0e25
19	Mixed Tundra	Summer	15.0	0.50	0.92	10.0	5.0	2.67	9.0e25
		Winter	15.0	0.60	0.92	10.0	5.0	1.75	9.0e25
20	Barren Tundra	Summer	15.0	0.02	0.90	6.0	2.0	1.60	12.0e5
		Winter	15.0	0.05	0.90	6.0	5.0	1.80	12.0e5
21	Lakes	Summer	8.0	0.02	0.98	0.01	3.0	1.67	18.9e5
		Winter	8.0	0.02	0.98	0.01	3.0	1.67	18.9e5

Table A.1: 21-MODIS Land use categories and associated parameters (Summer and Winter values). Parameters from left to right are: albedo (ALBD %), soil moisture availability (SLMO $\times 100\%$), surface emissivity (SFEM $\times 100\%$), surface roughness length (SFZ0 $\times 10^{-2}$m), thermal inertia (THERIN $\times 4.184 \times 10^2 \mathrm{J \cdot m^{-2} \cdot K^{-1} \cdot s^{-1/2}}$), and surface heat capacity (SFHC $\times 10^6 \mathrm{J \cdot m^{-2} \cdot K^{-1}}$).

B

Ship Categories

French	English
AUTRES VRACS	Other Bulk
BARGE OCEA.VERTI	Ocean Vertical Barge
BARGE OCEAN.RORO	Ocean RoRo Barge
CAR CARRIER	Car Carrier
CAR-FERRIE	Car Ferry
CARGO CONVENT.	Conventional Cargo
CARGO REFRIGE.	Refrigerated Cargo
CARGO SPECIALIS.	Specialized Cargo
GAZ LIQUEFIES	Liquefied Gas
MINERALIER	Bulk Carrier
NAV. CIT. VRAC ALIM	Bulk Food
NAV. DE PLAISANCE	Pleasure Craft
NAV. NON CLASSE	Unclassified Navigation
NAV. PROD. CHIM	Chemical Tanker
NAV. TRAV. MARITIME	Maritime Work Vessel
NAVIRE DE GUERRE	Warship
NAVIRE DE PECHE	Fishing Vessel
NAVIRE RO-RO	Ro-Ro Vessel
NAVIRE SCIENTIFI	Scientific Vessel
P. CONT. CELLULAIRE	Cellular Container Ship
P.CONTE.POLYVALE.	Multipurpose Container Ship
P.CONTENEURS ROUL	Rolling Containers
PAQUEBOT	Cruise Ship
PETROLIER	Tanker
PORTE-BARGES	Barge Carrier
REMOR.RAVITAI	Supply Tug
ROPAX	Roll-on/Roll-off Passenger
VRAQUIER SOLIDE	Solid Bulk Carrier

C

NO_2 and SO_2 Ship Contribution

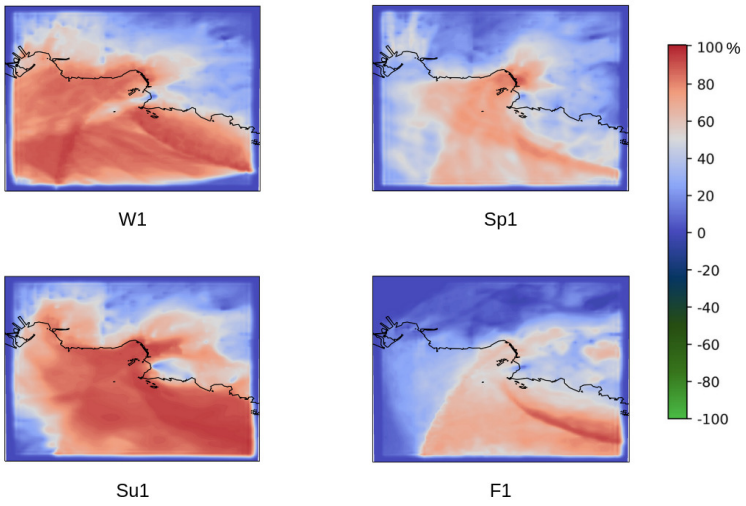

Figure C.1: 72-hour averaged contribution of ships to NO_2 concentration in % for weak winds.

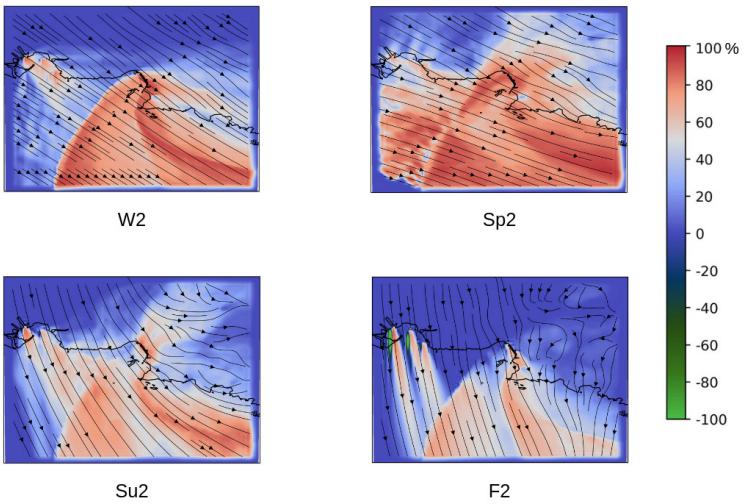

Figure C.2: 72-hour averaged contribution of ships to NO_2 concentration in % for moderate winds.

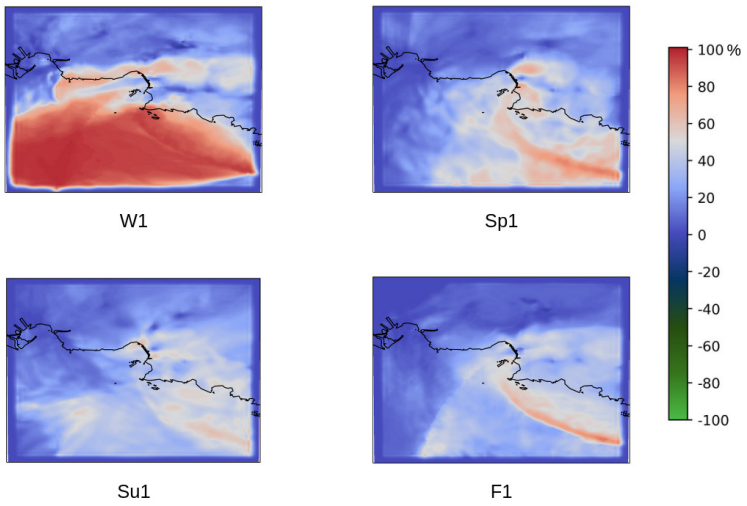

Figure C.3: 72-hour averaged contribution of ships to SO_2 concentration in %
for weak winds.

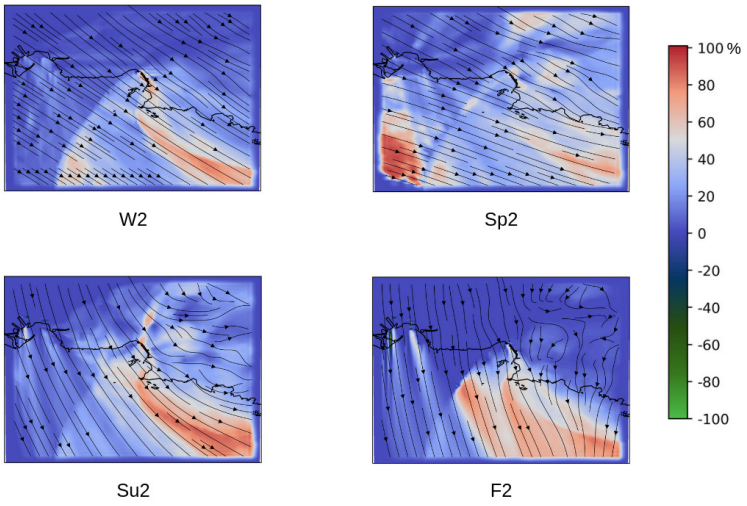

Figure C.4: 72-hour averaged contribution of ships to SO_2 concentration in %
for moderate winds.

D

Features Correlations

Pearson inter-correlations were calculated for each grid cell, and the average across the grid cells was used as the final correlation values, which are presented in Fig.D.1.

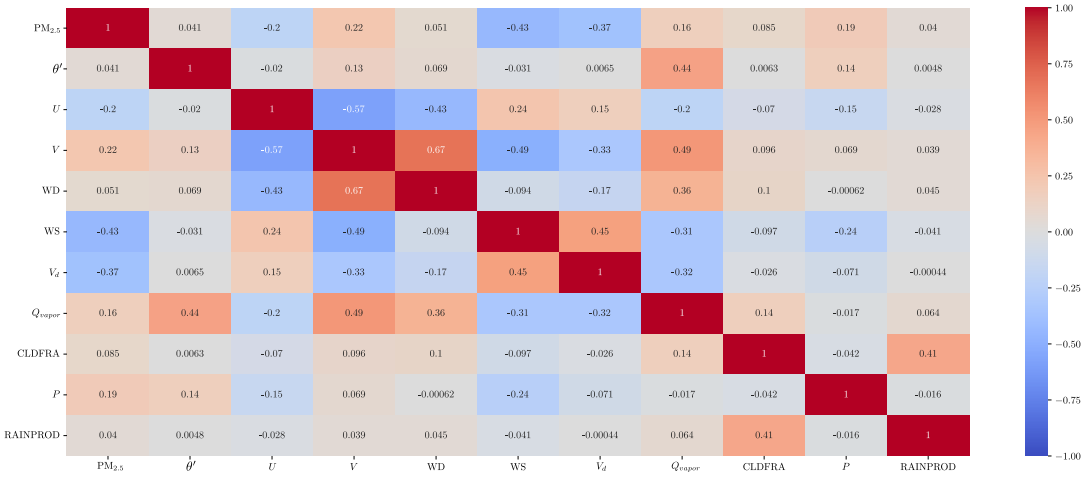

Figure D.1

E

Publication

Atmospheric Environment 329 (2024) 120542

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

Modeling of air pollution due to marine traffic in Marseille

E. Chevet *, O. Boiron, F. Anselmet

Aix Marseille University, CNRS, Centrale Méditerranée, IRPHE UMR 7342, Marseille, France

HIGHLIGHTS

- Ship contribution to Pollution in Marseille area is estimated with WRF-Chem.
- A high ship traffic scenario analyzed under eight typical meteorological conditions.
- Relative ship contribution is much higher for NO_2 and SO_2 than for $PM_{2.5}$ and PM_{10}.
- Under low-wind conditions, ship pollution is more widespread than in moderate wind.
- Sea–land breeze transitions seem to enhance $PM_{2.5}$ accumulation above Marseille.

ARTICLE INFO

Keywords:
Atmospheric pollution
Ship traffic
Numerical weather prediction
Chemical emissions
Sea–land breeze

ABSTRACT

Marine traffic is constantly increasing and is unlikely to decrease in the years to come. However, merchant and tourist ships still use fuels with a high pollutant content, particularly in the Mediterranean sea where regulations are more flexible than those applied in northern European waters. The pollutants of interest in this work, which include particulate matter (PM_{10}, $PM_{2.5}$), NO_2, SO_2 and ozone, are responsible for respiratory diseases such as airway irritation, cardiopulmonary and lung cancers. The aim of this study is to estimate the impact of maritime traffic on air quality in Marseille in case of high port occupancy. The same ship traffic emission scenario will be studied under a series of typical meteorological conditions selected on the basis of a statistical seasonal analysis. This approach therefore differs from studies based on annualized ship traffic scenarios, which do not capture worst-case pollution. In order to do so, we have parameterized the numerical weather forecasting model WRF (Weather Research and Forecasting) that allows the inclusion of chemistry and atmospheric transport equations. The relative difference in concentration between a situation with and without ship traffic allows us to evaluate the influence of ships on air quality. The results show an average ship contribution of between 1.5% and 5.7% for $PM_{2.5}$, with some local maximum of about 25% near the port. Contribution levels for gaseous species are much higher, with values reaching 40% for SO_2 and 80% for NO_2. The analysis of various meteorological conditions has highlighted a significant influence of wind speed and sea–land breeze dynamic on ship contribution.

1. Introduction

The seaways are increasingly frequented, both for development of international trade and for tourism. Although maritime activity has experienced a recent slowdown due to COVID-19, projections for 2050 anticipate a growth in global international maritime trade, with estimates ranging from a 25% to 250% increase (Anon, 2022). Besides, 70% of emissions related to maritime transport occur within 400 km from the coast (Endresen et al., 2003). In fact, most cruise ships continue to run their engines at berth, directly affecting the health of populations living near port areas. Various studies have highlighted the adverse effects of ship emissions on public health (Corbett et al., 2007; Viana et al., 2020; Contini and Merico, 2021). Since 2020, the International Maritime Organization (IMO) has implemented measures to curb ship fuel emissions under Annex VI of the MARPOL (Marine Pollution Convention) (Anon, 2016), imposing a sulfur content limit of 0.5% m/m (mass to mass). Despite these efforts to regulate fuel pollutant content, elevated levels of sulfur oxides (SOx), nitrogen oxides (NOx), and fine particles (PM_{10} and $PM_{2.5}$) persist in ship emissions. The Mediterranean Sea is under consideration for designation as a Sulfur Emission Control Area (SECA) zone in 2025, which would lower the sulfur content limit down to 0.1% m/m. A recent study conducted in the context of the Baltic Seas, where the SECA is already in effect, assessed the benefits of adopting such a policy, resulting in an estimated reduction in premature deaths attributable to maritime traffic of about

* Corresponding author.
E-mail address: elliot.chevet@centrale-marseille.fr (E. Chevet).

https://doi.org/10.1016/j.atmosenv.2024.120542
Received 6 December 2023; Received in revised form 12 April 2024; Accepted 25 April 2024
Available online 30 April 2024
1352-2310/© 2024 Elsevier Ltd. All rights reserved.

Figure E.1: First page of publication in the journal of Atmospheric Environment.

References

IMO. RESOLUTION MEPC.280(70). Amendments to the Annex of the Protocol of 1997 to Amend the International Convention for the Prevention of Pollution from Ships, 1973, as Modified by the Protocol of 1978 Relating There to (Amendments to MARPOL Annex VI). Technical report, 2016. URL https://www.imo.org/en/KnowledgeCentre/IndexofIMOResolutions/Pages/MEPC-2016-2017.aspx.

DNV GL. MARITIME FORECAST TO 2050—Energy Transition Outlook 2022. Technical report. Available online:. Technical report, 2022. URL https://www.dnv.com/maritime/publications/maritime-forecast-2022/download-the-report.html.

Abirami, S. and Chitra, P. Regional air quality forecasting using spatiotemporal deep learning. *Journal of Cleaner Production*, 283:125341, 2021.

Ackermann, I. J., Hass, H., Memmesheimer, M., Ebel, A., Binkowski, F. S., and Shankar, U. Modal aerosol dynamics model for europe: Development and first applications. *Atmospheric environment*, 32(17):2981–2999, 1998.

AtmoSud. Qualité de l'air 2021. Technical report, AtmoSud, 2022.

Barregard, L., Molnàr, P., Jonson, J. E., and Stockfelt, L. Impact on population health of baltic shipping emissions. *International Journal of Environmental Research and Public Health*, 16(11), 2019. ISSN 1660-4601. doi: 10.3390/ijerph16111954.

Bastin, S., Drobinski, P., Dabas, A., Delville, P., Reitebuch, O., and Werner, C. Impact of the rhône and durance valleys on sea-breeze circulation in the marseille area. *Atmospheric Research*, 74(1):303–328, 2005. ISSN 0169-8095. doi: https://doi.org/10.1016/j.atmosres.2004.04.014.

Binkowski, F. S. and Shankar, U. The regional particulate matter model: 1. model description and preliminary results. *Journal of Geophysical Research: Atmospheres*, 100(D12):26191–26209, 1995.

Boikos, C., Siamidis, P., Oppo, S., Armengaud, A., Tsegas, G., Mellqvist, J., Conde, V., and Ntziachristos, L. Validating cfd modelling of ship plume dispersion in an urban environment with pollutant concentration measurements. *Atmospheric Environment*, 319:120261, 2024.

Chen, B., You, S., Ye, Y., Fu, Y., Ye, Z., Deng, J., Wang, K., and Hong, Y. An interpretable self-adaptive deep neural network for estimating daily spatially-continuous pm2.5 concentrations across china. *Science of The Total Environment*, 768:144724, 2021. ISSN 0048-9697. doi: https://doi.org/10.1016/j.scitotenv.2020.144724.

Chen, D., Zhao, N., Lang, J., Zhou, Y., Wang, X., Li, Y., Zhao, Y., and Guo, X. Contribution of ship emissions to the concentration of pm2. 5: A comprehensive study using ais data and wrf/chem model in bohai rim region, china. *Science of the Total Environment*, 610:1476–1486, 2018.

Cheng, M., Fang, F., Navon, I. M., Zheng, J., Tang, X., Zhu, J., and Pain, C. Spatio-temporal hourly and daily ozone forecasting in china using a hybrid machine learning model: Autoencoder and generative adversarial

networks. *Journal of Advances in Modeling Earth Systems*, 14(3):e2021MS002806, 2022. doi: https://doi.org/10.1029/2021MS002806. e2021MS002806 2021MS002806.

Cheng, Q., Wang, X., Chen, D., Ma, Y., Zhao, Y., Hao, J., Guo, X., Lang, J., and Zhou, Y. Impact of ship emissions on air quality in the guangdong-hong kong-macao greater bay area (gba): With a particular focus on the role of onshore wind. *Sustainability*, 15(11):8820, 2023.

Chevet, E., Boiron, O., and Anselmet, F. Modeling of air pollution due to marine traffic in marseille. *Atmospheric Environment*, 329:120542, 2024.

Chin, M., Rood, R. B., Lin, S.-J., Müller, J.-F., and Thompson, A. M. Atmospheric sulfur cycle simulated in the global model gocart: Model description and global properties. *Journal of Geophysical Research: Atmospheres*, 105(D20):24671–24687, 2000.

CIGALE. [dataset] CIGALE database - AtmoSud inventory 2019. Technical report, AtmoSud, 2019. URL `https://cigale.atmosud.org/index.php`.

Cimorelli, A., Perry, S., Venkatram, A., Weil, J., Paine, R., Wilson, R., Lee, R., Peters, W., Brode, R., and Paumier, J. Aermod: Description of model formulation. 2004.

Cohen, A. J., Brauer, M., Burnett, R., Anderson, H. R., Frostad, J., Estep, K., Balakrishnan, K., Brunekreef, B., Dandona, L., Dandona, R., et al. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the global burden of diseases study 2015. *The lancet*, 389 (10082):1907–1918, 2017.

Colette, A., Aas, W., Banin, L., Braban, C. F., Ferm, M., Gonzalez Ortiz, A., Ilyin, I., Mar, K., Pandolfi, M., Putaud, J.-P., et al. Air pollution trends in the emep region between 1990 and 2012. 2016.

Corbett, J. J., Winebrake, J. J., Green, E. H., Kasibhatla, P., Eyring, V., and Lauer, A. Mortality from Ship Emissions: A Global Assessment. *Environmental Science & Technology*, 41(24):8512–8518, December 2007. ISSN 0013-936X, 1520-5851. doi: 10.1021/es071686z.

Di, Q., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., Sabath, M. B., Choirat, C., Koutrakis, P., Lyapustin, A., et al. An ensemble-based model of pm2. 5 concentration across the contiguous united states with high spatiotemporal resolution. *Environment international*, 130:104909, 2019.

Drewil, G. I. and Al-Bahadili, R. J. Air pollution prediction using lstm deep learning and metaheuristics algorithms. *Measurement: Sensors*, 24:100546, 2022.

Duffourg, F. and Ducrocq, V. Origin of the moisture feeding the heavy precipitating systems over southeastern france. *Natural Hazards and Earth System Sciences*, 11(4):1163–1178, 2011. doi: 10.5194/nhess-11-1163-2011.

Eivazi, H., Veisi, H., Naderi, M. H., and Esfahanian, V. Deep neural networks for nonlinear model order reduction of unsteady flows. *Physics of Fluids*, 32(10), 2020.

Emmons, L. K., Walters, S., Hess, P. G., Lamarque, J.-F., Pfister, G. G., Fillmore, D., Granier, C., Guenther, A., Kinnison, D., Laepple, T., et al. Description and evaluation of the model for ozone and related chemical tracers, version 4 (mozart-4). *Geoscientific Model Development*, 3(1):43–67, 2010.

Endresen, , Sørgård, E., Sundet, J., Dalsøren, S., Isaksen, I., Berglen, T., and Gravir, G. Emission from international sea transportation and environmental impact. *J. Geophys. Res. Atmos.*, December 2003. doi: 10.1029/2002JD002898.

European Environment Agency. Air quality in Europe 2022. Technical report, 2022. URL `https://www.eea.europa.eu/publications/air-quality-in-europe-2022/`.

Fang, L., Jin, J., Segers, A., Liao, H., Li, K., Xu, B., Han, W., Pang, M., and Lin, H. X. A gridded air quality forecast through fusing site-available machine learning predictions from rfsml v1. 0 and chemical transport

model results from geos-chem v13. 1.0 using the ensemble kalman filter. *Geoscientific Model Development*, 16 (16):4867–4882, 2023.

Federico, S., Pasqualoni, L., Sempreviva, A. M., De Leo, L., Avolio, E., Calidonna, C. R., and Bellecci, C. The seasonal characteristics of the breeze circulation at a coastal mediterranean site in south italy. *Advances in Science and Research*, 4(1):47–56, 2010. doi: 10.5194/asr-4-47-2010.

GBD 2015 Mortality and Causes of Death Collaborators. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990-2015: a systematic analysis for the global burden of disease study 2015. *The Lancet*, 388(10053):1659–1724, Oct 2016. ISSN 0140-6736. doi: 10.1016/S0140-6736(16)31679-8.

Geng, G., Xiao, Q., Liu, S., Liu, X., Cheng, J., Zheng, Y., Xue, T., Tong, D., Zheng, B., Peng, Y., et al. Tracking air pollution in china: near real-time pm2. 5 retrievals from multisource data fusion. *Environmental Science & Technology*, 55(17):12106–12115, 2021.

Gensler, A., Henze, J., Sick, B., and Raabe, N. Deep learning for solar power forecasting—an approach using autoencoder and lstm neural networks. In *2016 IEEE international conference on systems, man, and cybernetics (SMC)*, pages 002858–002865. IEEE, 2016.

Gilik, A., Ogrenci, A. S., and Ozmen, A. Air quality prediction using cnn+ lstm-based hybrid deep learning architecture. *Environmental science and pollution research*, pages 1–19, 2022.

Gonzalez, F. J. and Balajewicz, M. Deep convolutional recurrent autoencoders for learning low-dimensional feature dynamics of fluid systems. *arXiv preprint arXiv:1808.01346*, 2018.

Gourav, Rekhi, J. K., Nagrath, P., and Jain, R. Forecasting air quality of delhi using arima model. In *Advances in Data Sciences, Security and Applications: Proceedings of ICDSSA 2019*, pages 315–325. Springer, 2020.

Grell, G. A. and Dévényi, D. A generalized approach to parameterizing convection combining ensemble and data assimilation techniques. *Geophysical Research Letters*, 29(14):38–1, 2002.

Guenther, A., Karl, T., Harley, P., Wiedinmyer, C., Palmer, P. I., and Geron, C. Estimates of global terrestrial isoprene emissions using megan (model of emissions of gases and aerosols from nature). *Atmospheric Chemistry and Physics*, 6(11):3181–3210, 2006.

Hijazi, S., Stabile, G., Mola, A., and Rozza, G. Data-driven pod-galerkin reduced order model for turbulent flows. *Journal of Computational Physics*, 416:109513, 2020.

Huang, H., Zhou, C., Huang, L., Xiao, C., Wen, Y., Li, J., and Lu, Z. Inland ship emission inventory and its impact on air quality over the middle yangtze river, china. *Science of The Total Environment*, 843:156770, 2022.

Iacono, M. J., Delamere, J. S., Mlawer, E. J., Shephard, M. W., Clough, S. A., and Collins, W. D. Radiative forcing by long-lived greenhouse gases: Calculations with the aer radiative transfer models. *Journal of Geophysical Research: Atmospheres*, 113(D13), 2008.

Jacob, D. J. *Introduction to Atmospheric Chemistry*. Princeton University Press, 1999.

Jalkanen, J.-P., Brink, A., Kalli, J., Pettersson, H., Kukkonen, J., and Stipa, T. A modelling system for the exhaust emissions of marine traffic and its application in the baltic sea area. *Atmospheric Chemistry and Physics*, 9(23):9209–9223, 2009.

Janić, Z. I. Nonsingular implementation of the mellor-yamada level 2.5 scheme in the ncep meso model. 2001.

Jiang, J., Aksoyoglu, S., Ciarelli, G., Baltensperger, U., and Prévôt, A. S. Changes in ozone and pm2. 5 in europe during the period of 1990–2030: Role of reductions in land and ship emissions. *Science of the Total Environment*, 741:140467, 2020.

Jonsson, Å. M., Westerlund, J., and Hallquist, M. Size-resolved particle emission factors for individual ships. *Geophysical Research Letters*, 38(13), 2011.

Jung, G. and Choi, S.-Y. Forecasting foreign exchange volatility using deep learning autoencoder-lstm techniques. *Complexity*, 2021(1):6647534, 2021.

Karl, M., Ramacher, M. O. P., Oppo, S., Lanzi, L., Majamäki, E., Jalkanen, J.-P., Lanzafame, G. M., Temime-Roussel, B., Le Berre, L., and D'anna, B. Measurement and modeling of ship-related ultrafine particles and secondary organic aerosols in a mediterranean port city. *Toxics*, 11(9):771, 2023.

Kauffmann, A., Rouil, L., Real, E., Chang, J.-P., Serveau, L., Gueguen, C., Mathias, E., and Riviere, E. *National Pole of Coordination of Territorial Inventories. Methodological guide for the elaboration of territorial inventories of atmospheric emissions (air pollutants and greenhouse gases). Method of elaboration of territorial inventories of atmospheric emissions*, 2012.

Kumar, U. and Jain, V. Arima forecasting of ambient air pollutants (o 3, no, no 2 and co). *Stochastic Environmental Research and Risk Assessment*, 24:751–760, 2010.

Kusaka, H., Kondo, H., Kikegawa, Y., and Kimura, F. A simple single-layer urban canopy model for atmospheric models: Comparison with multi-layer and slab models. *Boundary-layer meteorology*, 101:329–358, 2001.

Kuzu, S. L., Bilgili, L., and Kiliç, A. Estimation and dispersion analysis of shipping emissions in bandirma port, turkey. *Environment, Development and Sustainability*, 23:10288–10308, 2021.

Li, X., Peng, L., Hu, Y., Shao, J., and Chi, T. Deep learning architecture for air quality predictions. *Environmental Science and Pollution Research*, 23:22408–22417, 2016.

Lumley, J., Yaglom, A., and Tatarski, V. Atmospheric turbulence and radio wave propagation. *Journal of computational Chemistry*, 23(13):1236–1243, 1967.

Ma, Y., Chen, D., Fu, X., Shang, F., Guo, X., Lang, J., and Zhou, Y. Impact of sea breeze on the transport of ship emissions: A comprehensive study in the bohai rim region, china. *Atmosphere*, 13(7), 2022. ISSN 2073-4433. doi: 10.3390/atmos13071094.

Mamarikas, S., Matthias, V., Karl, M., Fink, L., Simonen, P., Keskinen, J., Dal Maso, M., Fridell, E., Moldanova, J., Hallquist, Å., et al. Assessing shipping induced emissions impact on air quality with various techniques: Initial results of the scipper project. *Transportation Research Procedia*, 72:2141–2148, 2023.

Marelle, L., Thomas, J. L., Raut, J.-C., Law, K. S., Jalkanen, J.-P., Johansson, L., Roiger, A., Schlager, H., Kim, J., Reiter, A., et al. Air quality and radiative impacts of arctic shipping emissions in the summertime in northern norway: from the local to the regional scale. *Atmospheric Chemistry and Physics*, 16(4):2359–2379, 2016.

Maulik, R., Mohan, A., Lusch, B., Madireddy, S., Balaprakash, P., and Livescu, D. Time-series learning of latent-space dynamics for reduced-order model closure. *Physica D: Nonlinear Phenomena*, 405:132368, 2020.

McConnell, J. R., Wilson, A. I., Stohl, A., Arienzo, M. M., Chellman, N. J., Eckhardt, S., Thompson, E. M., Pollard, A. M., and Steffensen, J. P. Lead pollution recorded in greenland ice indicates european emissions tracked plagues, wars, and imperial expansion during antiquity. *Proceedings of the National Academy of Sciences*, 115(22):5726–5731, 2018. doi: 10.1073/pnas.1721818115.

Méndez, M., Merayo, M. G., and Núñez, M. Machine learning algorithms to forecast air quality: a survey. *Artificial Intelligence Review*, 56(9):10031–10066, Sep 2023. ISSN 1573-7462. doi: 10.1007/s10462-023-10424-4.

Miola, A. and Ciuffo, B. Estimating air emissions from ships: Meta-analysis of modelling approaches and available data sources. *Atmospheric Environment*, 45(13):2242–2251, April 2011. ISSN 13522310. doi: 10.1016/j.atmosenv.2011.01.046.

Mohan, A. T. and Gaitonde, D. V. A deep learning based approach to reduced order modeling for turbulent flow control using lstm neural networks. *arXiv preprint arXiv:1804.09269*, 2018.

Morrison, H., Curry, J., and Khvorostyanov, V. A new double-moment microphysics parameterization for application in cloud and climate models. part i: Description. *Journal of the atmospheric sciences*, 62(6):1665–1677, 2005.

Mueller, N., Westerby, M., and Nieuwenhuijsen, M. Health impact assessments of shipping and port-sourced air pollution on a global scale: A scoping literature review. *Environmental Research*, 216:114460, 2023. ISSN 0013-9351. doi: https://doi.org/10.1016/j.envres.2022.114460.

Murena, F., Mocerino, L., Quaranta, F., and Toscano, D. Impact on air quality of cruise ship emissions in naples, italy. *Atmospheric Environment*, 187:70–83, 2018.

Nair, N. J. and Goza, A. Leveraging reduced-order models for state estimation using deep learning. *Journal of Fluid Mechanics*, 897:R1, 2020.

Nakanishi, M. and Niino, H. Development of an improved turbulence closure model for the atmospheric boundary layer. *Journal of the Meteorological Society of Japan. Ser. II*, 87(5):895–912, 2009. doi: 10.2151/jmsj.87.895.

Nazvanova, A., Ong, M. C., and Yin, G. A data-driven reduced-order model based on long short-term memory neural network for vortex-induced vibrations of a circular cylinder. *Physics of Fluids*, 35(6), 2023.

Nunes, R. A., Alvim-Ferraz, M. C., Martins, F. G., Penuelas, A. L., Durán-Grados, V., Moreno-Gutiérrez, J., Jalkanen, J.-P., Hannuniemi, H., and Sousa, S. I. Estimating the health and economic burden of shipping related air pollution in the iberian peninsula. *Environment international*, 156:106763, 2021.

Oppo, S., Karl, M., Christos, B., Otto Paul Ramacher, M., Tsegas, G., Moldanova, J., Jutterström, S., Gustafsson, M., Majamäki, E., Jalkanen, J.-P., and Schwarzkopf, D. The scipper project shipping contributions to inland pollution push for the enforcement of regulations. 2023.

Partanen, A., Laakso, A., Schmidt, A., Kokkola, H., Kuokkanen, T., Pietikäinen, J.-P., Kerminen, V.-M., Lehtinen, K., Laakso, L., and Korhonen, H. Climate and air quality trade-offs in altering ship fuel sulfur content. *Atmospheric Chemistry and Physics*, 13(23):12059–12071, 2013.

Piga, D. and Armengaud, A. Air quality impact and green house gases assessment for cruise and passenger ships. *CAIMANS, Marseille*, 2015.

Qi, Q., Yu, F., Nair, A. A., Lau, S. S., Luo, G., Mithu, I., Zhang, W., Li, S., and Lin, S. Hidden danger: The long-term effect of ultrafine particles on mortality and its sociodemographic disparities in new york state. *Journal of Hazardous Materials*, 471:134317, 2024. ISSN 0304-3894. doi: https://doi.org/10.1016/j.jhazmat.2024.134317.

Ramacher, M. O. P., Matthias, V., Aulinger, A., Quante, M., Bieser, J., and Karl, M. Contributions of traffic and shipping emissions to city-scale nox and pm2. 5 exposure in hamburg. *Atmospheric environment*, 237:117674, 2020.

Rao, K. S., Devi, G. L., Ramesh, N., et al. Air quality prediction in visakhapatnam with lstm based recurrent neural networks. *International Journal of Intelligent Systems and Applications*, 11(2):18–24, 2019.

Rapún, M.-L. and Vega, J. M. Reduced order models based on local pod plus galerkin projection. *Journal of Computational Physics*, 229(8):3046–3063, 2010.

Raut, J.-C., Law, K. S., Onishi, T., Daskalakis, N., and Marelle, L. Impact of shipping emissions on air pollution and pollutant deposition over the barents sea. *Environmental Pollution*, 298:118832, 2022.

Rowley, C. W., Colonius, T., and Murray, R. M. Model reduction for compressible flows using pod and galerkin projection. *Physica D: Nonlinear Phenomena*, 189(1-2):115–129, 2004.

Sagheer, A. and Kotb, M. Unsupervised pre-training of a deep lstm-based stacked autoencoder for multivariate time series forecasting problems. *Scientific reports*, 9(1):19038, 2019.

Salameh, D., Pey, J., Bozzetti, C., El Haddad, I., Detournay, A., Sylvestre, A., Canonaco, F., Armengaud, A., Piga, D., Robin, D., et al. Sources of pm2. 5 at an urban-industrial mediterranean city, marseille (france): Application of the me-2 solver to inorganic and organic markers. *Atmospheric Research*, 214:263–274, 2018.

Samal, K. K. R., Babu, K. S., Das, S. K., and Acharaya, A. Time series based air pollution forecasting using sarima and prophet model. In *proceedings of the 2019 international conference on information technology and computer communications*, pages 80–85, 2019.

Samal, K. K. R., Babu, K. S., Panda, A. K., and Das, S. K. Data driven multivariate air quality forecasting using dynamic fine tuning autoencoder layer. In *2020 IEEE 17th India Council International Conference (INDICON)*, pages 1–6. IEEE, 2020.

Schell, B., Ackermann, I. J., Hass, H., Binkowski, F. S., and Ebel, A. Modeling the formation of secondary organic aerosol within a comprehensive air quality model system. *Journal of Geophysical Research: Atmospheres*, 106 (D22):28275–28293, 2001.

Schlesinger, W. H. and Bernhardt, E. S. Chapter 3 - the atmosphere. In Schlesinger, W. H. and Bernhardt, E. S., editors, *Biogeochemistry (Fourth Edition)*, pages 51–97. Academic Press, fourth edition edition, 2020. ISBN 978-0-12-814608-8. doi: https://doi.org/10.1016/B978-0-12-814608-8.00003-7. URL `https://www.sciencedirect.com/science/article/pii/B9780128146088000037`.

Seng, D., Zhang, Q., Zhang, X., gourav, G., and gourav, X. Spatiotemporal prediction of air quality based on lstm neural network. *Alexandria Engineering Journal*, 60(2), 2021.

Shin, M., Kang, Y., Park, S., Im, J., Yoo, C., and Quackenbush, L. J. Estimating ground-level particulate matter concentrations using satellite-based data: a review. *GIScience & Remote Sensing*, 57(2):174–189, 2020.

Siami-Namini, S., Tavakoli, N., and Namin, A. S. A comparison of arima and lstm in forecasting time series. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)*, pages 1394–1401. Ieee, 2018.

Sirovich, L. Turbulence and the dynamics of coherent structures. i. coherent structures. *Quarterly of applied mathematics*, 45(3):561–571, 1987.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Duda, M. G., Huang, X.-Y., Wang, W., Powers, J. G., et al. A description of the advanced research wrf version 3. Technical report, 2008.

Sofiev, M., Winebrake, J. J., Johansson, L., Carr, E. W., Prank, M., Soares, J., Vira, J., Kouznetsov, R., Jalkanen, J.-P., and Corbett, J. J. Cleaner fuels for ships provide public health benefits with climate tradeoffs. *Nature communications*, 9(1):406, 2018.

Soulhac, L., Salizzoni, P., Cierco, F.-X., and Perkins, R. The model sirane for atmospheric urban pollutant dispersion; part i, presentation of the model. *Atmospheric environment*, 45(39):7379–7395, 2011.

Tang, D., Zhan, Y., and Yang, F. A review of machine learning for modeling air quality: Overlooked but important issues. *Atmospheric Research*, page 107261, 2024.

Tie, X., Madronich, S., Walters, S., Zhang, R., Rasch, P., and Collins, W. Effect of clouds on photolysis and oxidants in the troposphere. *Journal of Geophysical Research: Atmospheres*, 108(D20), 2003.

Toscano, D. The impact of shipping on air quality in the port cities of the mediterranean area: A review. *Atmosphere*, 14(7):1180, 2023.

Toscano, D., Murena, F., Quaranta, F., and Mocerino, L. Assessment of the impact of ship emissions on air quality based on a complete annual emission inventory using ais data for the port of naples. *Ocean Engineering*, 232:

109166, 2021.

Van der Tak, C. and Cotteleer, A. Emissions 2009: Netherlands continental shelf, port areas, and ospar region ii. Technical report, Report, 2011.

Viana, M., Hammingh, P., Colette, A., Querol, X., Degraeuwe, B., Vlieger, I. d., and van Aardenne, J. Impact of maritime transport emissions on coastal air quality in Europe. *Atmospheric Environment*, 90:96–105, June 2014. ISSN 13522310. doi: 10.1016/j.atmosenv.2014.03.046.

Viana, M., Rizza, V., Tobías, A., Carr, E., Corbett, J., Sofiev, M., Karanasiou, A., Buonanno, G., and Fann, N. Estimated health impacts from maritime transport in the mediterranean region and benefits from the use of cleaner fuels. *Environment international*, 138:105670, 2020.

Wang, R., Tie, X., Li, G., Zhao, S., Long, X., Johansson, L., and An, Z. Effect of ship emissions on o3 in the yangtze river delta region of china: Analysis of wrf-chem modeling. *Science of The Total Environment*, 683: 360–370, 2019.

Wesely, M. Parameterization of surface resistances to gaseous dry deposition in regional-scale numerical models. *Atmospheric Environment (1967)*, 23(6):1293–1304, 1989. ISSN 0004-6981. doi: https://doi.org/10.1016/0004-6981(89)90153-4.

Wicker, L. J. and Skamarock, W. C. Time-splitting methods for elastic models using forward time schemes. *Monthly weather review*, 130(8):2088–2097, 2002.

Winther, M., Christensen, J. H., Plejdrup, M. S., Ravn, E. S., Eriksson, Ó. F., and Kristensen, H. O. Emission inventories for ships in the arctic based on satellite sampled ais data. *Atmospheric Environment*, 91:1–14, 2014.

Wu, J. Introduction to convolutional neural networks. *National Key Lab for Novel Software Technology. Nanjing University. China*, 5(23):495, 2017.

Xayasouk, T., Lee, H., and Lee, G. Air pollution prediction using long short-term memory (lstm) and deep autoencoder (dae) models. *Sustainability*, 12(6):2570, 2020.

Xu, X. and Yoneda, M. Multitask air-quality prediction based on lstm-autoencoder model. *IEEE transactions on cybernetics*, 51(5):2577–2586, 2019.

Xu, Y., Sha, Y., Wang, C., Cao, W., and Wei, Y. Comparative studies of predictive models for unsteady flow fields based on deep learning and proper orthogonal decomposition. *Ocean Engineering*, 272:113935, 2023.

Yan, X., Zang, Z., Jiang, Y., Shi, W., Guo, Y., Li, D., Zhao, C., and Husi, L. A spatial-temporal interpretable deep learning model for improving interpretability and predictive accuracy of satellite-based pm2.5. *Environmental Pollution*, 273:116459, 2021. ISSN 0269-7491. doi: https://doi.org/10.1016/j.envpol.2021.116459.

Yang, J., Yin, P., Ye, S., Wang, S., Zheng, J., and Ou, J. Marine emission inventory and its temporal and spatial characteristics in the city of shenzhen. *Huan jing ke xue= Huanjing kexue*, 36(4):1217–1226, 2015.

Zang, Z., Guo, Y., Jiang, Y., Zuo, C., Li, D., Shi, W., and Yan, X. Tree-based ensemble deep learning model for spatiotemporal surface ozone (o3) prediction and interpretation. *International Journal of Applied Earth Observation and Geoinformation*, 103:102516, 2021. ISSN 1569-8432. doi: https://doi.org/10.1016/j.jag.2021.102516.

Zaveri, R. A., Easter, R. C., Fast, J. D., and Peters, L. K. Model for simulating aerosol interactions and chemistry (mosaic). *Journal of Geophysical Research: Atmospheres*, 113(D13), 2008.