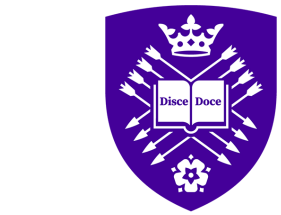

University of Sheffield

Air Pollution Modelling and Source Inference

Nur Izfarwiza Binti Mohd Talib

Supervisor: Michael Smith

A report submitted in fulfilment of the requirements
for the degree of your degree (e.g. BSc in Computer Science)

in the

School of Computer Science

May 14, 2025

Declaration

All sentences or passages quoted in this report from other people's work have been specifically acknowledged by clear cross-referencing to author, work and page(s). Any illustrations that are not the work of the author of this report have been used with the explicit permission of the originator and are specifically acknowledged. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure in this project and the degree examination as a whole.

Name: Nur Izfarwiza Binti Mohd Talib

Signature:

Date: 14 May 2025

Acknowledgement

The first time I heard my supervisor say, "I'm sorry for giving you a hard project," I did not fully understand what he meant. However, as the project progressed, and especially toward the end, I started to feel the weight of that sentence.

I am deeply grateful to my supervisor, Dr. Michael Smith, for his kindness, patience, and constant support. In every supervision meeting, he took the time to explain concepts carefully, never making me feel inadequate for what I did not know. There was never a meeting where he did not draw diagrams or patiently walk me through the ideas. His guidance was invaluable and made an immense difference to my learning experience.

I would also like to thank my father, an air pollution expert, who—despite being on the other side of the world in Malaysia—listened to my weekly calls after each supervisor meeting. Even though he had no background in Computer Science or Gaussian processes, he was always willing to hear me out, offering encouragement when I needed it most.

Thank you to my mother, whose emotional support has carried me through university, and to my friends who stood beside me through the ups and downs of this journey. I am forever grateful for your presence and unwavering belief in me.

Abstract

Inferring pollution sources from observed pollutant concentrations and meteorological fields is a challenging inverse problem. The Advection Gaussian Process (Advection GP) adjoint approach, which combines an adjoint-based solution of the advection–diffusion equation with a Gaussian Process prior over the source distribution, offers robust uncertainty quantification at high spatial resolution[13]. However, it has previously only been tested on synthetic datasets.

This project evaluates the application of the Advection GP framework to realistic data, focusing on Victoria, Australia, during the 2019–20 bushfire season. Satellite-derived Aerosol Optical Depth (AOD) from MERRA-2 was used to approximate $PM_{2.5}$ concentrations, and wind fields were incorporated to model pollutant advection and diffusion. The analysis was geographically constrained to the region of Victoria to study localised dispersion, and vertical transport was approximated by focusing on near-surface wind layers relevant to $PM_{2.5}$ transport.

Results demonstrate that the Advection GP approach can qualitatively recover major pollution source regions when compared with satellite fire anomaly data. While quantitative accuracy remains limited by uncertainties in input data and vertical transport effects, the findings represent an important step toward applying probabilistic source inference methods to real-world environmental events.

Key contributions include the development of data preprocessing workflows, adaptations of the modelling framework for real observations, and new insights into the challenges of vertical representation and pollutant proxy selection.

This study lays the groundwork for future advances in environmental monitoring and air quality management. Future work could extend this methodology to dynamic atmospheric conditions, incorporate real-time remote sensing data, and develop fully three-dimensional models to better capture vertical transport processes.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Air pollution has become an urgent global issue, with significant consequences for public health, ecosystems, and climate change. According to the World Health Organisation, exposure to fine particulate matter ($PM_{2.5}$) contributes to over 4.2 million premature deaths annually [39]. $PM_{2.5}$ refers to tiny airborne particles that can be inhaled, typically measuring 2.5 micrometres or less in diameter [38]. The ability to accurately identify and attribute pollution sources is essential for designing effective mitigation strategies and informing environmental policies. However, this task is highly challenging due to the complexity of atmospheric transport processes, geographical variability, and the limited availability of observational data.

The inverse problem of pollution source attribution—determining unknown emission sources from observed pollutant concentrations—is fundamentally probabilistic due to uncertainties in atmospheric transport, limited measurement coverage, and model errors [31]. Pudykiewicz (1998) demonstrated that even with adjoint methods, source localisation remains constrained by data sparsity—particularly the lack of vertical profiles—and the limited density of monitoring networks, making exact source determination difficult. The adjoint method is a widely used mathematical technique for efficiently computing how changes in model inputs—such as pollutant emission sources—affect outputs like atmospheric concentrations. This makes it particularly suitable for high-dimensional inverse problems governed by physical models such as advection–diffusion equations, where direct computation of sensitivities would be computationally expensive [30].

Studies incorporating chemical data assimilation, such as Arellano et al. (2007), reveal biases in inventory-based emissions estimates, with systematic overpredictions in source regions and underpredictions in remote areas [4]. Traditional inventory-based approaches rely on predefined emissions estimates, yet Kopacz et al. (2009) found that inverse modelling often requires substantial adjustments to these inventories, highlighting their limitations in real-world applications [23]. Deterministic optimisation-based methods attempt to minimise model error but are computationally intensive and lack robust uncertainty quantification, making them impractical for large-scale applications [3].

To address these challenges, this study builds upon the Adjoint-Aided Gaussian Process

(Advection GP) framework developed by Gahungu et al. (2022) [13]. Their work introduced an adjoint-based approach combined with a Gaussian Process (GP) prior, demonstrating that a truncated basis expansion enables exact Bayesian inference while significantly reducing computational costs. A Gaussian Process is a flexible, non-parametric model that defines a distribution over functions, fully specified by a mean function and a covariance function [32]. In the Advection GP framework, this prior is used to represent the unknown pollution source distribution, enabling the model to capture spatial variability and quantify uncertainty in the inferred sources. However, their method was primarily validated using synthetic datasets, and its effectiveness in real-world pollution events remains largely unexplored. This dissertation represents the first application of the Advection GP framework to real environmental satellite data, focusing on pollution events during the 2019–20 Australian bushfire season.

This dissertation extends their framework by applying Gaussian Process-based inverse modelling to pollution source attribution in a real-world setting, specifically focusing on the 2019–20 Victoria bushfires. This study utilises Aerosol Optical Depth (AOD) data from NASA's MERRA-2 reanalysis [16] as a proxy for PM2.5 concentrations. Aerosol Optical Depth (AOD) is a satellite-derived measure of the extinction of solar radiation due to aerosols in the atmosphere, quantifying how much light is absorbed or scattered as it travels vertically through the atmospheric column [29]. It reflects the total aerosol load in the column and is commonly used as a proxy for surface-level air pollution, especially in regions where ground-based monitoring is limited. In addition to AOD data, wind field information is incorporated to model pollutant transport and diffusion processes.

By integrating Gaussian Processes (GPs) with adjoint-based inference, this approach models pollution sources as continuous spatial distributions, capturing variations in source intensity over time and space. Unlike traditional source attribution methods, this framework offers robust uncertainty quantification, enabling source inference even in the presence of sparse or noisy observational data.

Ultimately, this dissertation seeks to demonstrate whether probabilistic source inference models like Advection GP can be effectively transitioned from controlled synthetic environments to complex real-world scenarios — a critical step toward operationalising probabilistic modelling frameworks in air quality management.

1.1 Aims and Objectives

The aim of this project is to test whether the approach by Gahungu et al. (2022)[13] to infer the sources of air pollution using observed atmospheric data, particularly Aerosol Optical Depth (AOD) and wind fields, while evaluating the accuracy of these inferences against qualitative and quantitative assessments of ground truth source data.

Objectives

- **Develop and validate the approach**: Implement forward and backward modelling techniques in Python to test whether the method reliably identifies pollution sources under controlled conditions using synthetic datasets.

- **Apply the AdvectionGP framework to real-world scenarios:** Use real AOD and wind datasets such as those from 2019-20 Australian bushfires, to evaluate the model's ability to infer pollution sources in practical settings
- **Qualitative validation:** Compare the predicted pollution sources against observed thermal anomaly data qualitatively to assess the model's alignment with known sources.

Chapter 2

Literature Review

2.1 Introduction

The attribution of air pollution sources is a critical area of research in atmospheric science, enabling policymakers to design targeted emission control strategies. Mathematical models grounded in atmospheric chemistry and physics are essential for tracing pollutants from their origins, understanding their atmospheric transport, and predicting their impact on air quality at specific locations [34].

However, the inverse problem of pollution source attribution—identifying unknown emission sources based on limited observational data—remains highly challenging. Atmospheric transport processes are complex, meteorological conditions are variable, and observational datasets are often sparse or noisy.

Over the past decades, researchers have proposed a variety of methodologies to tackle this problem, ranging from *deterministic optimisation techniques*, which seek to minimise model–observation discrepancies, to *probabilistic inference models* that explicitly account for uncertainties.

This chapter reviews the evolution of air pollution source attribution methods, focusing particularly on adjoint-based approaches and probabilistic models. It highlights key limitations, such as computational cost and the assumption of static emissions, and motivates the development of adjoint-aided Gaussian Process (GP) frameworks. Addressing these limitations forms the foundation for the real-world application explored in this dissertation.

2.2 Historical Approaches to Air Pollution Source Attribution

2.2.1 Single-Source Identification

Early research on air pollution source attribution focused on identifying *single sources* of contamination, particularly in *emergency scenarios* such as nuclear fallout or industrial accidents. Pudykiewicz (1998) [31] introduced an adjoint-based *tracer transport equation* to estimate the location of released radioisotopes. These early models assumed a *point-source*

emission, which, while effective for specific cases, lacked the ability to model *continuous emissions* across spatially distributed sources. While effective for isolated emission events, such as nuclear accidents or industrial leaks, this approach lacks the flexibility to account for **multiple, spatially distributed sources**.

Another approach to source attribution was through *Bayesian inference methods*, where *a single-source location* was estimated using Markov Chain Monte Carlo (MCMC) sampling [5] to reconstruct the most probable source location given downwind concentration measurements and wind field data. This approach effectively frames source estimation as an **inverse problem**, where the goal is to determine the original emission parameters from limited concentration data. While probabilistic, these methods remained computationally expensive and were often infeasible for *large-scale atmospheric transport* problems.

2.2.2 Multi-Source and Regional Attribution Methods

As research progressed, studies expanded to multi-source attribution, where emissions were distributed across a domain rather than originating from a single point. One common approach was to divide the study area into discrete regions and optimise emissions within each. Arellano et al. (2007) [4] applied this approach during the INTEX-B field mission, evaluating a chemical transport model by incorporating satellite CO observations.

Unlike traditional methods, Arellano et al. (2007) employed ensemble-based data assimilation, integrating real-time observations into the Community Atmosphere Model (CAM3). By assimilating MOPITT CO retrievals, they refined CO distributions and identified biases in existing emission inventories. Their findings revealed systematic biases, particularly:

- Overestimation of CO emissions in source regions (e.g., China).
- Underestimation of transported CO over remote areas (e.g., the Pacific Ocean).

These results underscored the limitations of static regional emission inventories, which often fail to account for transport uncertainties and observational errors. Consequently, more flexible probabilistic approaches—such as Gaussian Processes—have been explored to better quantify uncertainty while avoiding rigid spatial partitioning.

2.2.3 Adjoint-Based Methods for Source Attribution

The introduction of *adjoint models* significantly enhanced the computational efficiency of inverse modelling in atmospheric sciences. These models enable *gradient-based optimisation* by computing the sensitivities of an objective function with respect to all inputs in a *single backward pass*, avoiding the need to repeatedly solve the forward model for each parameter [23].

Adjoint-based approaches have been widely applied in air pollution source attribution:

- Kopacz et al. (2009) [23] compared adjoint-based inversion with analytical Bayesian inversion methods to constrain carbon monoxide emissions in Asia using satellite-derived

MOPITT CO measurements. Their study highlighted trade-offs between computational efficiency and uncertainty representation across the two approaches.

- Yee (2008) [41] developed a probabilistic framework capable of estimating an unknown number of contaminant sources by integrating atmospheric transport models with statistical inference. While not an adjoint method in itself, this work complements adjoint approaches by addressing the challenges of multiple-source reconstruction in turbulent atmospheric conditions.

Despite their computational advantages, traditional adjoint methods typically assume emissions are fixed or temporally averaged over a predefined window. This assumption limits their applicability in scenarios involving rapidly varying sources, such as wildfire emissions or irregular industrial activities. As a result, more flexible modelling frameworks that can accommodate time-varying and spatially distributed emissions are increasingly necessary for realistic pollution source inference.

2.3 Recent Advances in Probabilistic Inference for Air Pollution Attribution

2.3.1 Bayesian Inference and Markov Random Fields

To overcome the limitations of deterministic adjoint-based methods, researchers began incorporating *Bayesian inference* into source attribution models. Hwang et al. (2019) [21] proposed a *Bayesian inverse physics model* with a *Markov random field prior*, allowing neighbouring sources to be spatially correlated. This resulted in smoother and more realistic spatial source distributions. However, their framework did not capture *temporal dynamics*, meaning that sources were assumed to be constant over time.

Albani et al. (2021) [3] extended Bayesian inversion approaches by applying *adaptive MCMC sampling* to better quantify uncertainty in emission estimates. While this improved statistical robustness, MCMC methods remain computationally expensive, especially when applied to large spatial domains or high-dimensional source fields.

2.3.2 Gaussian Process-Based Source Attribution

Recent work has introduced *Gaussian Process (GP) priors* as a powerful alternative for modelling pollution sources. Gaussian Processes offer several advantages:

- They provide a continuous and flexible prior over source distributions, allowing the model to capture spatial variability and provide principled uncertainty estimates.
- They enable structured uncertainty quantification through tractable approximations, such as random Fourier features, which avoid the need for computationally expensive sampling methods like MCMC [13].

- They integrate naturally with adjoint-based solvers for efficient inference.

Gahungu et al. (2022) [13] proposed an *adjoint-aided Gaussian Process inference* approach, where the use of a *truncated basis expansion*—a dimensionality reduction technique—enabled exact Bayesian inference at significantly lower computational cost. This dissertation builds directly on their framework by applying it to real-world atmospheric pollution events, focusing specifically on Australia's 2019–20 bushfire season.

2.4 Focus on Australia

Australia provides an ideal setting for this study due to its seasonal bushfires, which release significant amounts of pollutants into the atmosphere. These events are particularly suitable for analysing pollutant dispersion because:

- **Known Source Locations:** Bushfire locations can be accurately identified through remote sensing, offering clear, observable ground truth data for validation of the model[8].
- **High AOD Data Coverage:** Australia experiences relatively low cloud cover during the bushfire season, ensuring consistent and reliable Aerosol Optical Depth (AOD) measurements for analysis.[16]
- **Scale of Pollution and Remote Sensing Compatibility:** The spatial scale of pollutant sources and their dispersion fits well with the resolution and coverage provided by satellite remote sensing data.

Although trends such as increasing smoke AOD levels in Northwestern Australia have been reported [26], this study does not aim to model long-term aerosol trends. Instead, the 2019–20 bushfire season is selected because it provides a clear, well-documented pollution event that enables testing of the Advection GP inference framework using real data. The increased frequency of bushfires in arid regions, such as Western Australia—where fire activity rose by approximately 40% between 2008 and 2013 [33]—and occasional smoke transport from nearby regions like Indonesia further illustrate the relevance of biomass burning as a source of pollution. However, these factors serve to support the selection of a representative case study, not to establish regional trends..

The availability of precise forest fire data during this period allows for model validation against known source locations. This makes the 2019–20 season a practical and well-suited case study to evaluate the performance of the pollution source inference model under realistic conditions, rather than a vehicle for exploring broader environmental dynamics.

2.5 Atmospheric Transport Mechanisms

Understanding how pollutants move through the atmosphere is essential for accurate source attribution. Recent research highlights the complex behaviour of aerosol transport, particularly

for anthropogenic sources. According to Jacob (2000), the phase state of aerosols—whether aqueous or solid—plays a critical role in their chemical reactivity, especially in the lower troposphere. In ozone modelling, it is often assumed that aerosols are aqueous due to the high relative humidity and the energy barrier for efflorescence, which hinders their transition to a dry phase [22].

Heald et al. (2006) found that the transpacific transport of anthropogenic aerosols occurs predominantly within the lower free troposphere, specifically between 900–700 hPa. This contrasts with the transport behaviour of carbon monoxide (CO), which Liang et al. (2004) showed to be primarily concentrated within the boundary layer. These differences illustrate how various pollutants follow distinct transport pathways depending on their chemical properties and atmospheric interactions [19, 25].

Vertical transport within the atmospheric boundary layer also plays a key role in pollutant distribution. A study conducted in Beijing by Ding et al. (2005) measured $PM_{2.5}$ and PM_{10} concentrations at altitudes of 100 m, 200 m, and 320 m, finding that concentrations decreased gradually with height. This pattern reflects limited vertical mixing and suggests pollutants remain relatively concentrated near their sources in urban areas. Their study also emphasized that vertical turbulence and boundary layer dynamics are closely linked to the distribution of aerosols [10].

Although these urban-based findings differ from rural or wildfire scenarios, they offer valuable insight into how vertical stratification influences pollutant behaviour. In contrast to stable, low-altitude emissions from traffic or factories, bushfire smoke often penetrates higher atmospheric layers, including the lower free troposphere. This allows for widespread regional and even intercontinental transport, particularly under strong convective conditions.

Drawing on these insights, it becomes clear that selecting the appropriate atmospheric height is critical for modelling pollution transport. The boundary layer is most relevant for capturing local and short-range effects, while the lower free troposphere better represents long-range transport processes. In this study, the focus is placed on the lower free troposphere to more effectively capture the dispersion of bushfire smoke across Australia and evaluate its regional transport dynamics.

Chapter 3

Requirements and Analysis

3.1 Project Overview

This project investigates the viability of inferring air pollution sources by leveraging the AdvectionGP framework, which models pollutant transport using observed wind fields and AOD data. The overall approach is grounded in the work of Gahungu et al.[13], with the goal of evaluating how effectively the methodology generalises to real-world atmospheric conditions, focusing on the 2019–20 bushfire season in Victoria, Australia as a representative case study.

To support this goal, the project incorporates several subcomponents:

- A wind model that retrieves and processes random points on physical domain to return wind speed from NASA's MERRA-2 dataset.
- A sensor model capable of simulating remote sensing observations, by generating particles on polygons across the physical domain for particle-based backward inference.
- Integration of real AOD observations to replace synthetic concentration data in the inference pipeline.
- Conversion of AOD values to $PM_{2.5}$ concentrations to provide pollution observations for source inference.
- Evaluation through visual and geographical validation against known bushfire ignition points and air quality measurements.

While some of the following content overlaps with system design decisions, these elements are included here to provide analytical justification for the coordinate system, vertical layer selection, and spatial-temporal modelling parameters used in this study. The goal is to outline how these considerations, along with the functional and non-functional requirements that follow shape the downstream implementation, which is detailed in Chapter 4.

3.2 Data and Input Parameters

3.2.1 Datasets Used

Meteorological Data: MERRA-2 for Wind Data

This study employs NASA's MERRA-2 dataset, specifically the `tavg3_3d_asm_Nv` (M2T3NVASM) collection provided by NASA's GMAO and hosted on GES DISC [17]. This is relevant to pollution transport and analysis atmospheric conditions. MERRA-2 is a modern reanalysis product generated by the NASA Global Modeling and Assimilation Office (GMAO) using the Goddard Earth Observing System Model (GEOS) version 5.12.4. It provides a high-resolution, assimilated dataset covering the satellite era from 1980 to the present, ensuring comprehensive temporal and spatial coverage. [6]

The M2T3NVASM dataset includes time-averaged, three-hourly meteorological fields across 72 vertical model layers. Key assimilated variables include layer height and wind components of both eastward wind and northward wind. These parameters are critical for modelling pollutant dispersion dynamics. The dataset's temporal resolution (three-hourly) and global spatial extent make it suitable for analysing the transport and diffusion of pollutants over Australia.[6]

To capture the impact of bushfire emissions, bushfire occurrence data and emissions inventories were cross-referenced with this atmospheric dataset.

Aerosol Optical Depth (AOD) Data: MERRA-2 inst3_2d_gas_Nx

This study uses the `MERRA-2 inst3_2d_gas_Nx` (M2I3NXGAS) dataset, which provides three-hourly, single-level AOD values from NASA's MERRA-2 reanalysis system [5]. AOD acts as a proxy for $PM_{2.5}$ concentration but has limitations—it cannot distinguish between aerosol types and is influenced by factors such as aerosol height, humidity, and mixing layer properties [24].

Despite these constraints, machine learning models like Random Forests can capture the AOD–$PM_{2.5}$ relationship effectively when combined with meteorological and land use data [37, 18].

Thanks to its high spatial and temporal resolution, M2I3NXGAS remains a valuable resource for studying aerosol transport during events like the 2019–20 Australian bushfire season.

3.2.2 Key Variables

The key variables used in this study are drawn from the MERRA-2 datasets and include spatial, temporal, and meteorological components:

- **Longitude (`lon`):** Measured in degrees east, representing the east-west spatial coordinate.
- **Latitude (`lat`):** Measured in degrees north, representing the north-south spatial coordinate.

- **Vertical Level (`lev`):** Atmospheric layers in the MERRA-2 meteorological dataset.
- **Time (`time`):** Measured in minutes since 2024-06-01 01:30:00, representing the temporal dimension of the data.
- **Eastward Wind (`U`):** Measured in meters per second (`m s`$^{-1}$), representing the component of wind blowing from west to east.
- **Northward Wind (`V`):** Measured in meters per second (`m s`$^{-1}$), representing the component of wind blowing from south to north.
- **Aerosol Optical Depth (`AODANA`):** Instantaneous assimilated values representing the intensity of atmospheric aerosols, used as a measure of particulate matter concentration.

3.2.3 Data Extraction

A combination of automated and manual data retrieval methods was considered for accessing MERRA-2 NetCDF files. Due to challenges with NASA's API authentication and download consistency, a manual approach was adopted. See Appendix .1 for details.

3.3 Physical Domain Considerations and Preprocessing

To adapt the AdvectionGP framework for real-world application, it is essential to establish a consistent and well-justified foundation for spatial and temporal modelling. This section outlines the key analytical considerations for data preprocessing, spatial parametrisation, and temporal resolution—each of which plays a critical role in enabling regional pollution source inference. These considerations inform and constrain the design decisions detailed in Chapter 4 and are consistently applied across the wind model, sensor model, and inference pipeline. In particular, the choice of coordinate system and map projection ensures alignment between data sources and modelling components throughout the system.

3.3.1 Coordinate System and Projections

This study represents particle positions and pollution values in a two-dimensional domain, requiring a suitable map projection to minimise distortion and ensure spatial consistency. Since wind speed is given in metres per second ($\mathrm{m\,s^{-1}}$), a projection that preserves distance is crucial for accurate simulation.

Latitude and longitude define points on Earth's curved surface but do not convey planar distances. To support numerical computation and visualisation, a projection that maintains local accuracy is essential. A full comparison of candidate projections is provided in Appendix .2.

Selected Projection: UTM Zone 56S UTM Zone 56S (EPSG:32756) was selected using the EPSG Geodetic Parameter Dataset [1, 2] for its ability to preserve distance and scale in eastern Australia—particularly Victoria, which was heavily affected during the 2019–20 bushfires. Figure 17 shows the coverage of this projection.

Figure 3.1: UTM Zone 56S projection covering the eastern regions of Australia, generated using EPSG's visualisation tool powered by MapTiler [2].

3.3.2 Vertical Height Selection

The analysis initially focused on the lower free troposphere, specifically at altitudes corresponding to 900–700 hPa, based on findings from the literature reviewed in Chapter 2. This altitude range is relevant for capturing long-range pollutant transport mechanisms, such as bushfire smoke dispersion across Australia.

In NASA's MERRA-2 dataset [17], vertical height is represented as discrete pressure levels [6]. Layers 56 to 67 were selected to represent this range and were planned to be averaged into a single vertical slice to simplify the modelling process. This would allow the simulation to operate in two spatial dimensions (x, y) while retaining the dominant transport features of the lower free troposphere.

However, due to changes in model design and practical constraints during implementation (see Chapter 6, Limitations), the final system was tested using wind averaged over a different vertical range. This section documents the initial design decisions that informed earlier data preprocessing and modelling considerations.

Table 4.2 Products on the native vertical grid will be output on the following levels. Pressures are nominal for a 1000 hPa surface pressure and refer to the top edge of the layer. Note that the bottom layer has a nominal thickness of 15 hPa.

Level	P(hPa)	Lev	P(hPa)	Lev	P(hPa)	Lev	P(hPa)	Lev	P(hPa)	Lev	P(hPa)
1	0.0100	13	0.6168	25	9.2929	37	78.5123	49	450.000	61	820.000
2	0.0200	14	0.7951	26	11.2769	38	92.3657	50	487.500	62	835.000
3	0.0327	15	1.0194	27	13.6434	39	108.663	51	525.000	63	850.000
4	0.0476	16	1.3005	28	16.4571	40	127.837	52	562.500	64	865.000
5	0.0660	17	1.6508	29	19.7916	41	150.393	53	600.000	65	880.000
6	0.0893	18	2.0850	30	23.7304	42	176.930	54	637.500	66	895.000
7	0.1197	19	2.6202	31	28.3678	43	208.152	55	675.000	67	910.000
8	0.1595	20	3.2764	32	33.8100	44	244.875	56	700.000	68	925.000
9	0.2113	21	4.0766	33	40.1754	45	288.083	57	725.000	69	940.000
10	0.2785	22	5.0468	34	47.6439	46	337.500	58	750.000	70	955.000
11	0.3650	23	6.2168	35	56.3879	47	375.000	59	775.000	71	970.000
12	0.4758	24	7.6198	36	66.6034	48	412.500	60	800.000	72	985.000

Figure 3.2: The vertical grid on following levels in NASA's data [6]

3.4 Spatial and Temporal Scale Definition

3.4.1 Bounding Box

The spatial extent of the analysis was initially defined by a wide bounding box over eastern Australia, aimed at capturing key bushfire-prone regions during the 2019–20 fire season. However, the final focus was narrowed to the state of **Victoria**. This refinement was motivated by both practical considerations and data availability:

- **Data Availability:** Victoria offered the richest and most reliable public datasets during the bushfire period, including shapefiles of bushfire ignition points, hourly $PM_{2.5}$ measurements from EPA stations, and high-resolution satellite AOD data.
- **Event Significance:** East Gippsland, Victoria, was one of the regions most severely impacted during the 2019–20 bushfire season, making it a valuable case study for pollution source inference.
- **Computational Efficiency:** Focusing on a smaller region reduced data size and computational load while preserving relevance and depth of analysis.

This geographic focus aligns with the availability of high-quality datasets published through the **DataVic** portal [36], including historical fire origin points [8] and EPA Victoria's hourly $PM_{2.5}$ measurements [12]. These datasets are later used for validation and evaluation.

This spatial focus enabled the model to integrate real observational data from Victoria while maintaining tractability for wind preprocessing and inference. The bounding box was applied consistently to subset wind, AOD, and observational data.

3.4.2 Temporal Scale

The temporal resolution of this project is derived from the MERRA-2 dataset, which provides meteorological variables at 3-hour intervals. Each NetCDF file includes eight time steps per day, recorded as minutes since a reference time (`2019-10-01 01:30:00 UTC`). This regular spacing allows for consistent temporal sampling across the dataset.

3.5 Project Requirements

The system is designed to infer pollution sources using satellite-based observations and meteorological data. The shading in the table indicates the priority: red for mandatory requirements, while yellow for desirable (optional) requirements. The following functional components are required to meet this objective.

3.5.1 Wind Model Requirements

Table 3.1: Functional Requirements for Wind Model

ID	Requirement	Description
1	Bounding Box Selection	Define the spatial limits of the analysis area using lat/lon coordinates and convert to UTM.
2	Vertical Layer Selection	Select MERRA-2 pressure levels (e.g., layers 56–67) to represent long-range atmospheric transport.
3	Wind Lookup Table	Preprocess and store wind data in a lookup structure (KD-tree or grid) for efficient retrieval.
4	Wind Lookup Function	Implement `getwind()` to return east/north wind vectors from spatial-temporal coordinates.
5	Tensor Shape Compliance	Ensure `getwind()` returns tensor with shape `(Nparticles, Nobs, 2)` for model compatibility.
6	Interpolation Between Space and Time	Enable linear interpolation between spatial locations and temporal steps to improve wind estimation accuracy. This provides smoother wind transitions but increases computational complexity.

3.5.2 Sensor Model Requirements

Table 3.2: Functional Requirements for the Sensor Model

ID	Requirement	Description
1	Grid Polygon Generation	Convert real latitude and longitude into projected UTM coordinates and generate sensor grid polygons for particle placement.
2	Particle Generation	Generate particles randomly within each polygon using a Poisson distribution to simulate sensor observations.
3	Coordinate Reprojection	Support accurate reprojection of particle coordinates from UTM back to lat/lon for map-based visualisation.
4	Logical Particle Grouping	Ensure particles remain logically grouped by polygon for consistency and clearer visual differentiation on the map.

3.5.3 Physical Domain and Observation Requirements

Table 3.3: Physical Domain and Observation Requirements

ID	Requirement	Description
1	Bounding Area Selection	Select Victoria as the spatial study domain due to the availability of high-quality fire, air quality, and satellite data relevant to the 2019–20 bushfire season.
2	AOD Observation Integration	Replace synthetic concentration data with real AOD values from MERRA-2 to simulate realistic pollution fields for inference.

3.5.4 AOD to PM2.5 Conversion Requirements

Table 3.4: Pollution Observation Generation Requirements

ID	Requirement	Description
1	Realistic Pollution Mocking	Replace synthetic values with real AOD observations to simulate realistic pollution concentration fields.
2	Wind-Aligned Observation	Ensure that pollution observations are compatible with and influenced by wind data for coherent simulation.
3	Visualisation Support	Visualise AOD or $PM_{2.5}$ fields to support spatial interpretation of pollution data during analysis.
4	AOD to $PM_{2.5}$ Conversion	Use a Random Forest regression model to convert AOD values into estimated $PM_{2.5}$ concentrations.
5	Bushfire Accuracy Alignment	Ensure inferred pollution sources spatially align with known bushfire locations and are extracted from appropriate vertical height layers (e.g., 900–700 hPa).

3.5.5 Non-Functional Requirements

Table 3.5: Non-Functional Requirements

ID	Requirement	Description
1	Projection Accuracy	Ensure that UTM Zone 56S maintains accurate local scale and distance fidelity when projecting geospatial coordinates.
2	Wind Model Performance	The wind model must offer a reasonable trade-off between runtime efficiency and spatial-temporal accuracy in wind vector retrieval.
3	Robustness	The system should be resilient to missing or undefined data (e.g., NaN values), avoiding runtime crashes during large-scale simulation.
4	Extensibility	The architecture should support easy integration of alternative wind sources, vertical heights, or sensor configurations for future use.

3.6 Evaluation Plan

The aim of this evaluation plan is to determine the effectiveness and validity of the pollution source inference framework. As this study transitions from synthetic to real data, several strategies are proposed to assess the model's realism and geographical accuracy.

3.6.1 Evaluation Objectives

The evaluation will focus on the following core aspects:

- **Geographical Accuracy:** Ensure that the inferred pollution source (represented by the source mean) spatially aligns with known bushfire regions in Victoria. This involves visualising the estimated source distribution using Cartopy and verifying that it lies within the state boundary.
- **Particle Trajectory Consistency:** Assess whether particle movements projected by the model follow realistic wind-driven transport paths. Cartopy visualisations will be used to overlay particle paths on geographical maps to check their coherence with wind flow patterns.
- **Qualitative Comparison with Real Fire Locations:** Overlay inferred sources onto real bushfire ignition points provided by Victoria's fire origins shapefile dataset [8]. This comparison will support visual validation of whether the model reasonably captures the spatial characteristics of known fire events. Visualisation will be performed using Cartopy throughout the evaluation process.

3.6.2 Limitations and Scope

At this stage, the evaluation remains qualitative due to the lack of precise ground truth data on pollution source strength and emission timing. While fire ignition points from official datasets provide valuable spatial references, source inference is inherently a spatial distribution problem, not a point estimation task. Thus, a direct comparison between inferred results and ignition point data must be interpreted with caution.

Additionally, the model currently uses a two-dimensional wind field, averaging horizontal wind components over a fixed vertical layer. This simplification omits the full complexity of atmospheric mixing, including vertical advection and turbulence, which may influence pollutant dispersion in real conditions. This has implications for how well the model can capture three-dimensional dynamics.

Furthermore, although satellite-derived AOD serves as a proxy for particulate matter concentration, it is limited by its nature as a top-down, column-integrated measurement. Unlike surface-level $PM_{2.5}$ sensors, AOD lacks specificity in vertical distribution and does not directly capture ground-level pollution intensity. This restricts the capacity for high-fidelity quantitative validation and represents a broader challenge in remote sensing-based source attribution.

Chapter 4

Design

4.1 Use of the AdvectionGP Framework

This project builds upon the open-source `AdvectionGP` framework developed by the Machine Learning group at the University of Sheffield (SheffieldML) [35]. The framework provides a modular system for simulating and inferring pollution sources using Gaussian processes under advection-diffusion dynamics. It includes core modules for wind modeling, sensor definition, kernel design, and model inference.The existing codebase, while robust, is primarily designed to operate on synthetic data over abstract coordinate systems. Specifically, it does not support real-world wind fields or the projection of particle movements onto geographical maps. Wind data in the original implementation is generated using simplified functions rather than extracted from real meteorological datasets. Furthermore, particles are not geospatially grounded (i.e., not mapped to Earth's surface).

Motivation for Extension

This project extends the `AdvectionGP` framework to work with real-world data by:

- Integrating real wind fields from NASA's MERRA-2 dataset [17].
- Supporting coordinate transformations and geographic projections (e.g., UTM Zone 56S) .
- Enabling the use of satellite-derived Aerosol Optical Depth (AOD) data [16] as pollution observations, which are subsequently converted to $PM_{2.5}$.

Repository Structure and Integration Overview

The original AdvectionGP repository structure was preserved for compatibility. Key components extended for this project include:

- `models/mfmodel.py`: Core inference logic via `MeshFreeAdjointAdvectionDiffusionModel`, selected for its flexibility with irregular, real-world data.

- `wind.py`: Real meteorological wind data was integrated by implementing multiple `RealWind` variants, replacing the framework's synthetic wind. See Section 4.3.
- `sensors.py`: A custom `RemoteSensingModel` was added to simulate satellite-style AOD observations through particle generation over spatial polygons (see Section 4.4).
- `kernels.py`: Existing kernel implementations, such as Exponentiated Quadratic (EQ) `EQ`, were reused to define spatial basis functions for Gaussian Process inference.

Table 4.1 summarises the modified components.

File / Folder	Description
`models/`	Inference logic (`MeshFreeAdjointAdvectionDiffusionModel`)
`wind.py`	Custom `RealWind` classes for real wind data
`sensors.py`	`RemoteSensingModel` for particle-based observation simulation
`kernels.py`	EQ kernel for GP basis function generation

Table 4.1: Modified files from the AdvectionGP framework

4.2 System Architecture and Inference Pipeline

This project adopts the `MeshFreeAdjointAdvectionDiffusionModel`, a particle-based inference system provided by the AdvectionGP framework, to simulate and trace the origin of air pollution observed during the 2019–20 bushfire season in Victoria, Australia. The model integrates Gaussian Processes (GPs), adjoint simulation techniques, and real meteorological data to support source attribution over complex spatial domains.

4.2.1 Model Pipeline and Execution Flow

The pollution source inference pipeline designed for this project is based on the `MeshFreeAdjointAdvectionDif` `(mfmodel)` provided by the AdvectionGP framework. The full inference process is structured into three main stages: **Inputs and Setup**, **Model and Inference Steps**, and **Validations and Outputs**, as illustrated in Figure 4.1.

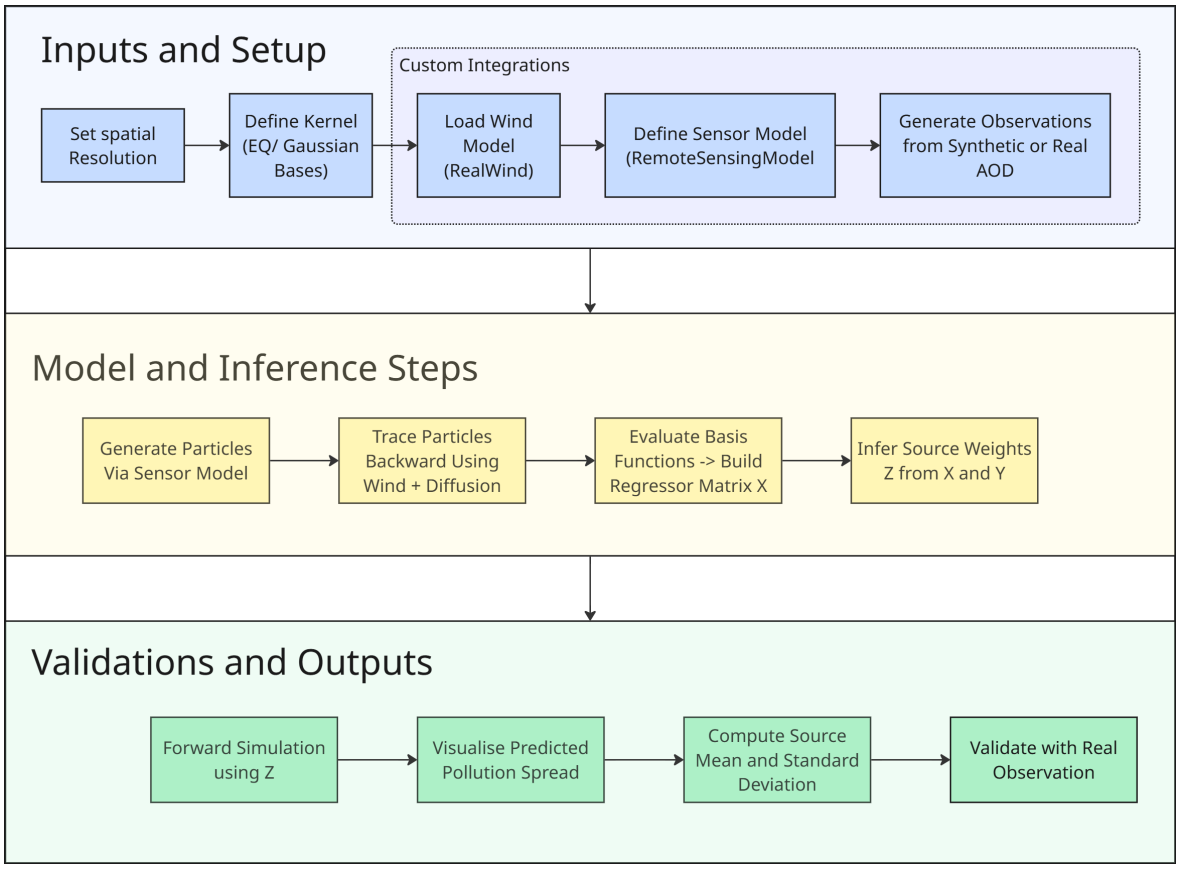

Figure 4.1: Flow diagram illustrating the full model pipeline using the `mfmodel`, including custom integrations of the `AdvectionGP` framework such as the wind model, sensor model, and real pollution observations.

Stage 1: Inputs and Setup

This initial stage defines all required model parameters and data sources. The function that is implemented in this project is the wind model, the sensor model and the pollution observation and the remainder was pre-implemented in the base framework. The parameters were selected based on exploratory testing and are critical to the performance and robustness of the model's inference.

- **Set Spatial Resolution**: The temporal and spatial resolution of the domain is defined here, e.g. `[T, X, Y] = [120, 10, 10]`, which determines the number of basis functions used in each dimension.
- **Boundary Conditions**: The simulation domain is constrained by a four-dimensional bounding box:

$$\texttt{boundary} = [(t_{\min}, x_{\min}, y_{\min}),\ (t_{\max}, x_{\max}, y_{\max})]$$

This defines the spatial and temporal extent of the model. Both particle trajectories and kernel basis functions are restricted to this domain, ensuring that simulation and inference are performed only within the region of interest.

- **Define Kernel**: The kernel determines how basis functions are distributed over the domain. Two types are available in the AdvectionGP codebase:
  - `EQ` (Exponentiated Quadratic): Generates smooth, global basis functions using random Fourier features. The EQ kernel was chosen for its ability to capture smooth, continuous variations across space, suitable for gradual pollution dispersion patterns.
- **Load Wind Model**: The wind model (e.g. `RealWind`) provides wind vectors for each timestep and location. It influences particle trajectories during both backward and forward simulations.
- **Define Sensor Model**: The sensor model (e.g. `RemoteSensingModel`) defines how and where synthetic particles are placed based on the observation grid (or real satellite data coverage).
- **Generate Observations (Y)**: Pollution concentrations are simulated from our main dataset for pollution which is AOD [16] which then would be converted into $PM_{2.5}$
- **Hyperparameters:**
  - **Prior Variance** (`k_0`): Governs the strength of the prior belief over the source weights Z. A small value implies a strong prior (more regularisation), while a large value gives more flexibility to match the data.
  - **Noise Standard Deviation** (`noiseSD`): Captures measurement noise or uncertainty in the pollution observations Y. This is important because AOD measurements (or synthetic proxies) may be affected by errors, sensor limitations, or preprocessing artefacts.
  - **Number of Basis Features** (`N_feat = 1000`): Determines the number of basis functions used to approximate the pollution source. A higher number increases expressiveness but also computational cost.

Stage 2: Model Execution and Inference

This section performs the core Gaussian Process inference to estimate the pollution source distribution:

- **Generate Particles via Sensor Model**: For each observation point, multiple particles are sampled and placed on the sensor grid, each associated with a timestamp.

- **Trace Particles Backward Using Wind + Diffusion**: Particles are propagated backward in time according to wind dynamics (`getwind()`) and random diffusion, simulating possible origins of the observed pollution.
- **Evaluate Basis Functions to Build Regressor Matrix** X: As particles travel, they are evaluated against the basis functions defined by the kernel. This generates the regressor matrix X, which represents the influence of each possible source location on the observations.
- **Infer Source Weights** Z: A posterior distribution over the source weights Z is computed by solving a linear Gaussian model $Y \approx X^T Z$. The resulting mean Z represents the estimated source strength over time and space.

Stage 3: Validations and Outputs

Once the source has been inferred, the model simulates how pollution spreads forward in time:

- **Forward Simulation using** Z: The inferred source is propagated forward through the same wind model, simulating expected pollutant concentrations at later time steps.
- **Compute the Source Mean using** Z: Combining the inferred coefficients with the model's basis functions to compute the source mean across the domain.
- **Visualise Predicted Pollution Spread**: The resulting concentration fields can be visualised over a spatial map. These can be presented side by side with real AOD measurements and historical fire ignition points in Victoria [8] for qualitative comparison.
- **Validate with Real Observations**: Model outputs can be compared against actual AOD values [16], wind vector directions, and bushfire start locations [8] to assess the plausibility and accuracy of the inferred sources.

4.3 Wind Model Design

Wind modelling plays a critical role in the AdvectionGP pipeline, as it governs the movement of particles through space over time. This directly affects how pollution is inferred and how the results are simulated. The custom wind model implemented in this project — RealWind — replaces the synthetic wind used in the original framework with real meteorological data from NASA's MERRA-2 dataset.

The wind model is invoked in two core stages of the pipeline:

- **Backward Propagation during Particle Generation (Adjoint Inference)**: Determines the trajectory of particles traced backward in time from observation points to estimate possible source origins.

- **Forward Simulation of Pollution Spread**: Simulates how the inferred pollution source propagates forward through the atmosphere to produce concentration predictions.

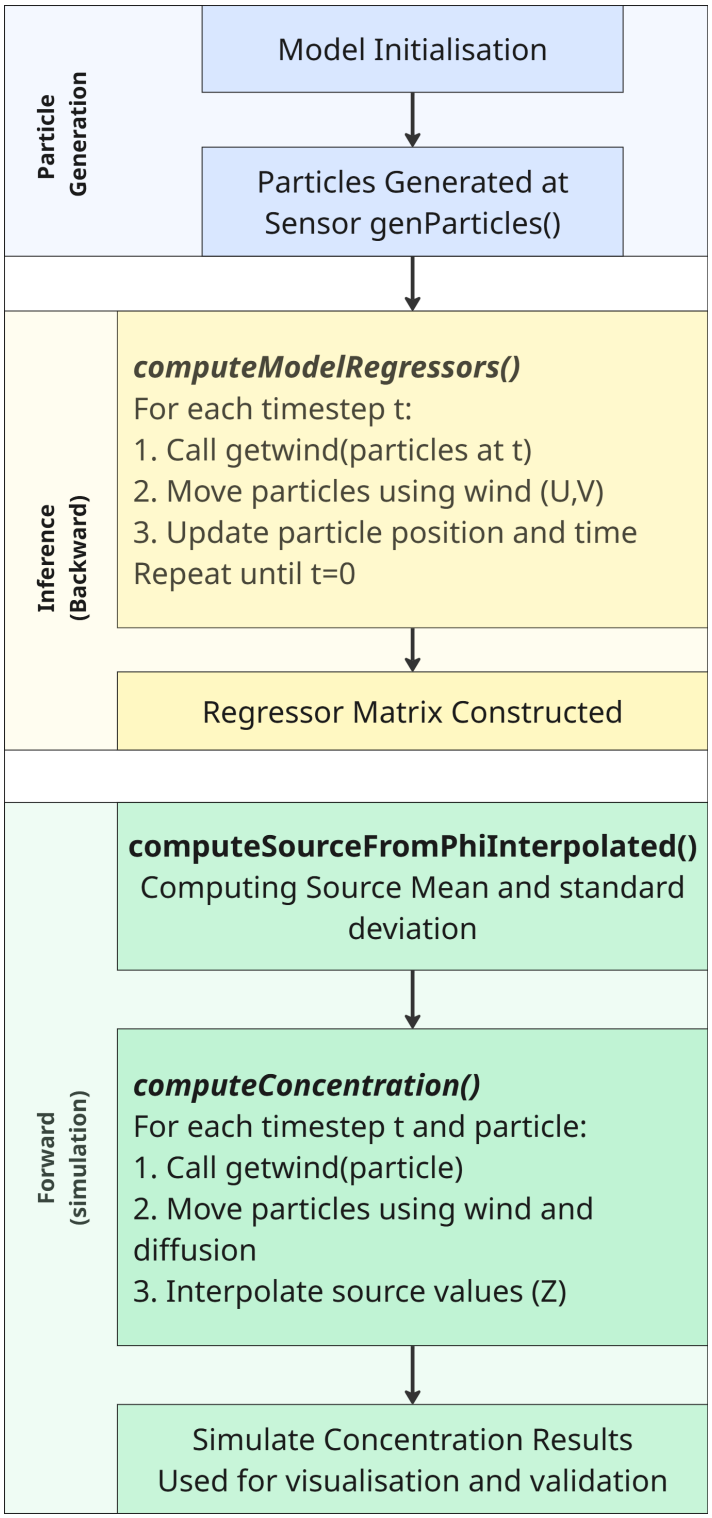

Figure 4.2: Flow diagram showing the role of the wind model in the AdvectionGP pipeline. The `getwind()` function is called repeatedly in both backward inference (`computeModelRegressors`) and forward simulation (`computeConcentration`) to update particle positions based on real wind data.

As shown in Figure 4.2, getwind() is the key method responsible for querying preprocessed wind values at a given particle's spatiotemporal coordinates. Internally, the RealWind class constructs a spatial lookup table and aligns temporal indices to ensure efficient and accurate wind vector retrieval.

This design enables the model to reflect realistic atmospheric transport patterns observed during the 2019–20 bushfire season in Australia, thereby improving source inference accuracy when applied to AOD data.

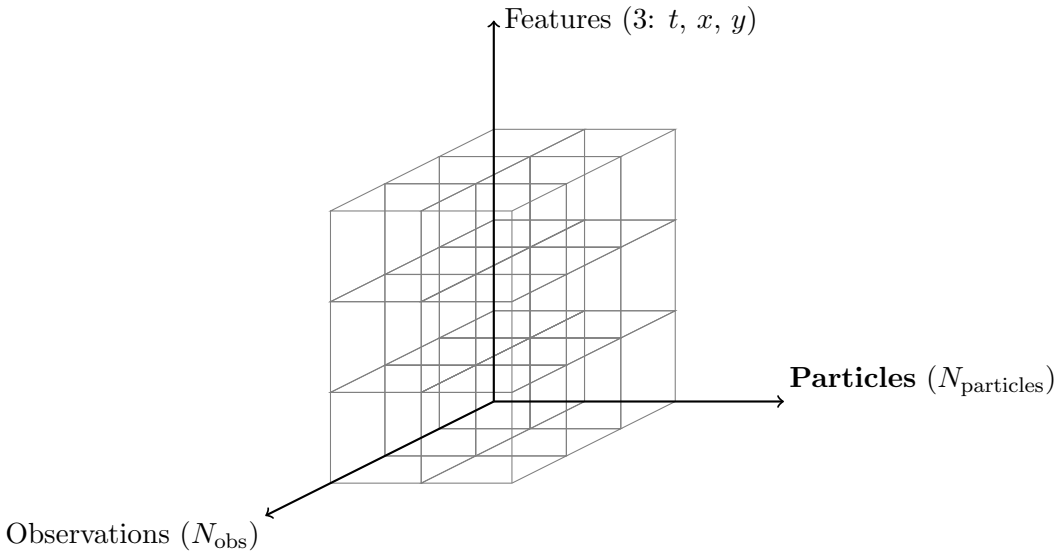

Figure 4.3: Tensor representation of shape $(N_{\text{particles}}, N_{\text{obs}}, 3)$ used for wind model input. Each cube encodes wind query coordinates where t represents time,x represents eastings and y represents northings

represents

Particle i / Obs j	t	x	y
(0,0)	900	354321	5790000
(0,1)	930	354900	5790100
(1,0)	901	354410	5790050
⋮	⋮	⋮	⋮

Figure 4.4: Example of wind query tensor with shape $(N_{particles}, N_{obs}, 3)$. Each row represents one spatiotemporal query.

4.3.1 Data Source: NASA MERRA-2 Wind Fields

This project replaces the synthetic wind fields of the original AdvectionGP framework with real wind data from NASA's MERRA-2 `tavg3_3d_asm_Nv` dataset [17], which provides a global 3-hour reanalysis of wind at 72 vertical levels.

Key variables used include longitude, latitude, vertical layer index, time, and eastward (`U`) and northward (`V`) wind components. Wind vectors were averaged over levels 56–67 (approx. 900–700 hPa) to represent the lower free troposphere—identified in Chapter 2 as crucial for long-range aerosol transport.

Data was accessed via NetCDF using `xarray` and `netCDF4`, with KD-tree methods enabling efficient spatiotemporal wind lookup during particle tracing. Incorporating real wind improves the simulation's realism and alignment with bushfire-period atmospheric conditions in Victoria.

4.3.2 Challenges in Wind Model Lookup

Wind data plays a pivotal role in the AdvectionGP framework, as particle trajectories—both during backward source inference and forward pollution simulation—are directly influenced by wind speed and direction. However, while NASA's MERRA-2 dataset provides meteorological fields sampled on a regular 1° grid at 3-hour intervals, particles in this project are placed randomly over a polygon representing the observation domain (e.g., AOD coverage). This introduces a spatial-temporal mismatch between the locations where wind is known and where wind information is required.

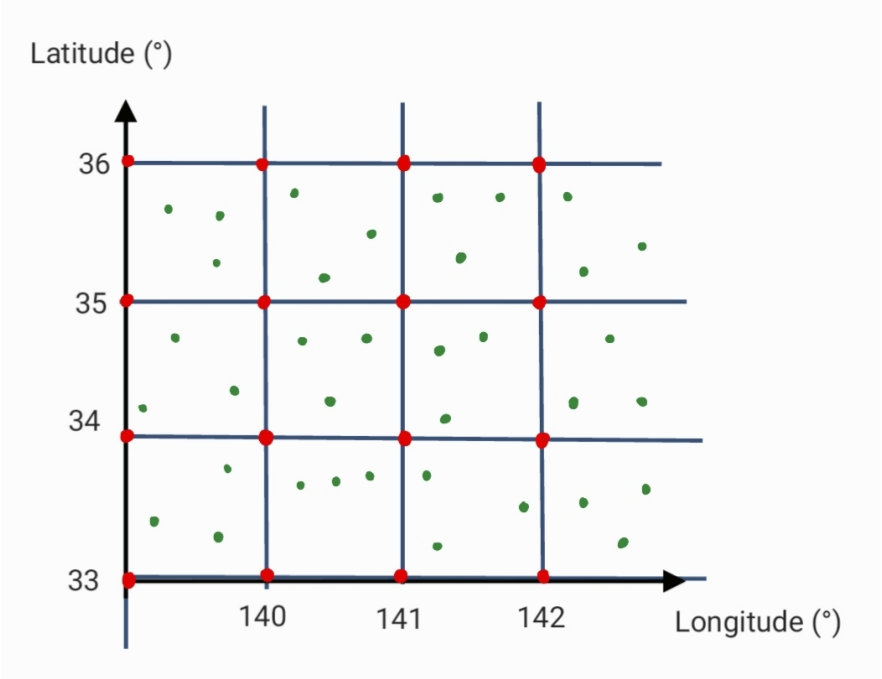

Figure 4.5: The red dots represent wind grid points where eastward and northward wind components can be accurately retrieved from NASA data. In contrast, the green dots represent particle positions, which are randomly generated and do not perfectly align with the wind grid. As a result, their wind values must be estimated either through interpolation or nearest-neighbour assignment.

To address this mismatch, a lookup mechanism is required that can match arbitrary

`(time, easting, northing)` coordinates of particles to the closest or most representative wind vector `(U, V)`. Due to the large number of particles generated during each simulation (often in the thousands), this lookup process must be both computationally efficient and reasonably accurate.

The following challenges shaped the wind model design:

- **Irregular Particle Placement**: Unlike the wind grid, particle locations are continuous and irregular, necessitating interpolation or nearest-neighbour approximation.
- **High Volume of Queries**: Each particle may require wind lookup at multiple time steps, resulting in tens of thousands of queries per simulation run.
- **Trade-off between Accuracy and Speed**: Full interpolation offers high accuracy but is computationally expensive; nearest-neighbour approaches are faster but approximate.

4.3.3 Wind Lookup Strategy: Accuracy vs. Efficiency

Given the high volume of particles and the irregularity of their spatial-temporal positions, designing an efficient wind lookup mechanism was a key consideration. The NASA MERRA-2 wind data is provided on a regular 1-degree grid and sampled every 3 hours. However, particles are generated randomly across time and space, requiring wind values at arbitrary coordinates.

Two main priorities influenced the lookup design:

- **Accuracy:** Interpolation in time and space can improve physical realism but may increase computational complexity.
- **Efficiency:** Approximate methods such as nearest-neighbour lookup or direct grid indexing enable faster simulations but may reduce precision.

Rather than committing to a single method at the design stage, the approach taken was to remain flexible and support multiple implementations. The final selection was deferred until Chapter 5, where methods are benchmarked and evaluated based on compatibility with the inference pipeline and runtime performance. This will involve measuring both the runtime and the deviation in inferred source locations using each method. Such evaluation will help justify the final design decision based on empirical results.

4.3.4 Robustness and Configurability

The wind model was designed with configurable parameters—such as `start_date`, `num_days`, `layer_range`, and bounding box—allowing flexible reuse. Users can easily adjust the temporal range, vertical layers, or spatial extent to:

- Analyse different bushfire events or time periods,
- Test the impact of varying atmospheric heights,

- Modify the study region as needed.

This adaptability supports broader experimentation without changing core code.

4.3.5 Planned Testing

To ensure the wind model was functioning as expected, planned tests included:

- Visualising particle movement using Cartopy to confirm that trajectories reflect plausible wind dynamics.
- Comparing particle flow under different wind model variants (e.g., Nearest Neighbour, Grid Indexing) to assess consistency and completeness (e.g., no missing wind values).
- Benchmarking runtime performance across models using a fixed number of particles and observations to evaluate computational feasibility.

These tests aimed to verify both physical realism and integration compatibility within the full inference pipeline.

4.4 Sensor Design and Particle Generation

The sensor model is a critical component of the source inference pipeline. In this project, a custom `RemoteSensingModel` was developed to simulate satellite-based remote sensing measurements, particularly those derived from Aerosol Optical Depth (AOD) [16] observations.

Unlike traditional ground-based sensors that collect discrete point measurements, satellite data provides spatially aggregated readings over broader areas. To reflect this, the model generates synthetic observations over a grid of polygons covering the region of interest.

This design supports realistic simulation of remote sensing data by enabling:

- **Spatial Sampling:** Particles are randomly distributed within each grid cell using a Poisson point process, mimicking the irregular but widespread coverage of satellite sensors.
- **Temporal Flexibility:** Each particle is assigned a timestamp sampled from a user-defined interval, ensuring compatibility with the time resolution of wind and AOD datasets.
- **Mesh-Free Compatibility:** Output particles follow the `(time, easting, northing)` format and shape `(N_particles, N_obs, 3)` (similar as Figure 4.3, meeting the input requirements of the `MeshFreeAdjointAdvectionDiffusionModel`.

Overall, this sensor design enables scalable and flexible integration of remote sensing-like observations into the source inference model, supporting both synthetic and real-world data use cases.

4.4.1 Backward Propagation via Adjoint Inference

The inference model relies on an adjoint-based approach to track particles backward in time and reconstruct the source distribution. This logic is implemented in the `computeModelRegressors()` method of the `MeshFreeAdjointAdvectionDiffusionModel`.

1. Particles are traced backward using wind vectors obtained from the `getwind()` function.
2. At each timestep, the position and time of each particle are updated.
3. Kernel basis functions are evaluated at each particle's location and accumulated to form the regressor matrix $\mathbf{X}$.

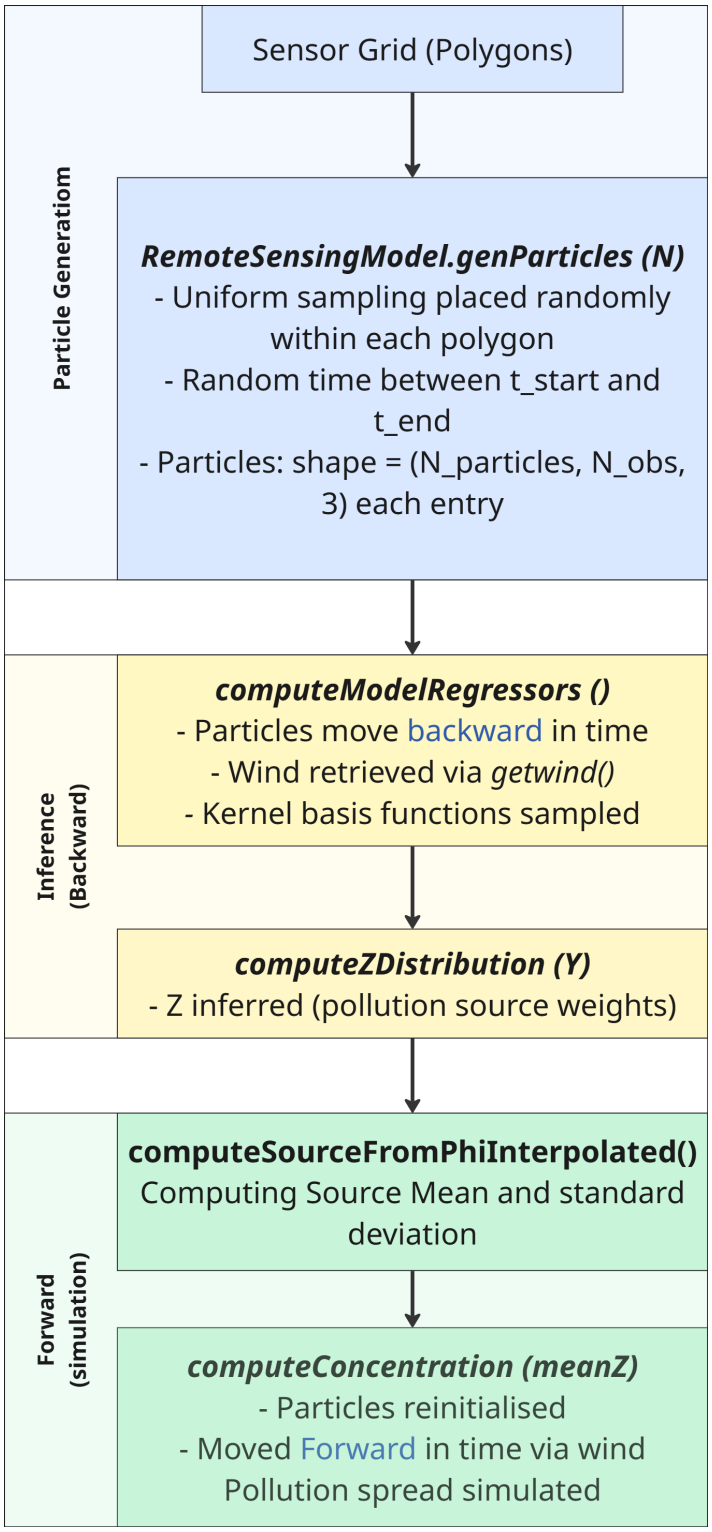

Figure 4.6: Illustration of backward particle movement from sensor polygons, tracing potential origins of observed pollution under wind dynamics.

4.4.2 Sensor Grid and UTM Projection

To simulate satellite-style coverage, the study area over Victoria was divided into a grid of non-overlapping polygons, each representing a sensor footprint. Geographic coordinates were converted to UTM Zone 56S (EPSG:32756) to ensure consistent spatial scaling in metres, aligning with wind data units.

A fixed bounding box:

```
(min_lon=140.5, min_lat=-39, max_lon=150, max_lat=-34)
```

was used to generate the grid based on MERRA-2's 1° resolution. The projection and polygon generation were implemented using Shapely and PyProj, enabling accurate particle placement and motion in a physically meaningful coordinate system.

4.4.3 Particle Generation via `genParticles()`

Polygon-Based Particle Generation

For each polygon in the grid, synthetic particles are sampled using a Poisson point process (via `pointpats.random.poisson`). This stochastic sampling mimics the irregular nature of satellite coverage while maintaining spatial coverage consistency. Each particle is also assigned a random timestamp within a defined time window `[t_start, t_end]`, scaled to seconds to match the model's internal temporal resolution.

The `RemoteSensingModel` defines a method `genParticles(N)` that generates particles from each polygonal sensor region. These particles serve as backward tracers for inferring possible pollution source locations.

- **Uniform Spatial Sampling:** Particles are uniformly distributed across each polygon within bounds defined by the polygon geometry to represent AOD-like remote sensing measurements
- **Random Temporal Assignment:** Each particle is assigned a random timestamp uniformly sampled from the observation window $[t_{\text{start}}, t_{\text{end}}]$. This reflects the temporal uncertainty associated with remote sensing data, which are often time-averaged.
- **Particle Array Shape:** The resulting particle matrix has the shape `(N_particles, N_obs, 3)`, representing time, x (easting), and y (northing).

4.4.4 Planned Testing

To validate the RemoteSensingModel design, testing focused on:

- Confirming that particles were correctly seeded within their respective grid polygons.
- Ensuring temporal values fell within the expected range.

- Visualising grouped particles to check for logical separation between observation cells.
- Reprojecting coordinates back to latitude–longitude to confirm spatial accuracy.

These tests helped confirm the sensor's readiness for use in particle-based backward inference.

4.4.5 Summary

The `RemoteSensingModel` enables realistic simulation of satellite observations by seeding particles across a projected polygon grid in Victoria. Its configurable spatial and temporal sampling ensures compatibility with mesh-free backward inference.

4.5 Pollution Observation Design

To simulate realistic pollution observations, this project incorporates satellite-based Aerosol Optical Depth (AOD) data from NASA's MERRA-2 `inst3_2d_gas_Nx` product [16]. These replace the synthetic values used in earlier tests, enabling application to real-world events like the 2019–20 bushfire season in Victoria, Australia.

The observation vector Y is used during the inference step `computeZDistribution()` to inform backward particle tracing and source reconstruction.

4.5.1 Integration with Real AOD Data

$PM_{2.5}$ is the most relevant surface-level pollutant, but station-based data is sparse. AOD, available globally from satellites, was therefore selected as the main pollution input, later converted to estimated $PM_{2.5}$ for validation.

The integration process included:

- **Temporal Alignment:** Matching the model time with the nearest AOD snapshot (e.g., 30 Dec 2019, 00:00 UTC).
- **Coordinate Projection:** Converting UTM Zone 56S centroids to lat/lon via `pyproj`.
- **Index Lookup:** Using NumPy `argmin` to find the closest AOD grid cell.
- **Value Extraction:** Retrieving values from the `AODANA` variable.

4.5.2 AOD to $PM_{2.5}$ Estimation

To improve interpretability, AOD was converted to $PM_{2.5}$ using statistical models informed by prior studies. $PM_{2.5}$ is more relevant for surface-level pollution analysis.

This study adopts the approach of Tian et al. [37], applying a Random Forest regression model that integrates AOD with meteorological variables such as temperature, humidity, and boundary layer characteristics. Their work highlights how machine learning (ML) methods—particularly ensemble tree-based models—can extract meaningful nonlinear relationships from noisy and multi-dimensional environmental datasets.

4.5.3 Combining AOD, Wind, and Station-Based $PM_{2.5}$ Observations

Three datasets were integrated:

- **AOD:** NASA MERRA-2 `inst3_2d_gas_Nx` [16]
- **Meteorology:** Wind components (`U`, `V`), temperature (`T`), pressure (`ps`), and humidity (`RH`) from `tavg3_3d_asm_Nv` [17]
- **Ground truth:** Hourly $PM_{2.5}$ data from Victorian EPA stations [12]

Model Features

- AODANA (Aerosol Optical Depth)
- Wind speed (derived from `U`, `V`)
- Relative humidity (`RH`)
- Specific humidity (`QV`)
- Temperature (`T`)

These features were selected based on availability in MERRA-2 and relevance to $PM_{2.5}$ estimation.

4.5.4 Planned Testing

Observation data will be tested by:

- Overlaying extracted AOD values on Cartopy maps to check spatial consistency.
- Checking for missing values (e.g., NaNs) and ensuring values were scaled to [0, 1].
- Evaluating $PM_{2.5}$ predictions via Random Forest using R^2 and RMSE metrics.
- Analysing feature importance to confirm that AOD and meteorological variables contributed meaningfully.

These checks ensured the data could be reliably integrated into the inference model.

4.6 System Integration Testing Plan

A final integration test was conducted to ensure that all modules—wind model, sensor model, observation data, and inference kernel—worked cohesively within the AdvectionGP framework. This involved verifying:

- Successful wind-driven particle advection without runtime errors or NaNs.

- Correct regressor matrix construction and source inference outputs.
- Geospatial alignment of particle paths and source maps with Victoria's boundaries.
- Plausible pollution concentration maps from forward simulation.

Initial tests used synthetic inputs (e.g., constant wind), followed by real AOD and $PM_{2.5}$ data for final validation.

4.7 Summary of Design Choices

While Chapter 4 introduced the initial design considerations, most of the final decisions were solidified through practical implementation and iterative testing. These were guided by empirical observations related to model compatibility, runtime performance, and output accuracy. The final configuration adopted in this project is summarised below:

- **Model Type:** `MeshFreeAdjointAdvectionDiffusionModel` selected for its flexibility in supporting irregular, sparse sensor data and particle-based inference logic.
- **Wind Model:** A custom `RealWind` class was implemented to incorporate real MERRA-2 wind data, supporting efficient spatial and temporal lookups.
- **Kernel:** The Exponentiated Quadratic (EQ) kernel was used to define spatial basis functions for Gaussian Process inference.
- **Sensor Model:** A custom `RemoteSensingModel` was developed to simulate satellite-style observations and generate particles from polygonal sensor regions.
- **Projection:** UTM Zone 56S (EPSG:32756) was used to minimise spatial distortion over southeastern Australia.
- **Pollution Source:** Initially modelled using a synthetic Gaussian blob; later replaced with real AOD data from the MERRA-2 dataset to improve physical realism. Then would be converted into PM2.5 using Random Forest.
- **Vertical Scope:** Wind data averaged over vertical levels 56–67 to represent the lower free troposphere, consistent with aerosol transport literature.
- **Wind Lookup Strategy:** Nearest neighbour implemented for initial speed, with interpolation subjected to comparative evaluation in Chapter 5.
- **Validation Strategy:** Source estimates and predicted concentrations are visually compared with known bushfire hotspots and AOD spatial patterns.

These design choices reflect a careful balance between realism and computational feasibility, supporting rigorous testing while ensuring practical deployment over the bushfire-affected regions of Australia.

Chapter 5

Implementation and Testing

5.1 Overview

This chapter describes the implementation of the system architecture and design described in Chapter 4. It covers the practical steps taken to translate theoretical design elements into executable Python code and to integrate them into the existing `AdvectionGP` framework[1]. The goal of implementation is not only to operationalise the pollution source inference model, but also to visualise, validate, and test it in a geographically meaningful context—specifically over Victoria, Australia, during the 2019–20 bushfire season.

To support this, the model outputs—such as particle trajectories, inferred sources, and pollution concentrations—are visualised using `Cartopy`, overlaid on map projections of Australia. This ensures that source inference is not only numerically verified, but also spatially reasonable.

All source code was developed and tested in a forked version of the `AdvectionGP` repository, hosted on GitHub: `https://github.com/efamelody/advectionGP`.

5.2 Environment Setup

The implementation was carried out in Python, using open-source libraries for geospatial data processing, scientific computation, and visualisation. The following tools were essential:

- **Python Version:** 3.9
- **Key Libraries:**
  - `netCDF4` – for accessing MERRA-2 NetCDF files
  - `xarray` – for manipulating multi-dimensional geophysical arrays
  - `numpy`, `scipy` – for numerical operations and KD-tree-based lookup
  - `pandas` – for handling tabular wind data and efficient filtering

[1] `https://github.com/SheffieldML/advectionGP`

- `pyproj` – for coordinate transformations between geographic (WGS84) and projected (UTM Zone 56S) systems
- `matplotlib` – for general-purpose plotting
- `cartopy` – for map-based visualisation of pollution spread, wind vectors, and observational data
- `shapely` – for geometric operations and polygon generation in sensor modelling
- `geopandas` – for reading shapefiles such as fire ignition points
- `time`, `datetime`, `os` – for managing timestamps, file loading, and temporal alignment

Project Directory Structure

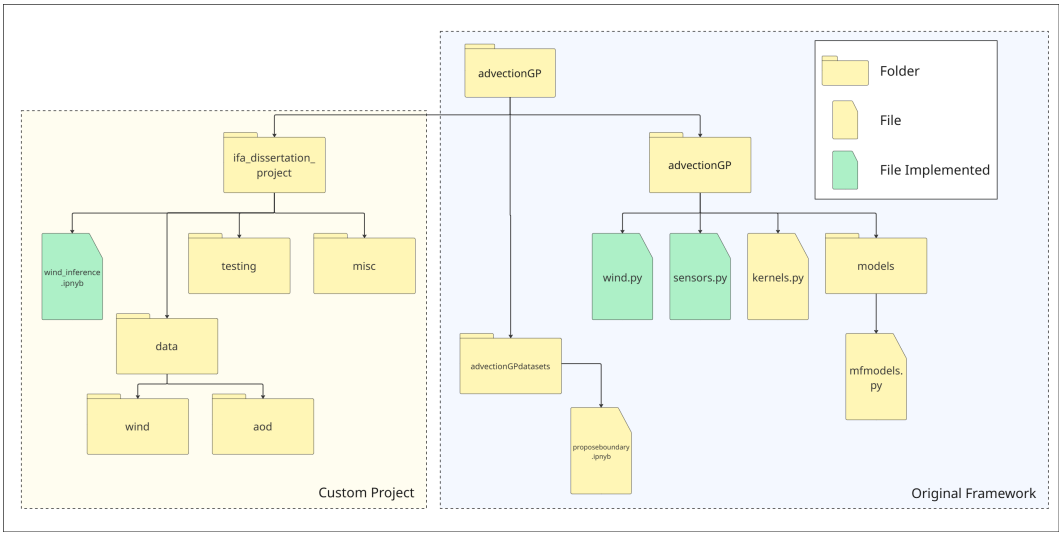

Figure 5.1: Modified Directory Structure of the AdvectionGP Framework

This diagram illustrates the directory layout used in this project. Yellow folders and files are part of the original AdvectionGP repository. Green files indicate where new implementations or customisations were added for this dissertation. The left section shows the structure of the custom experimental project folder, while the right shows the base framework.

Data Access and Preprocessing

Due to the size and nature of NASA's MERRA-2 datasets, all data was downloaded manually from NASA's GES DISC portal[2] and stored locally. The data was preprocessed using the procedures outlined in Chapter 4, including:

- Spatial subsetting via bounding box filters (Victoria, Australia)

[2] `https://disc.gsfc.nasa.gov/`

- Temporal alignment and layer averaging
- Conversion to UTM for consistent spatial indexing
- Construction of lookup tables or KD-trees depending on the wind model variant

With this environment in place, the next sections detail the implementation of core components such as the custom wind models, particle-based sensor generation, and integration of real AOD observations into the inference pipeline.

5.3 Wind Model Implementation and Testing

Wind modelling was the first component implemented in this project due to its central role in controlling particle movement and, consequently, pollution source inference. This section outlines how the real wind model (`RealWind`) was implemented and iteratively refined to balance accuracy, speed, and compatibility with the AdvectionGP framework.

5.3.1 Implementation of `RealWind`

The wind model loads wind vectors from NASA's MERRA-2 dataset, averaging over selected vertical layers and a user-defined bounding box to reduce data volume. The core functionality is split into two stages:

1. Initialisation and Preprocessing

Wind data were extracted from MERRA-2 NetCDF files within the defined bounding box and vertical layer range (e.g., layers 56–68) over the selected simulation period. Files were loaded iteratively by date, and eastward (`U`) and northward (`V`) wind components were spatially subsetted and averaged across vertical layers to approximate free troposphere winds.

2. Data Structuring and Lookup

The processed wind vectors were stored in a Pandas DataFrame containing timestamps, UTM Easting, UTM Northing, and wind components. A KD-tree was constructed over spatial coordinates for efficient nearest-neighbour lookup, while temporal matching was performed by finding the closest timestamp. This enabled rapid retrieval of local wind vectors during particle tracing.

5.3.2 Wind Lookup Methods

The nearest-neighbour method (`RealWindNearestNeighbour`) provided a fast and interpretable baseline but became computationally inefficient during large-scale particle tracing. To address this, several alternative wind lookup strategies were explored:

- **Binned KD-Trees (`RealWindBinned`):** Wind values were grouped into 3-hour bins, with a dedicated KD-tree built for each timestamp. This drastically reduced the temporal search cost and improved performance. However, issues with timestamp mismatches sometimes resulted in missing wind data (e.g., NaNs) during simulation.
- **Hybrid Interpolation (`RealWindHybrid`):** Interpolated wind values across time bins to smooth transitions. This improved accuracy but significantly increased runtime due to repeated tree queries and linear blending.
- **Fast Grid Indexing (`FastWindGrid`)**: Ultimately, a regular grid-based indexing method was developed to provide fast wind vector lookup using array slicing. Wind data were preloaded into structured NumPy arrays with dimensions `[time, y, x]`, allowing rapid retrieval using integer index conversion from projected coordinates. This method achieved the best balance of speed and reliability, especially when embedded in the full model pipeline.

Despite the conceptual benefits of interpolation-based methods, extended runtimes—exceeding one hour per inference run—presented a significant limitation. Although higher accuracy was theoretically achievable through temporal interpolation, the computational burden impeded iterative testing and model experimentation. A detailed comparison of all implemented wind models, including `RealWindNearestNeighbour`, `RealWindBinned`, and `RealWindHybrid`, is provided in Appendix 6.7.

5.3.3 Computational time

While theoretical complexity analysis offers insights into algorithmic performance, it is equally important to measure runtime in a real-world setting, where compatibility with the simulation model, data volume, and memory access patterns affect actual speed.

To that end, each wind model variant was benchmarked using a consistent configuration: 30 particles across 50 observation points over a 5-day simulation. The metrics recorded include the time to generate particles (`genParticles`), construct the regressor matrix (`computeModelRegressor` and compute the predicted pollution concentrations (`computeConcentration`). The results are summarised in Table **??**.

Wind Model	**Time to genParticles**	**Time to computeModelRegressors**	**Time to computeConcentration**
RealWindNearestNeighbour	18s	105s	190s
RealWindBinned	7s	32s	Error (NaNs)
RealWindHybrid	14s	90s	260s
FastWindGrid	**3s**	**24s**	**36s**

Table 5.1: Wind Model Runtime Comparison (30 particles, 50 observation points)

5.3.4 Visual Testing of Wind Models

Each version of the wind model was evaluated through visual inspection of particle trajectories on a Cartopy map over Victoria, Australia. A total of 500 particles were generated, with 10 particles per observation point across 50 grid cells (Figure 5.2). These particles were projected and visualised (Figure 5.3).

Visual inspection is essential when using real wind data, as it helps verify that wind vectors behave realistically across time and space, ensuring simulation credibility beyond numerical accuracy.

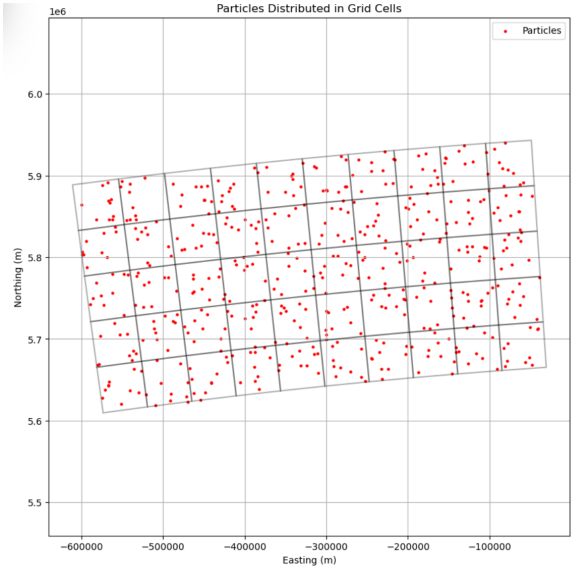

Figure 5.2: Random particles generated across 50 grid polygons. Each grid represents one observation, and each observation spawns 10 particles.

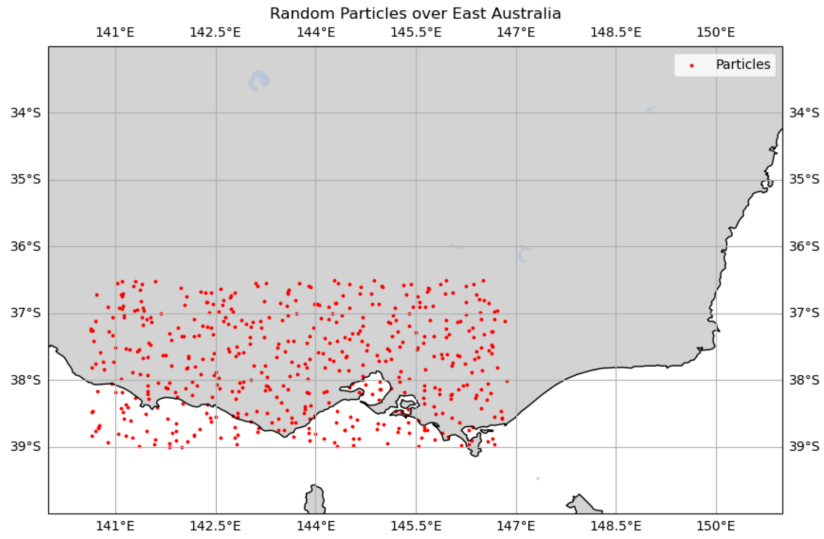

Figure 5.3: Initial particle positions projected across Victoria, used as the starting state for evaluating different wind models.

WindSimple

A constant-wind model was used during early debugging to verify pipeline functionality.

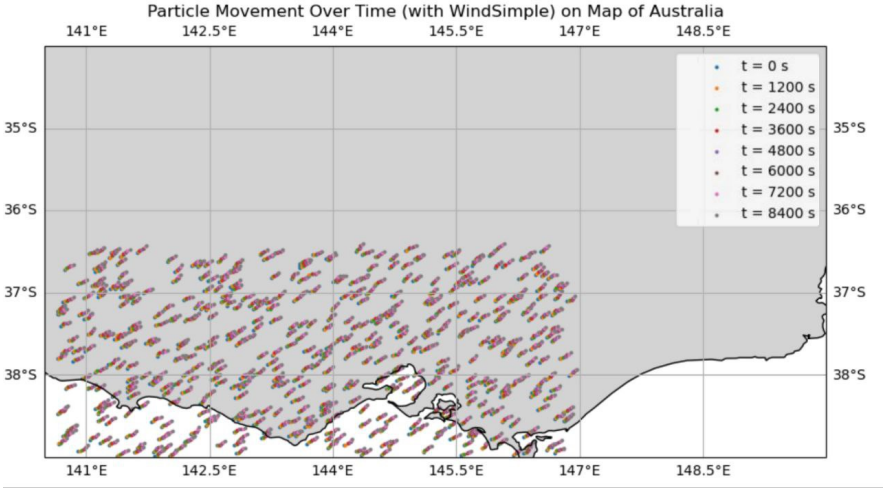

Figure 5.4: Particles moved diagonally and uniformly, confirming pipeline stability.

RealWindNearestNeighbour

This model uses spatial nearest-neighbour lookup for wind retrieval. Visually, particles consistently move south-eastward, aligning with expected meteorological patterns. Some irregular spacing between particle paths suggests minor interpolation artifacts, but all particles remain present and the flow appears realistic.

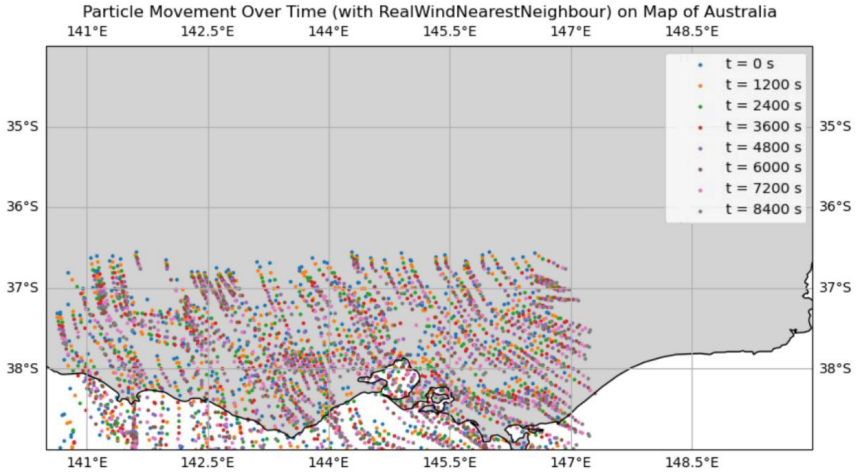

Figure 5.5: Particle motion under RealWindNearestNeighbour. Slight spacing irregularities are visible, but overall consistency is maintained.

RealWindBinned

Wind is grouped into 3-hour temporal bins with a KD-tree per timestamp. While this method improves speed, some particles, especially in western Victoria, appear to vanish. This indicates missing wind values (e.g., NaNs), which impacted model performance.

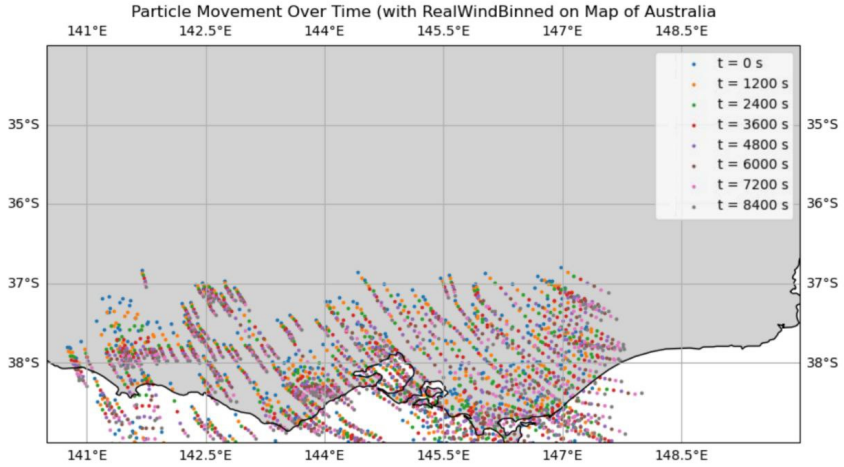

Figure 5.6: Particle trajectories under RealWindBinned model. Missing wind data in some regions caused particles to disappear.

RealWindHybrid

This method interpolates wind between time bins to increase temporal accuracy. While particle trajectories look smoother, some still disappear. Additionally, this method introduced high computational cost and still suffered from particle loss due to lookup gaps.

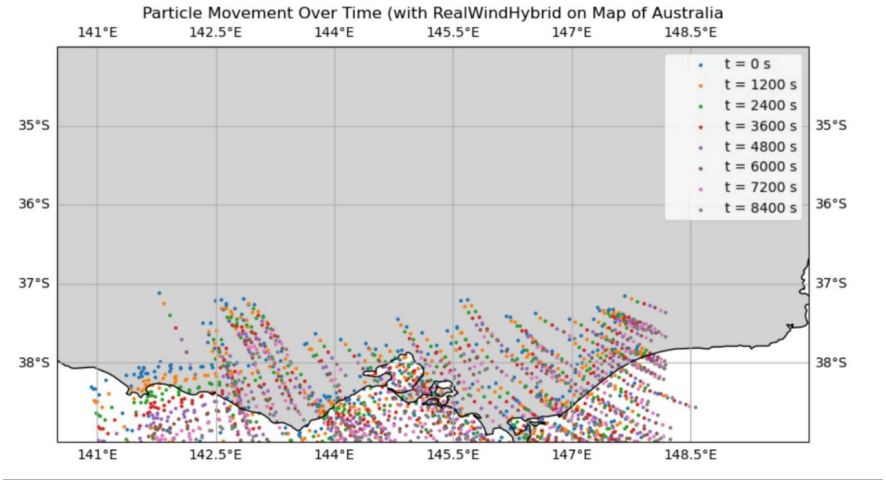

Figure 5.7: RealWindHybrid model with temporal interpolation. Some particles still vanish, and computation is significantly slower.

FastWindGrid

This model uses pre-indexed NumPy arrays to retrieve wind values directly by grid index. It achieved the fastest runtime and visually consistent, dense particle paths. All particles remain visible, and their flow appears meteorologically plausible.

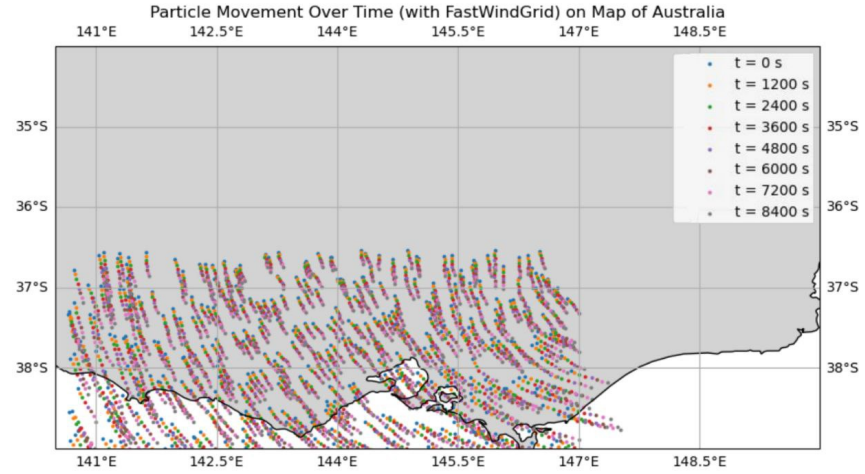

Figure 5.8: FastWindGrid model with indexed wind retrieval. All particles appear and flow uniformly, indicating accuracy and efficiency.

5.3.5 Model Selection Trade-offs

Summary of key trade-offs observed during testing:

Table 5.2: Comparison of Wind Lookup Methods

Method	**Speed**	**Accuracy**	**Reliability**
Nearest Neighbour	Slow	Visually realistic	All particles present
Binned KD-tree	Fastest	Incomplete (NaNs)	Some particles missing
Hybrid Interpolation	Very slow	Moderate	Some particles missing
Fast Grid Indexing	Fastest	Consistent	All particles present

5.3.6 Supplementary Wind Accuracy Analysis

While visual inspection confirmed that wind fields produced plausible particle motion, a brief exploratory analysis was also performed to assess spatial consistency. Wind components (U and V) were extracted from four nearby locations over a 24-hour period (e.g., December 20, 2019). Both eastward and northward winds showed stable, smooth trends with minimal spatial variation (Figures 5.9 and 5.10), supporting the use of nearest-neighbour and binned methods in this context.

Due to time constraints, full quantitative validation—such as Root Mean Square Error (RMSE) between interpolated and true wind vectors—was not conducted. This remains an area for future work to improve physical credibility, especially when generalising to other regions or longer simulation windows.

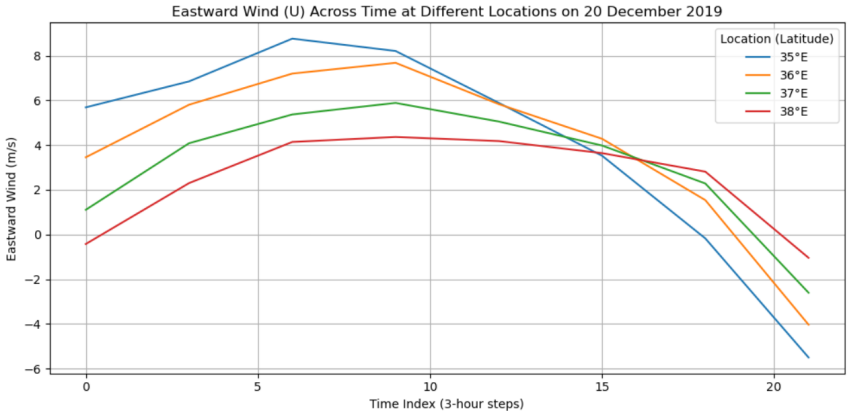

Figure 5.9: Eastward wind (U component) over time at four adjacent locations.

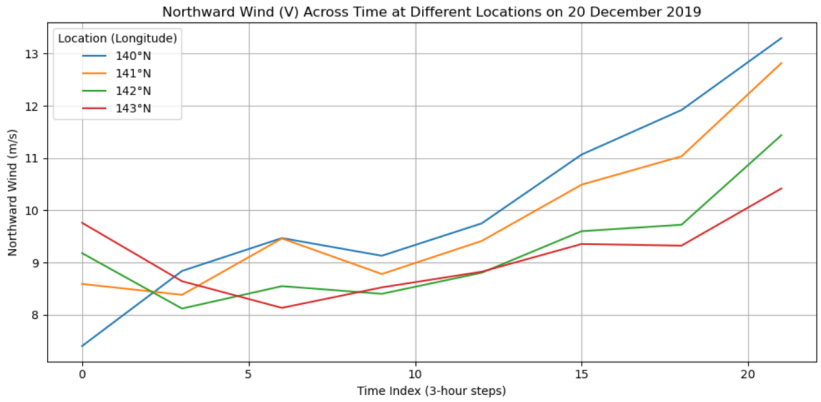

Figure 5.10: Northward wind (V component) over time at the same locations.

5.3.7 Final Selection: `FastWindGrid`

The `FastWindGrid` method was ultimately selected due to its optimal balance between speed and accuracy. As discussed in Section 5.3.3, it integrates efficiently within the broader model pipeline, enabling rapid wind vector lookup without sacrificing visual or spatial consistency.

Its efficiency allows for more extensive testing across other components of the model. In contrast, slower alternatives—such as `RealWindHybrid`—significantly increase runtime, limiting the feasibility of repeated experimentation and refinement. Thus, `FastWindGrid` was chosen as the most suitable option for large-scale particle simulations and iterative development. Although `FastWindGrid` achieved the best balance, future work could explore hybrid grid-interpolation schemes for even finer spatial fidelity.

5.4 Sensor and Particle System

This section documents the implementation and validation of the custom `RemoteSensingModel`, which emulates satellite-style observation using particles distributed over spatial polygons. The focus here is on how the sensor grid was generated, how particles were seeded across this grid, and how the resulting system was visually and logically tested.

5.4.1 Sensor Grid Creation

To replicate the spatial distribution of remote sensing data, a grid of observation polygons was created over the region of Victoria, Australia. Latitude and longitude coordinates from the MERRA-2 dataset were first projected into Universal Transverse Mercator (UTM) coordinates (EPSG:32756) to enable consistent distance calculations. Using a resolution of 1°, a meshgrid of points was constructed and projected to (Easting, Northing) coordinates, from which a series of square polygons were generated using the `Shapely` library.

This process resulted in a grid of 50 spatial regions, each representing a potential remote sensing observation zone. These polygons serve as the spatial domains within which particles

are sampled. The successful construction of this grid was confirmed through visual inspection (Figure 5.11), showing clear alignment between the expected region and the generated polygons.

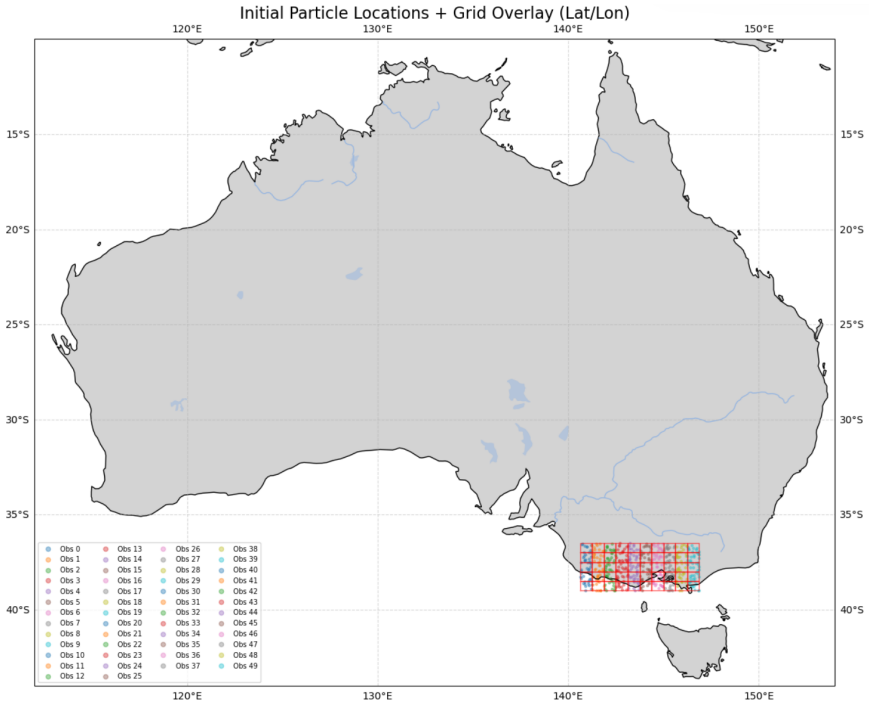

Figure 5.11: Initial Particle Locations with Grid Overlay in Lat/Lon. Grid cells are defined over Victoria.

5.4.2 Particle Generation

The `RemoteSensingModel` class defines a `genParticles()` method to populate each grid polygon with particles. Particle seeding is done using a Poisson distribution (`pointpats.random.poisson`), reflecting the irregular but spatially distributed nature of real satellite observations. Each particle is also assigned a random timestamp within a user-defined window (e.g., $[1260, 1440]$ minutes), converted into seconds to match internal model conventions.

The particle array is structured as $(N_{\text{particles}}, N_{\text{obs}}, 3)$, where each entry corresponds to $[t, x, y]$—the time (in seconds), Easting, and Northing coordinates. This structure ensures compatibility with the downstream mesh-free inference pipeline.

5.4.3 Visual Testing and Validation

To ensure the sensor model functions as expected, the following validation steps were performed:

1. **Grid Formation Validation:** Verified that polygons were correctly formed from the MERRA-2 grid and projected into UTM.

2. **Particle Distribution:** Ensured that all particles were spatially contained within their respective polygons using visual plots.

3. **Logical Grouping:** Confirmed that particles from each observation region are distinct and maintain association with their grid cell, as shown in Figure 5.12.

4. **Reprojection Accuracy:** Performed back-projection to confirm that particles could be transformed back into lat/lon for plotting and validation.

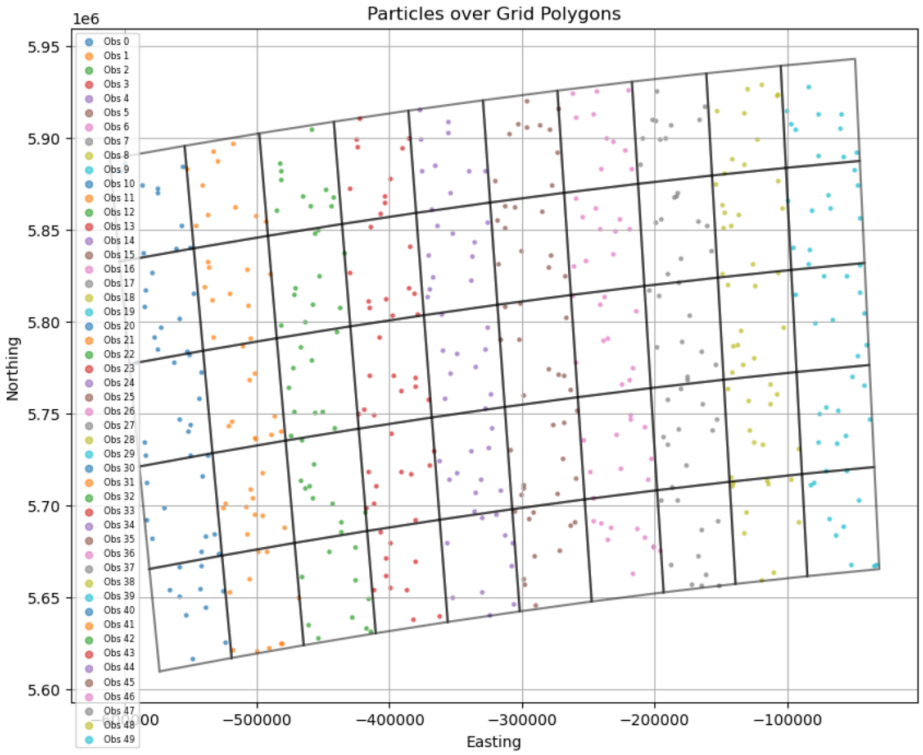

Figure 5.12: Particles grouped by grid polygon in UTM coordinates. Each color represents particles from a different observation cell.

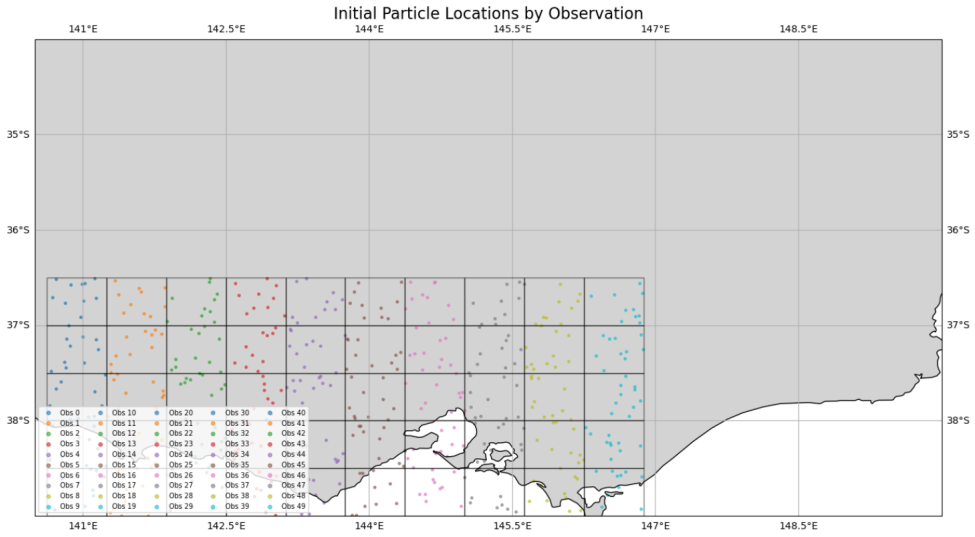

Figure 5.13: Particles distributed across Victoria in UTM coordinates.

5.4.4 Discussion

This sensor and particle generation framework allows realistic, high-resolution emulation of remote sensing observations. The grid flexibility, UTM-based consistency, and reproducible sampling mechanisms enable a scalable setup for testing pollution inference scenarios. Furthermore, visual testing confirmed the correctness of polygon definitions, particle placement, and spatial grouping, providing confidence in the model's readiness for integration into the full inference pipeline.

5.5 Pollution Observation (Y)

5.5.1 AOD Implementation

To replace synthetic observations with real-world pollution data, Aerosol Optical Depth (AOD) values were extracted from the MERRA-2 `inst3_2d_gas_Nx` [16] dataset. These values served as the pollution observation matrix Y used in the inference model.

Observation Locations: Observation locations were derived directly from the centroids of the grid polygons generated by the `RemoteSensingModel`. These polygons were designed to simulate the spatial coverage of remote sensing instruments over Victoria. Each centroid was transformed from UTM Zone 56S (EPSG:32756) to geographic coordinates (longitude and latitude) using PyProj. These transformed coordinates were then used to locate the nearest AOD values in the MERRA-2 dataset's spatial grid.

Time Selection: A single timestamp was selected from the AOD dataset by identifying the closest available time index to the start of the simulation. This provided a consistent snapshot of atmospheric aerosol concentration for the defined observation window.

AOD Extraction: For each sensor location, the nearest grid point in the AOD data was found by comparing geographic coordinates with the latitude and longitude arrays in the NetCDF file. AOD values were extracted for the corresponding locations and time step. To prepare the data for inference:

- Missing values (`NaN`) were replaced with zeros using `np.nan_to_num()`.
- Values were min-max normalised to the range $[0, 1]$, ensuring compatibility with the inference model.

This normalised AOD vector was used as the observational input Y, enabling spatial source inference using real satellite-derived aerosol data.

5.5.2 Challenges with AOD-Derived Observations

AOD reflects column-integrated aerosol density rather than surface-level pollution. It is affected by vertical mixing, aerosol type, and humidity, making direct interpretation within a backward particle framework non-trivial [37].

These limitations are discussed in more detail in Chapter 6, including implications for source inference and the need for conversion models such as AOD-to-$PM_{2.5}$ to improve surface relevance.

5.6 Conversion of AOD to $PM_{2.5}$

Following the design proposed in Chapter 4, this section outlines the conversion of Aerosol Optical Depth (AOD) to $PM_{2.5}$ using a Random Forest regression model. The goal was to generate a spatially resolved surface-level pollution field based on satellite and meteorological features, enabling a more physically relevant observation matrix Y for the source inference model.

The Random Forest model was trained on features such as AOD, wind speed, and relative humidity, and achieved an R^2 score of 0.69 on a held-out test set (see Appendix .0.1). Feature importance analysis revealed AOD, wind speed, and relative humidity as the most predictive variables. Once trained, the model was applied across the spatial grid of Victoria to predict $PM_{2.5}$ concentrations.

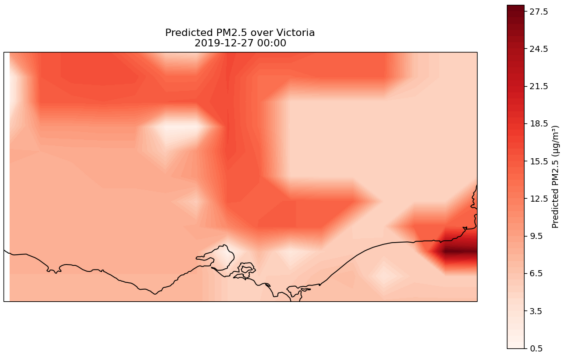

Figure 5.14: Predicted $PM_{2.5}$ concentrations across Victoria on 27 Dec 2019, generated using a Random Forest regression model trained on AOD and meteorological features (see Appendix .0.1). These values formed the observation vector Y for the source inference model.

These predictions were then used as the observation vector Y in the inference pipeline, replacing synthetic values. This transformation allowed the model to utilise satellite-derived data in a more physically interpretable form, aligning observations with surface-level pollution exposure.

5.7 Full Pipeline Integration

Once individual modules such as the wind model and sensor model were implemented and tested in isolation, it was essential to verify their interoperability within the full inference pipeline. This section documents the process of integrating and validating key components in a unified simulation framework, including the sensor particle generation, wind-driven advection, and eventual inference.

Throughout development, the `WindSimple` model was initially employed for its simplicity and ease of debugging. Once the overall structure was confirmed to be working, more realistic wind models such as `FastWindGrid` were incorporated into the integrated system. A synthetic blob source was also used in early testing phases to validate forward diffusion before transitioning to real AOD observations and later $PM_{2.5}$

5.7.1 Wind–Sensor Compatibility Testing

While unit testing individual components is a vital aspect of system validation, ensuring that these modules interact correctly is critical to achieving reliable results in an end-to-end inference setting. This subsection focuses on verifying that particles generated by the sensor model can be accurately propagated forward in time under the influence of the wind model.

Figure 5.15 illustrates a test scenario in which synthetic particles are seeded from sensor grid polygons and advected through time using a realistic wind field. This visualisation confirms that the integration between spatial sampling (via grid polygons) and temporal evolution (via wind-driven advection) is functioning as expected.

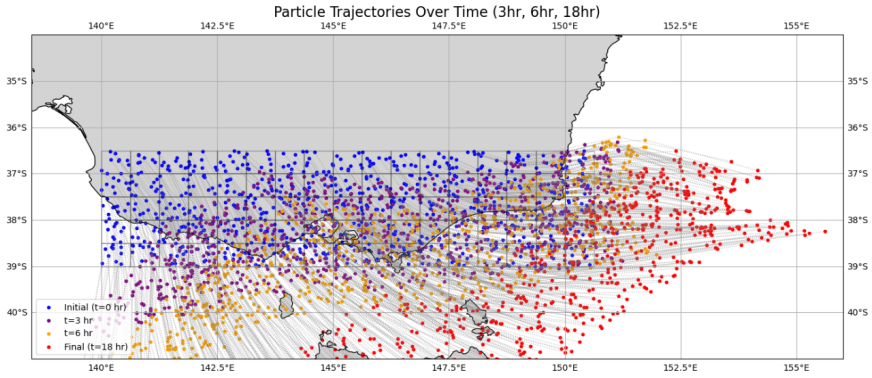

Figure 5.15: Particles generated by the sensor model and advected forward in time using the wind model. This visual test confirms correct interaction between particle generation and environmental dynamics.

From Figure 5.15, it is evident that particles advected under wind influence can travel significantly beyond their original sensor polygons, especially over longer durations such as a full day. This observation was instrumental in informing the spatial design of the inference domain.

In this context, the sensor grid defines the regions where pollution observations—such as Aerosol Optical Depth (AOD)—are collected, analogous to the footprint of a satellite sensor. In contrast, the inference boundary represents the broader spatial domain within which the pollution source is estimated.

Given that particles may originate from locations far beyond their observed positions due to advection and diffusion processes, it is necessary for the inference boundary to be larger than the sensor grid. This ensures that potential source regions are not excluded from consideration and that the model can accurately trace backward trajectories over time.

5.7.2 Validation Techniques and Model Confidence

Throughout the implementation and integration stages, model validation was primarily performed through visual inspection and logical consistency checks. Key validation steps included ensuring that:

- Particles remained within the defined geographic boundaries during advection.
- No particles exhibited NaN (Not-a-Number) values during forward and backward propagation.
- Wind vectors produced by the real wind models were meteorologically plausible when overlaid on map projections.
- Pollution concentration fields evolved smoothly and consistently over time.

Visual validation provided rapid, intuitive feedback during development, allowing the detection of major structural or dynamical issues in particle movement and source estimation.

Logical checks—such as verifying that all particles received valid wind vectors and that mass conservation trends were plausible—further reinforced confidence in model behaviour.

However, a full quantitative validation, such as computing interpolation errors between predicted and ground-truth wind fields or comparing predicted and observed $PM_{2.5}$ values using RMSE or correlation coefficients, remains an area for future work. While qualitative validation was sufficient for iterative development, incorporating quantitative performance metrics would provide a more rigorous assessment of model accuracy and reliability across diverse conditions.

Component-Level Testing. In addition to visual and logical validation, several components were explicitly tested in isolation to ensure correctness:

- **Wind Model:** Verified that specific (time, x, y) queries returned expected wind vectors from the preprocessed dataset, confirming the accuracy of spatial-temporal lookup and KD-tree/indexing strategies.
- **Sensor Model:** Confirmed that generated particles were strictly contained within their respective polygons and that timestamps matched the defined observation window.
- **$PM_{2.5}$ Conversion:** Validated that AOD and meteorological features were correctly matched to EPA ground station data using coordinate rounding and timestamp alignment, with results assessed using R^2 and visual correlation plots.

While not structured using an automated unit testing framework, these functional checks served as de facto unit tests, ensuring reliability and correctness across key modules.

Final Inference Model Configuration

The final inference step uses the `MeshFreeAdjointAdvectionDiffusionModel` to integrate sensor layout, wind model, kernel, and $PM_{2.5}$ observations for source estimation.

Kernel and Features. An `EQ` kernel with lengthscales `[8×24×3600, 50000, 50000]` (8 days, 50 km) and `N_feat` = 1500 Fourier features captures smooth spatiotemporal patterns in pollution dispersion.

Resolution. The model grid resolution is `[40, 30, 30]` in time, x, and y dimensions—balancing spatial detail with runtime over a 5-day window.

Noise and Scaling. Observation noise (`noiseSD` = 0.001) accounts for pre-processed $PM_{2.5}$ uncertainty. The prior variance (`k_0` = 0.5) controls the strength of the source prior.

Wind and Sensor Models. `FastWindGridSurface` was used for near-surface wind, initialised from 25 Dec 2019 over 8 days. The sensor model (`RemoteSensingModel`) generated 10 particles per polygon, with observation times spanning the full inference window.

Execution. Model inference was carried out using `computeModelRegressors`, `computeZDistribution`, and `computeConcentration`. Inferred source maps were visually validated against fire ignition points and $PM_{2.5}$ spread.

5.8 Summary

This chapter presented the practical implementation of the pollution source inference system using the AdvectionGP framework, focusing on operationalising theoretical designs into executable Python modules. Major components implemented include the development of several real wind models, culminating in the selection of the `FastWindGrid` method for its optimal speed-accuracy trade-off, and the creation of a custom sensor model simulating remote sensing observations through particle-based techniques.

Real-world aerosol observations were integrated by extracting AOD data from the MERRA-2 dataset and subsequently applying a Random Forest regression model to estimate surface-level $PM_{2.5}$ concentrations. These enhanced observations were incorporated into the inference pipeline to improve the physical relevance of source estimation.

Key trade-offs were encountered and addressed during implementation, particularly balancing computational efficiency against physical accuracy in the wind model and ensuring practical feasibility through visual validation techniques.

With the full pipeline successfully integrated and validated, the next chapter will present the results of the pollution source inference, including spatial and temporal source estimates, comparison with known bushfire ignition points, and a discussion of model performance and limitations.

Chapter 6

Results and Discussion

6.1 Inference Results Overview

The pollution source inference model was applied over Victoria, Australia, focusing on the period from 23 December 2019 to 2 January 2020. Initial analyses were conducted over smaller windows, specifically from 25 to 27 December 2019 and from 29 December 2019 to 1 January 2020, before extending to the full period to capture broader fire activity across the region, as documented in Victoria's Fire Origin Data [8].

This extended time window was chosen to encompass multiple fire ignition events, leading to higher pollution intensity and a more diverse set of validation points. A greater number of emission sources enhances the robustness of source attribution by enabling cross-validation and improving model calibration against observed fire activities. The inferred pollution source values represent the rate of emission required to explain observed concentrations, expressed in units of $\mu g/m^3/s$.

6.1.1 Discrepancy between Inferred Pollution Sources and Concentration Maps

The **inferred pollution source** map represents estimated emissions from fire events, derived from AOD data and fire ignition points. In contrast, the **concentration map** depicts $PM_{2.5}$ levels after atmospheric transport and dispersion.

Since the concentration map reflects pollution spread over time, it does not directly match the locations of initial emissions. Wind direction, diffusion, and cloud masking all shape this distribution, often displacing peak concentrations from original fire zones.

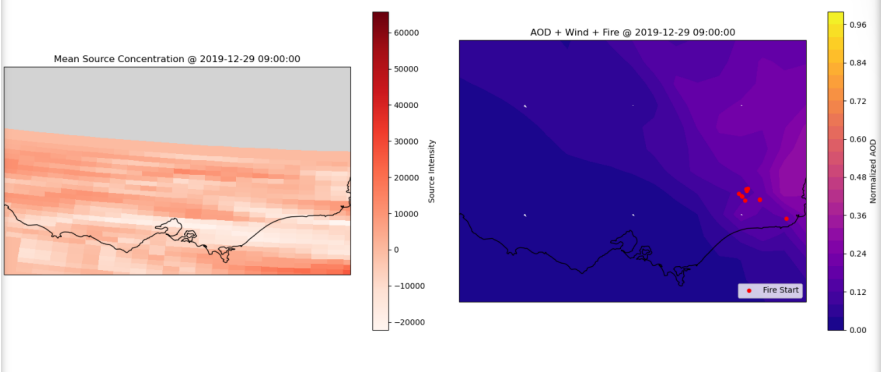

Figure 6.1: Comparison between simulated $PM_{2.5}$ and MERRA-2 reanalysis AOD on 27 December 2019. **(Left)** Simulated surface-level $PM_{2.5}$ concentration field after forward propagation from the inferred source using the AdvectionGP model. **(Right)** AOD values from NASA's MERRA-2 `inst3_2d_gas_Nx` dataset, visualised for the same timestamp and region. While the model predicts surface pollution spreading inland from fire zones, the reanalysis AOD fails to capture the visible smoke plume due to cloud masking and temporal smoothing. This illustrates the limitations of using reanalysis AOD as a proxy for surface pollution in highly dynamic fire events.

The forward simulation does not accurately reproduce the observed aerosol distribution, largely due to the simplifications inherent in the model, the effects of atmospheric transport, and the limitations of AOD as a proxy for surface-level pollution.The challenges of AOD-cloud masking during extreme events have also been reported in Che et al. [7], reinforcing the limitations observed in this study. While this highlights the challenges of forward modelling in complex environments, it does not undermine the utility of the source inference approach, which remains the primary focus of this study.

6.1.2 Validation of Source Inference with Fire Ignition and Standard Deviation Over Time

To validate the spatial accuracy of the inferred pollution sources, the inferred source maps were overlaid with recorded fire ignition points from the Victoria fire dataset. Validation focused on several key dates, including 25, 29, 30, and 31 December 2019. To enhance interpretation reliability, areas with standard deviation exceeding 0.5 were masked. High standard deviation values in this region indicated greater model uncertainty.

On 25 December, despite a recorded fire ignition, the model did not infer a significant pollution source, possibly due to low emission intensity during the early phase of the fire. Nevertheless, the surrounding area exhibited lower standard deviation values, and the region near the ignition point appeared notably darker, suggesting a more confident inference in that location.(see Figure 6.2) From 29 December onwards, the number of fire ignitions increased substantially, leading to noticeable intensification in the inferred source maps, particularly near fire ignition points. Although some areas remained associated with high uncertainty, the overall pollution source signal became more pronounced.

Notably, the spatial pattern of inferred sources from 29 to 31 December shows increased standard deviation near the centre of the domain—closer to the sensor region—reflecting rising uncertainty as the model attempts to infer emissions further back in time. This behaviour is expected in backward models: near the edges (e.g., borders), the wind field is more constrained by entry conditions, while closer to the present (i.e., fire date), particle origins are more uncertain due to the compounding effects of diffusion and diverse possible paths. Thus, although the inferred intensity maps align well with recorded fire ignitions, the corresponding standard deviation maps reveal that this alignment is not equally reliable across space. In future, quantifying these uncertainties in a more intuitive visual form—such as overlaying confidence intervals or shading—is crucial for better communicating model reliability in 2D spatial domains.

Figures 6.2 to 6.9 illustrate the progression of inferred source intensities and associated uncertainties across these dates. Several hotspots of elevated source intensity are evident close to known ignition locations. This correspondence supports the plausibility of the model's output. However, due to underlying uncertainties, some alignments with fire points may be coincidental rather than indicative of true source accuracy. All source inference plots follow a consistent colour scale, with darker shades indicating higher values. Fire ignition points are overlaid for validation reference.

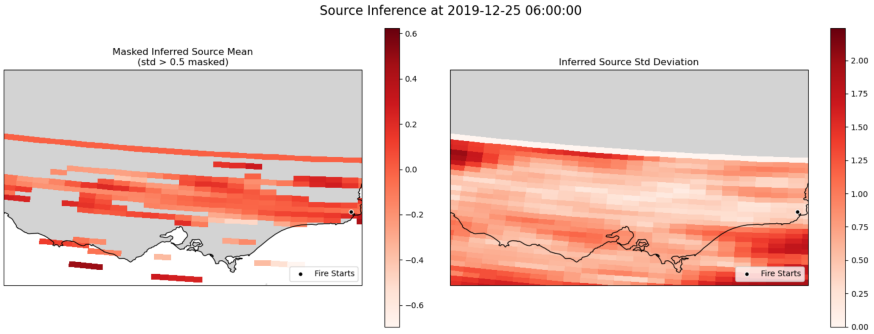

Figure 6.2: Inferred pollution source map (left) and corresponding standard deviation (right) for 25 December 2019 at 06:00. The left panel shows the masked inferred source mean, where regions with high uncertainty (standard deviation 0.5) are greyed out. The right panel presents the spatial distribution of uncertainty in source estimation. Fire start locations are overlaid for reference, highlighting regions where inferred sources are both present and reliable. Units of inferred source intensity are $\mu g/m^3/s$.

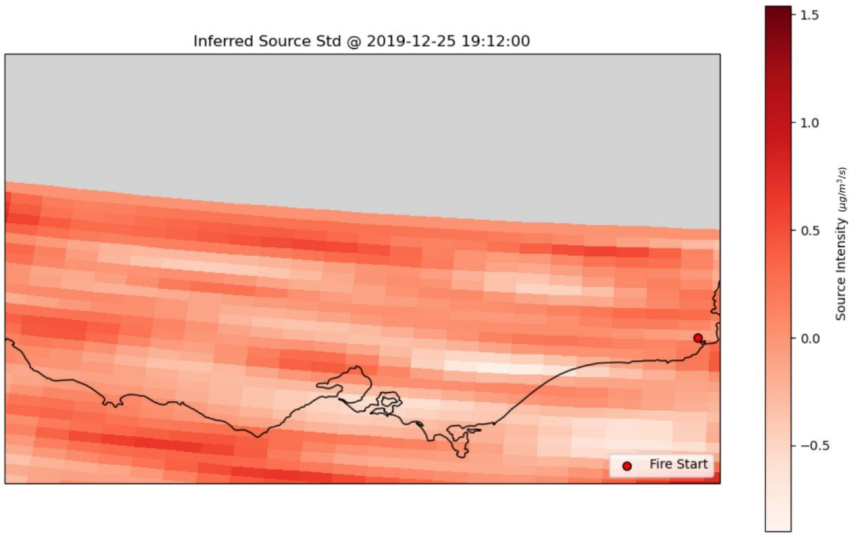

Figure 6.3: Inferred pollution source intensity map for 25 December 2019 at 19:12, generated by the adjoint-GP model. Fire start locations (red dots) are overlaid for reference. The map shows no notable emission strength near fire ignition points, indicating a weak model response at this timestamp.

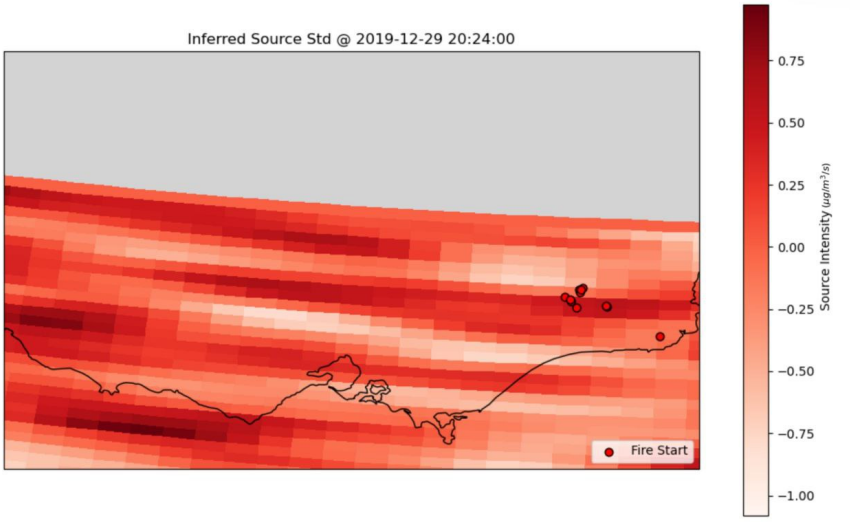

Figure 6.4: Inferred pollution source map for 29 December 2019, showing increased source intensity near fire start regions and surroundings.

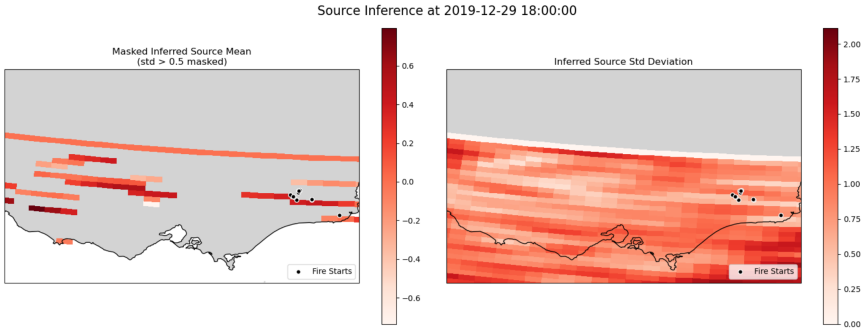

Figure 6.5: Most regions on 29 December show high uncertainty, but the few high-confidence areas align with fire ignition points, suggesting locally confident source estimation. Units of inferred source intensity are $\mu g/m^3/s$.

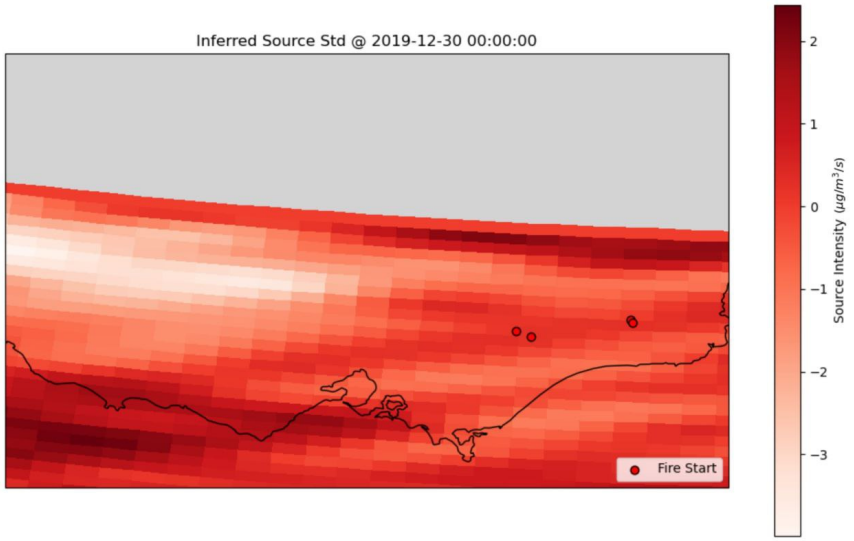

Figure 6.6: Inferred pollution source map for 30 December 2019, showing strong source intensity near central fire start regions.

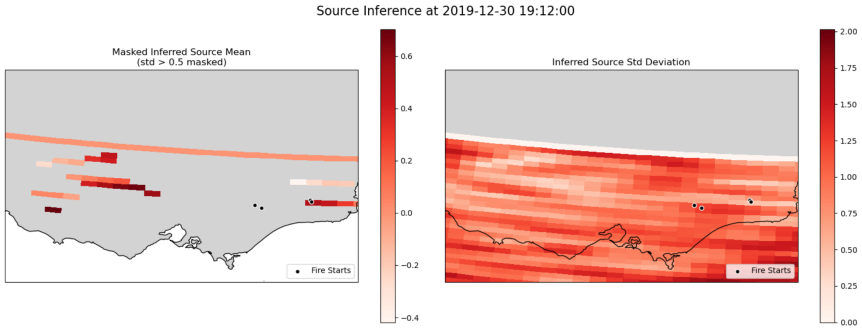

Figure 6.7: Standard deviation map for 30 December 2019. While two central fire ignition points were not captured, the eastern fire aligns with a low-uncertainty, well-defined source region.Units of inferred source intensity are $\mu g/m^3/s$

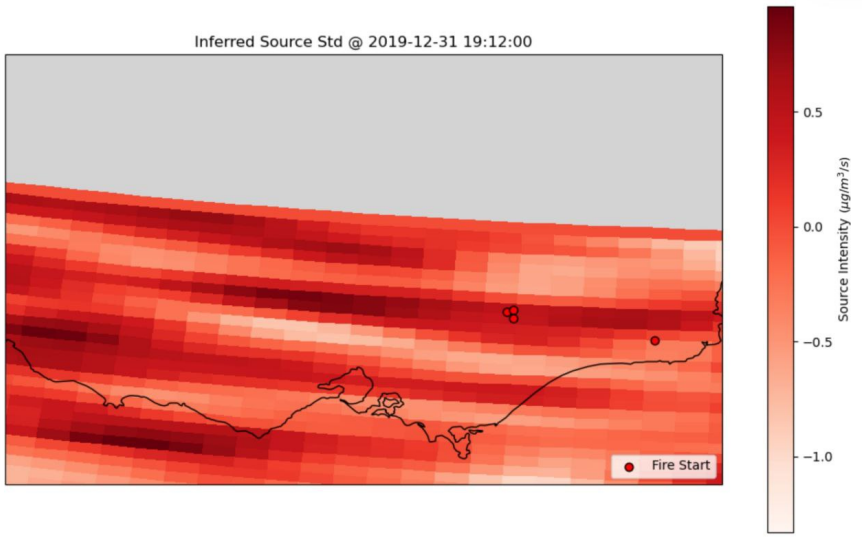

Figure 6.8: Inferred pollution source map for 31 December 2019, showing strong alignment with fire ignition points.

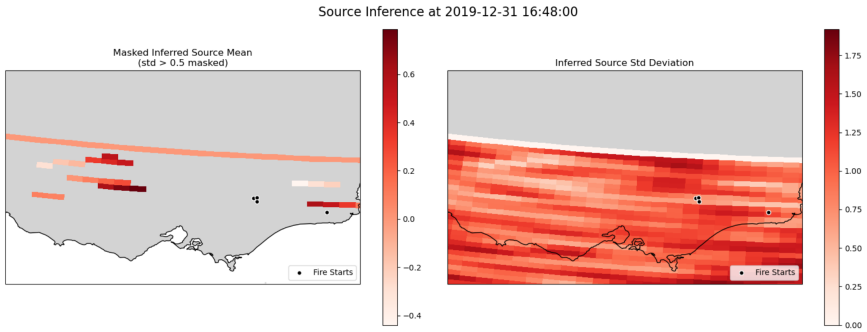

Figure 6.9: Standard deviation map for 31 December 2019, indicating few high-reliability source estimates.Units of inferred source intensity are $\mu g/m^3/s$

Notably, the spatial pattern of inferred sources from 29 to 31 December remains highly consistent across days, with limited variation in source distribution. This may reflect the fact that fire ignition data are reported as single points, whereas real fire activity often spans broader regions—making nearby areas valid contributors to emissions. However, the persistent similarity in inferred maps also suggests potential limitations in the model's temporal sensitivity, reducing its ability to resolve short-term changes in emission intensity. Despite this, several hotspots of elevated source intensity align well with recorded ignition points, supporting the plausibility of the model's output. Some spatial spread is observed, likely reflecting atmospheric transport and diffusion over time. This balance of strengths and limitations is discussed further in the limitations section.

These observed discrepancies between expected and measured AOD patterns motivate a deeper exploration of how AOD interacts with atmospheric dynamics—particularly wind flow and fire activity—which is addressed in the following section.

6.2 Discrepancies Between AOD Levels and Bushfire Intensity

Despite the intensity of the 2019–2020 Australian bushfires, global AOD visualisations (Figure 6.10) and regional analysis over Victoria show relatively weak aerosol signals compared to global hotspots. For example, AOD levels across southeastern Australia were substantially lower than those observed over Central Africa or the Amazon Basin during the same period.

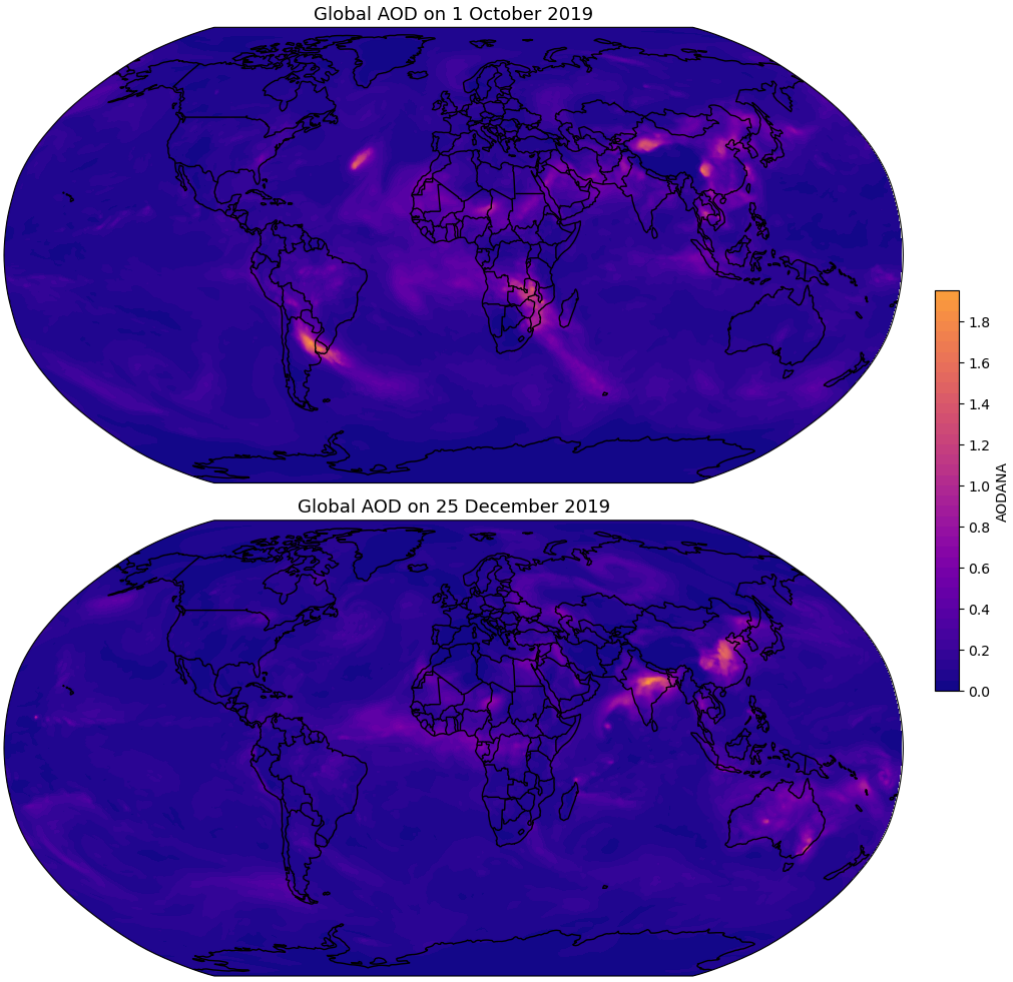

Figure 6.10: Global AOD from NASA MERRA-2 reanalysis on 1 October 2019 at 21:00 UTC, showing spatial variation in aerosol loading. Despite the early stages of the 2019–20 bushfire season, southeastern Australia exhibits relatively low AOD compared to major global hotspots like Central Africa and the Amazon Basin. This reanalysis product integrates multiple observational sources but may underestimate fire-driven aerosols in regions lacking dense ground observations.

This highlights a key limitation of AOD as a proxy for surface-level pollution. As a column-integrated satellite measurement, AOD reflects total aerosol loading throughout the atmosphere and may not capture near-surface concentrations relevant to public health or emission source attribution. Aerosol Optical Depth (AOD), while offering broad spatial coverage, reflects the total atmospheric column rather than near-surface concentrations relevant to public health and source attribution [20]. Its interpretation is further affected by cloud interference, viewing geometry, and the vertical distribution of aerosols.

To assess the bushfires' impact on aerosol loading, AOD values over Victoria were compared on two representative dates: 1 October 2019 (pre-fire) and 25 December 2019 (during active fire). As shown in Figure 6.11, while AOD levels increased during the fire period, they

remained spatially variable with relatively modest median values.

On 1 October, AOD values were tightly clustered between 0.05 and 0.15. By 25 December, although the minimum remained similar, the upper distribution shifted significantly, with the 99th percentile reaching 1.51 and a maximum of 2.37. This suggests that while bushfire emissions did elevate aerosol levels, high AOD events were sparse and not uniformly distributed.

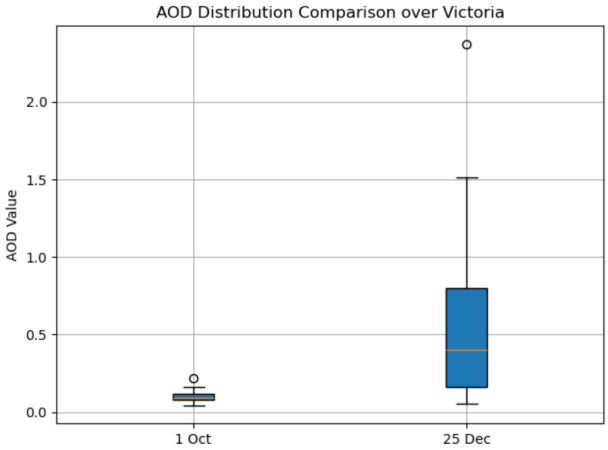

Figure 6.11: Boxplot comparison of AOD distributions over Victoria on 1 October and 25 December 2019. The 25 December dataset exhibits a wider spread and higher extreme values, reflecting the influence of bushfire emissions on atmospheric aerosol concentrations.

Visual inspection using NASA Worldview on 25 December 2019 [27] reveals that thick smoke plumes over southeastern Australia coincided with missing AOD data, likely due to cloud cover or sensor limitations; these manifest as striped gaps across the domain (see Figure 6.12). Unlike satellite products, MERRA-2 reanalysis fills these observational gaps using model-based assimilation, which may smooth over or underestimate the severity of fire-driven aerosol events [7, 14].

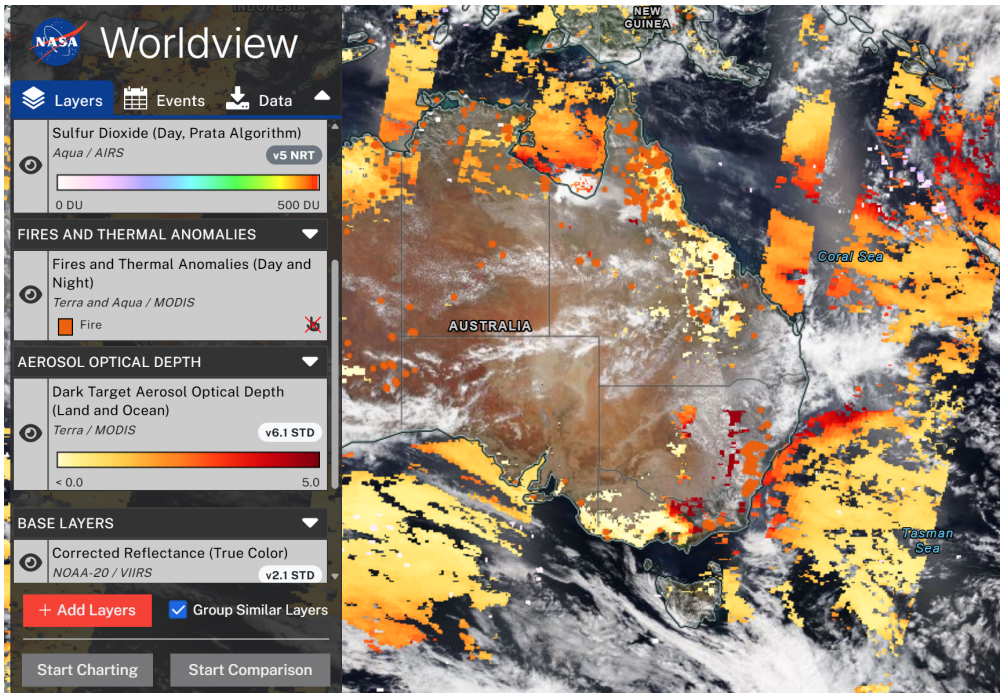

Figure 6.12: NASA Worldview image on 25 December 2019 showing significant gaps in MODIS AOD coverage (e.g., over southeastern Australia), despite visible smoke plumes. These stripes reflect areas where the satellite failed to retrieve AOD data.

MERRA-2 assimilates ground-based observations from the Aerosol Robotic Network (AERONET), a global system of sun photometers operated by NASA that measures atmospheric aerosol properties by observing sunlight extinction [28]. While this assimilation improves MERRA-2's accuracy near AERONET sites, it introduces spatial inconsistencies: Che et al. [7] found that MERRA-2 AOD was approximately 22.5% higher than MODIS-Deep Blue (MODIS-DB) near AERONET stations, but significantly lower in regions without AERONET coverage when AOD exceeded 0.1, particularly during dust events. Furthermore, during severe aerosol outbreaks, MERRA-2 has been shown to underestimate the magnitude and spatial extent of AOD relative to MODIS-DB, partly due to misclassification of thick dust as cloud during assimilation [7]. These findings suggest that while MERRA-2 performs reasonably where ground constraints exist, its reliability diminishes in remote areas—such as southeastern Australia during the bushfires—where satellite retrieval gaps and model smoothing dominate.

Nevertheless, the adjoint-GP model remains capable of handling missing AOD data points, since it only uses the available observations and does not require complete spatial coverage. While this introduces potential bias, especially when high-AOD areas are missed due to cloud cover, the model's probabilistic framework is resilient to such sparsity. Future work could improve accuracy by expanding the spatial domain to encompass thousands of kilometres—consistent with the estimated 1000 km/day advection range of fire plumes—to compensate for these observation gaps.

Therefore, although AOD serves as a valuable observational input in this study, its

limitations reinforce the importance of combining it with meteorological modelling, probabilistic inference, and independent ground-based validation. These observations are consistent with prior findings that highlight the complexity of interpreting AOD measurements due to factors such as cloud interference, viewing geometry, and the vertical distribution of aerosols [20].

6.3 Observed Misalignment Between AOD, Wind, and Fire Activity

Figure 6.13 presents a series of subplots across various timestamps, overlaying AOD distributions, wind vectors, and fire ignition points. These overlays reveal an important observational limitation: the AOD plumes do not follow the expected wind-driven transport direction.

While the modelled wind vectors suggest a clear flow path, the actual aerosol distribution appears offset or misaligned. This discrepancy may reflect the influence of vertical wind shear, unresolved local dynamics, or limitations in the AOD data itself. It highlights the difficulty of relying solely on observational AOD to infer emission sources or transport mechanisms.

These visual inconsistencies reinforce the need for a model-based inference approach that incorporates meteorological dynamics and uncertainty. Figure 6.13 underscores this need by showing that intuitive, wind-following dispersion is not always evident in AOD observations.

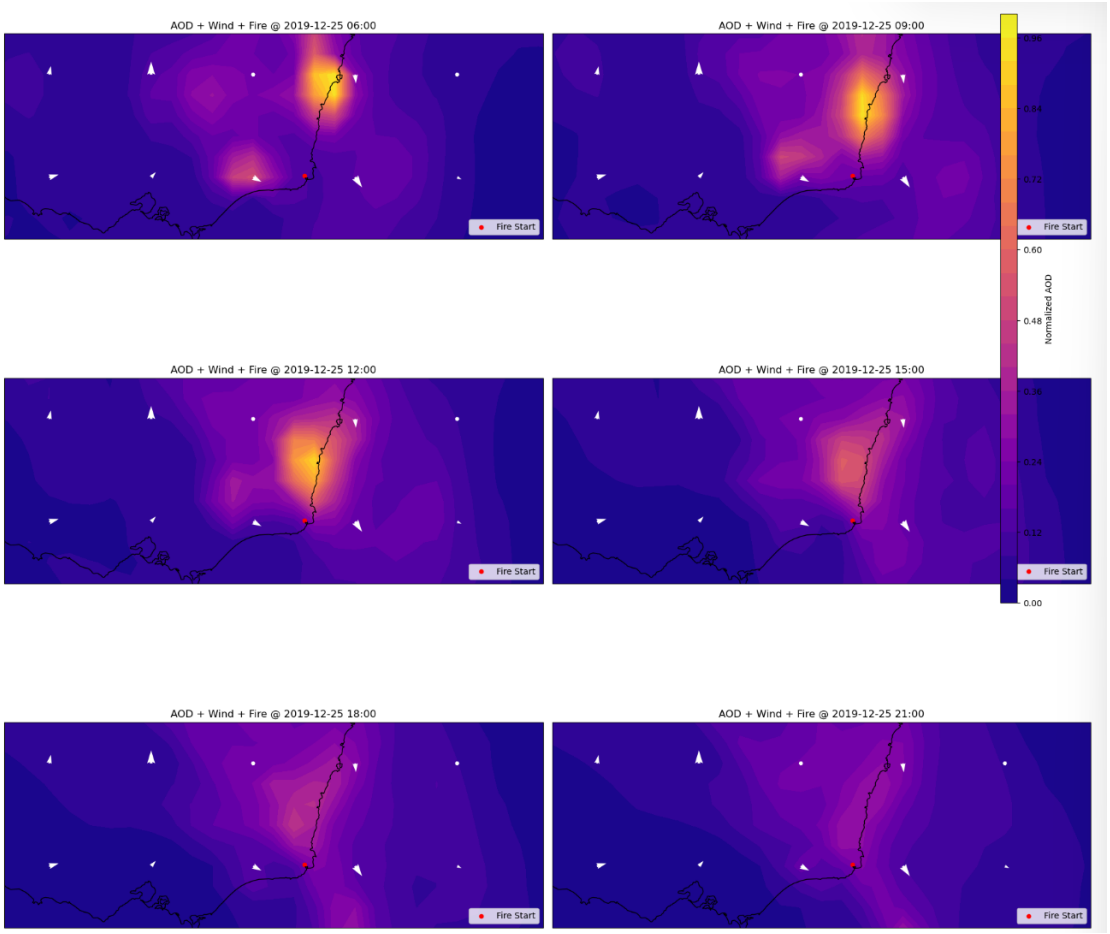

Figure 6.13: Spatial overlays showing AOD distributions, wind vectors, and recorded fire ignition points on 25 December 2019. Each subplot corresponds to a three-hour interval throughout the day, highlighting temporal variation in plume dispersion and atmospheric flow.

It is important to note that the AOD visualised here is from the MERRA-2 reanalysis dataset. This product fills gaps using model-based assimilation, which may smooth or misrepresent the true dispersion patterns seen in raw satellite imagery (e.g., MODIS).

These inconsistencies not only reinforce the limitations of AOD as a standalone validation tool, but also highlight a deeper modelling challenge: determining the appropriate vertical layer for inferring pollution transport. Since AOD reflects column-integrated aerosol loading and wind fields vary with altitude, misalignment between the two complicates source attribution. This issue is explored further in the following section, which examines the vertical layer assumptions and implications of 2D modelling.

6.4 Accuracy of $PM_{2.5}$ as a Pollution Proxy

Aerosol Optical Depth (AOD), while offering broad spatial coverage, remains a column-integrated measure of total aerosol load throughout the atmosphere. This complicates its use as a proxy for ground-level $PM_{2.5}$, especially in regions with limited station coverage like Australia.

A key challenge in this study involved the spatial resolution mismatch between datasets. MERRA-2 AOD is provided at a coarse $1^\circ \times 1^\circ$ resolution, whereas $PM_{2.5}$ stations are often located at sub-degree positions. To enable collocation, station coordinates had to be rounded to the nearest grid cell. This approximation likely introduces spatial noise, particularly in regions with sharp pollution gradients or heterogeneous land cover.

Despite these limitations, $PM_{2.5}$ was used as the modelling target because it better represents near-surface conditions—the atmospheric layer most relevant for health outcomes and transport inference in a two-dimensional model. This choice implicitly avoids elevated aerosols that influence AOD but contribute less to surface-level exposure, supporting a more interpretable and consistent pollution inference framework

6.5 Vertical Layer Considerations and 2D Modelling Assumption

Vertical transport plays a critical role in aerosol dispersion but is simplified in this study by modelling pollution movement in two dimensions (2D). While computationally efficient, this assumption overlooks vertical advection, wind shear, and mixing processes that can significantly affect trajectory paths. Because AOD integrates aerosol content over the entire atmospheric column, it complicates alignment with wind fields extracted at a fixed height, introducing uncertainty into trajectory and source inference.

To highlight this, Figure 6.14 shows a HYSPLIT forward trajectory simulation for 25 December 2019 at a fire site in East Gippsland. Trajectories initialised at 300 m, 1500 m, and 3000 m above ground diverge significantly in their transport pathways over 72 hours, underscoring the importance of vertical positioning in atmospheric modelling.

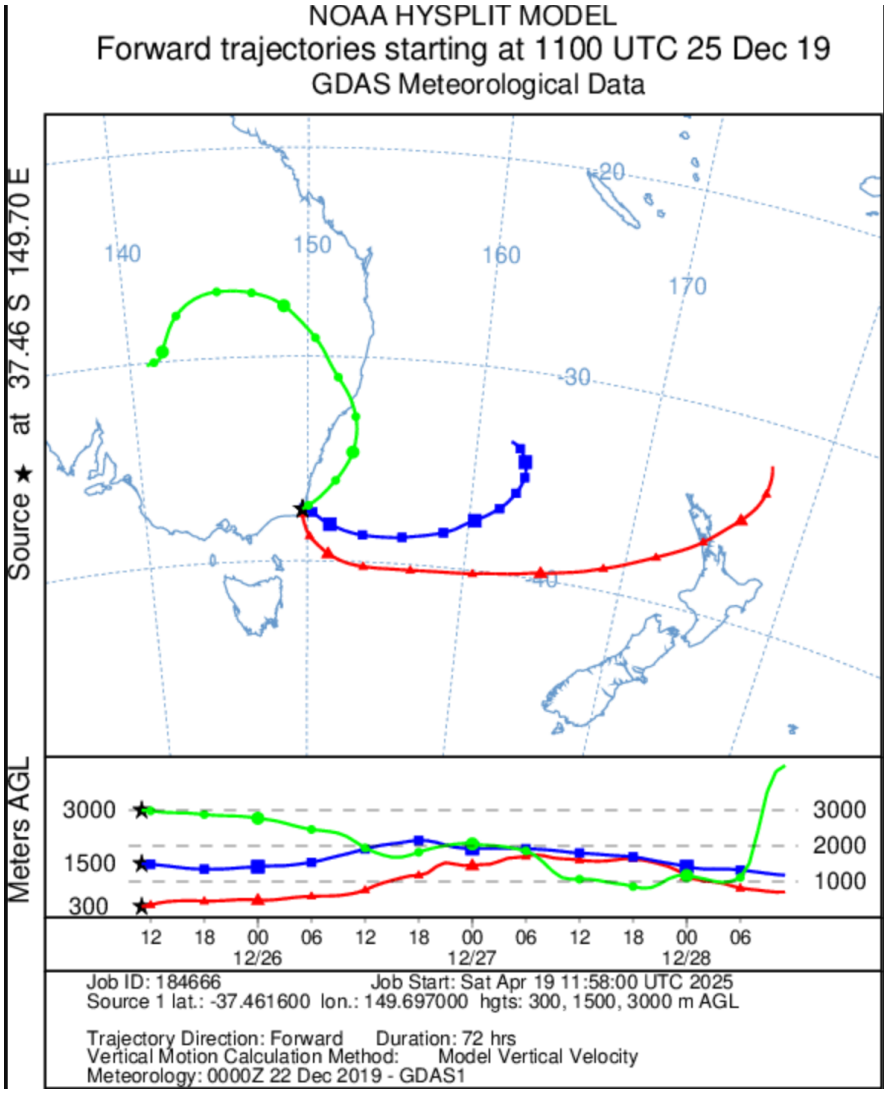

Figure 6.14: HYSPLIT forward trajectories from a fire location in East Gippsland on 25 December 2019 at three altitudes: 300 m (red), 1500 m (blue), 3000 m (green). Each path spans 72 hours, illustrating how dispersion depends on vertical origin.

The results highlight a dramatic divergence in transport direction and distance depending on the initial vertical level. The 300 m trajectory remains near the surface and moves eastward, aligning with local-scale dispersion. In contrast, the 3000 m trajectory veers north-west, likely influenced by upper-level winds. The 1500 m trajectory shows intermediate behaviour, curving south-east before looping. The lower panel of the plot illustrates how vertical motion also varies dynamically over time, with trajectories rising or falling due to atmospheric processes. Besides that, observations from the CALIPSO satellite confirm the presence of fine-mode aerosols in the stratosphere and more spherical particles in the troposphere during the 2019–2020 fires [40]. This stratification underscores the importance of accounting for vertical aerosol profiles in transport models.

To further investigate the effect of vertical wind layer selection on source inference, an additional sensitivity analysis was conducted using three different vertical ranges from the MERRA-2 dataset. The resulting source maps revealed noticeable variation depending on the selected layer, especially in spatial sharpness and alignment with known fire locations. These outcomes are consistent with findings from previous studies that emphasise the importance of vertical wind structure in pollution transport modelling [34]. Full details and visual results of this experiment are provided in Appendix E.0.5.

This divergence illustrates the sensitivity of pollution transport to vertical initialisation, confirming that choosing the wrong wind layer in a 2D model could result in inaccurate source-receptor mapping. Since AOD does not specify altitude, it becomes especially difficult to determine the most appropriate wind layer for transport modelling. As demonstrated in the trajectory divergence (Figure 6.14), pollutant transport is highly sensitive to vertical layer selection. Choosing the right pollutant proxy and corresponding wind layer is therefore critical. Future work should explore fully 3D modelling approaches that incorporate vertical dynamics to improve the realism and accuracy of source inference.

6.6 Limitations and Future Work

While the results of this study demonstrate the feasibility of inferring pollution sources using satellite-derived aerosol data and meteorological fields, several important limitations must be acknowledged:

- **Lack of Direct $PM_{2.5}$ Observations:** Due to sparse ground-based monitoring across Australia, the study relied on AOD as a proxy for $PM_{2.5}$, with a Random Forest model trained for conversion. This introduces uncertainty, as the AOD-to-$PM_{2.5}$ relationship is highly variable, sensitive to atmospheric conditions, and subject to satellite retrieval errors.

- **Vertical Transport and Mixing Effects:** A critical limitation is the omission of vertical transport processes. In reality, pollutants from fires often rise quickly into higher atmospheric layers, undergoing complex mixing. Without modelling this vertical behavior, source inference may misalign with actual transport pathways.

- **Simplifications in 2D Modelling:** For computational tractability, the model assumes two-dimensional (2D) horizontal transport, neglecting dynamic changes in particle altitude. As demonstrated in trajectory simulations, slight changes in initial height can lead to widely divergent plume pathways. Accurately capturing vertical motion remains a key future requirement.

- **Complexity of Fire Events:** The 2019–2020 bushfire season involved widespread, simultaneous ignitions of varying intensity. Representing these diverse sources as a smooth regional field inevitably oversimplifies real-world fire behavior, with smaller or short-lived fires likely underrepresented.

- **Challenges of Scaling Across Time and Space:** Extending the inference over longer periods would require dynamically adjusting the spatial domain to follow moving pollutant plumes. Maintaining high resolution while expanding coverage presents significant computational and modelling challenges.

6.6.1 Future Work Directions

Building on these limitations, several future improvements are proposed:

- **Explicit Vertical Modelling:** Incorporating vertical layers into particle simulations—either through multi-level advection schemes or stochastic vertical mixing models—would greatly improve realism. A fully three-dimensional (3D) framework would better capture fire-driven uplift and stratified transport.
- **Improved Surface Pollution Data:** Expanding the use of ground-based $PM_{2.5}$ measurements or leveraging emerging satellite $PM_{2.5}$ products could enhance the observational constraints. Hybrid data assimilation combining AOD and $PM_{2.5}$ may also be explored.
- **Longer-Term Simulations:** Extending simulations beyond a few days would allow study of pollution episodes at broader timescales. However, this would require adaptive domain management and careful treatment of uncertainties in long-range plume transport.
- **Refined Fire Source Attribution:** Weighting ignition points based on fire size, duration, or detected intensity could better differentiate between major and minor contributors in source inference.
- **Enhanced Uncertainty Quantification:** Future work could employ Bayesian frameworks or ensemble techniques to systematically characterise model uncertainty, providing more robust confidence estimates alongside inferred sources.
- **Larger Inference Domains to Handle Missing Data:** AOD data is often missing in key regions due to cloud cover, and these gaps are not randomly distributed—they frequently coincide with intense fire activity. Although the adjoint-GP model can still operate with partial observations, this introduces bias. Expanding the spatial domain (e.g., thousands of kilometres) could help capture the full extent of pollution transport over daily timescales (up to 1000 km/day), thereby mitigating the impact of missing observations and improving inference reliability.

Finally, although spatial masking and standard deviation overlays were used to indicate inference confidence in this study, future models should explore more advanced spatial uncertainty visualisations.

6.6.2 Summary

Overall, this study highlights the potential of combining satellite observations, meteorological data, and probabilistic models for pollution source estimation during extreme fire events. While the results are promising, they also expose critical challenges—particularly the need for vertical transport modelling and improved observation proxies. Addressing these gaps presents important opportunities for enhancing the realism, accuracy, and robustness of future pollution source inference frameworks.

6.7 Final Reflection

This project marks the first real-world application of the Advection Gaussian Process (Advection GP) framework using satellite and meteorological data from an actual environmental event, 2019–2020 Australian bushfires. Prior studies have largely relied on synthetic datasets, where system behaviour can be precisely controlled—wind fields can be made ideal, sources simplified, and vertical dynamics conveniently ignored. In contrast, working with real data revealed the full complexity of the problem: uncertainties in wind direction, gaps in satellite observations, challenges in vertical alignment, and limitations in observational proxies.

Translating a mathematically rich model—originally designed for controlled settings—into a functional tool for real atmospheric events was far from trivial. It demanded not only technical integration of disparate datasets, but also careful judgement in selecting assumptions, resolving spatial mismatches, and validating inferences. Despite these complexities, the model was successfully implemented, adapted, and validated on real AOD and meteorological data—demonstrating the framework's potential and setting a foundation for future work.

While there remains significant work to be done, this project represents a meaningful first step toward bridging data-driven inference and real-world environmental monitoring. This research not only bridges the gap between theory and application, but also marks a meaningful step toward scalable, data-driven air pollution source attribution with global relevance.

If further developed and validated, such tools could play a vital role in disaster response, air quality forecasting, and environmental policymaking. The potential societal impact is enormous—and this project proves that the first steps are not only possible, but within reach.

Bibliography

[1] Coordinate system change: Wgs 84 / utm zone 56s (epsg:32756), 2024.

[2] Epsg geodetic parameter dataset, 2024.

[3] Albani, R. A., Albani, V. V., Migon, H. S., and Neto, A. J. S. Uncertainty quantification and atmospheric source estimation with a discrepancy-based and a state-dependent adaptative MCMC. *Environmental Pollution 290* (2021), 118039.

[4] Arellano Jr, A., Raeder, K., Anderson, J., Hess, P., Emmons, L., Edwards, D., Pfister, G., Campos, T., and Sachse, G. Evaluating model performance of an ensemble-based chemical data assimilation system during INTEX-B field mission. *Atmospheric Chemistry and Physics 7*, 21 (2007), 5695–5710.

[5] Borysiewicz, M., Wawrzynczak, A., and Kopka, P. Bayesian-based methods for the estimation of the unknown model's parameters in the case of the localization of the atmospheric contamination source. *Foundations of Computing and Decision Sciences 37*, 4 (2012), 253.

[6] Bosilovich, M. G., Lucchesi, R., and Suarez, M. Merra-2: File specification. Tech. Rep. Office Note No. 9 (Version 1.1), NASA Global Modeling and Assimilation Office (GMAO), 2016. 73 pp.

[7] Che, Y., Yu, B., Parsons, K., Desha, C., and Ramezani, M. Evaluation and comparison of merra-2 aod and daod with modis deepblue and aeronet data in australia. *Atmospheric Environment 277* (2022), 119054.

[8] Department of Environment, Land, Water and Planning, Victoria. Fire origins - current and historical (firehistoryorigin). `https://datashare.maps.vic.gov.au/search?md=4e35399b-e558-5d34-95a0-87a7f24a6096`, 2019. Accessed: 11 April 2025.

[9] DiBiase, D., Sloan II, J. L., Baxter, R., Stroh, W., King, B. F., and many students. *The Nature of Geographic Information*. The Pennsylvania State University, 2006. [Chapter 2, Page 22].

[10] Ding, G., Chan, C., Gao, Z., Yao, W., Li, Y., Cheng, X., Meng, Z., Yu, H., Wong, K., Wang, S., and Miao, Q. Vertical structures of pm10 and pm2.5 and their dynamical character in low atmosphere in beijing urban areas. *Science in China, Series D: Earth Sciences 48* (2005), 38–54.

[11] Encyclopædia Britannica, I. Cartography, 2024.

[12] Environment Protection Authority Victoria. Epa airwatch: 2019 all sites air quality hourly averages. `https://apps.epa.vic.gov.au/datavic/Data_Vic/AirWatch/2019_All_sites_air_quality_hourly_avg_AIR-I-F-V-VH-O-S1-DB-M2-4-0.xlsx`, 2021. Accessed: 11 April 2025.

[13] Gahungu, P., Lanyon, C. W., Alvarez, M. A., Bainomugisha, E., Smith, M., and Wilkinson, R. D. Adjoint-aided inference of gaussian process driven differential equations.

[14] Gelaro, R., McCarty, W., Suárez, M. J., et al. The modern-era retrospective analysis for research and applications, version 2 (merra-2). *Journal of Climate 30*, 14 (2017), 5419–5454.

[15] Geography, G. Map distortion with tissot's indicatrix, 2015.

[16] Global Modeling and Assimilation Office (GMAO). Merra-2 inst3_2d_gas_nx: 2d, 3-hourly, instantaneous, single-level, assimilation, aerosol optical depth analysis v5.12.4. `https://doi.org/10.5067/HNGAOEWW0R09`, 2015. NASA Goddard Earth Sciences Data and Information Services Center (GES DISC), Greenbelt, MD, USA. Accessed: 2025-04-11.

[17] Global Modeling and Assimilation Office (GMAO). Merra-2 tavg3_3d_asm_nv: 3d, 3-hourly, time-averaged, model-level, assimilation, assimilated meteorological fields v5.12.4. `https://doi.org/10.5067/SUOQESM06LPK`, 2015. NASA Goddard Earth Sciences Data and Information Services Center (GES DISC), Greenbelt, MD, USA. Accessed: 2025-04-11.

[18] Handschuh, J., Erbertseder, T., and Baier, F. On the added value of satellite aod for the investigation of ground-level pm2.5 variability. *Atmospheric Environment 331* (2024), 120601.

[19] Heald, C. L., Jacob, D. J., Park, R. J., Alexander, B., Fairlie, T. D., Yantosca, R. M., and Chu, D. A. Transpacific transport of asian anthropogenic aerosols and its impact on surface air quality in the united states. *Journal of Geophysical Research Atmospheres 111* (7 2006).

[20] Holben, B. N., Eck, T. F., Slutsker, I., Tanré, D., Buis, J. P., Setzer, A., Vermote, E., Reagan, J. A., Kaufman, Y. J., Nakajima, T., Lavenu, F.,

Jankowiak, I., and Smirnov, A. Aeronet—a federated instrument network and data archive for aerosol characterization. *Remote Sensing of Environment 66*, 1 (1998), 1–16.

[21] Hwang, Y., Kim, H. J., Chang, W., Yeo, K., and Kim, Y. Bayesian pollution source identification via an inverse physics model. *Computational Statistics & Data Analysis 134* (2019), 76–92.

[22] Jacob, D. J. Heterogeneous chemistry and tropospheric ozone, 2000.

[23] Kopacz, M., Jacob, D. J., Henze, D. K., Heald, C. L., Streets, D. G., and Zhang, Q. Comparison of adjoint and analytical Bayesian inversion methods for constraining Asian sources of carbon monoxide using satellite (MOPITT) measurements of CO columns. *Journal of Geophysical Research: Atmospheres 114*, D4 (2009).

[24] Kumar, N. What can affect aod–pm2.5 association? *Environmental Health Perspectives 118*, 3 (Mar. 2010), A109–A110. Open access; NIH Public Domain.

[25] Liang, Q., Thompson, A. M., Jacob, D. J., Crawford, J. H., Chen, H.-R., Heikes, B. G., Sachse, G. W., Bradshaw, G. W., Diskin, G. S., Hall, S. R., and Shetter, R. R. Summertime influence of asian pollution in the free troposphere over north america. *Journal of Geophysical Research: Atmospheres 109*, D23 (2004), D23S11.

[26] Mehta, M., Singh, N., and Anshumali. Global trends of columnar and vertically distributed properties of aerosols with emphasis on dust, polluted dust and smoke - inferences from 10-year long caliop observations. *Remote Sensing of Environment 208* (4 2018), 120–132.

[27] NASA EOSDIS Worldview. NASA Worldview Snapshots - 25 December 2019. https://worldview.earthdata.nasa.gov/?v=104.48463967386093,-56.65052327606113,206.12995852287995,-5.051406554788279&l=Reference_Labels_15m,Reference_Features_15m,Coastlines_15m,AIRS_Prata_SO2_Index_Day,MODIS_Combined_Thermal_Anomalies_All,MODIS_Terra_Aerosol,VIIRS_NOAA20_CorrectedReflectance_TrueColor,VIIRS_SNPP_CorrectedReflectance_TrueColor(hidden),MODIS_Aqua_CorrectedReflectance_TrueColor(hidden),MODIS_Terra_CorrectedReflectance_TrueColor(hidden)&lg=true&t=2019-12-25-T10:00:00Z, 2019. Accessed: 29 April 2025.

[28] NASA Goddard Space Flight Center. AERONET - Aerosol Robotic Network, 2024. Accessed: 29 April 2025.

[29] NOAA STAR. Goes-r aerosol optical depth (aod). https://www.star.nesdis.noaa.gov/goesr/rework/product_aero_aod.php, 2023. Accessed 9 May 2025.

[30] Plessix, R.-E. A review of the adjoint-state method for computing the gradient of a functional with geophysical applications. *Geophysical Journal International 167*, 2 (2006), 495–503.

[31] Pudykiewicz, J. A. Application of adjoint tracer transport equations for evaluating source parameters. *Atmospheric environment 32*, 17 (1998), 3039–3050.

[32] Rasmussen, C. E., and Williams, C. K. I. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

[33] Salleh, A. Bushfire frequency has increased by 40 per cent over five years, scientists say, 2016. Accessed: 2025-04-29.

[34] Seinfeld, J. H. *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*. John Wiley & Sons, Incorporated, Newark, 2016. Available from: ProQuest Ebook Central. [Accessed 1 November 2024].

[35] SheffieldML. Advectiongp: Gaussian process framework for advection-diffusion. `https://github.com/SheffieldML/advectionGP`, 2021. Accessed: 2025-04-11.

[36] State Government of Victoria. Datavic - about datavic, 2024. Accessed: 11 April 2025.

[37] Tian, Z., Wei, J., and Li, Z. How important is satellite-retrieved aerosol optical depth in deriving surface pm2.5 using machine learning? *Remote Sensing 15*, 15 (2023), 3780. Open access under CC BY license.

[38] U.S. Environmental Protection Agency. Particulate matter (pm) basics. `https://www.epa.gov/pm-pollution/particulate-matter-pm-basics`, 2024. Accessed: 2025-05-09.

[39] World Health Organization. Ambient (outdoor) air quality and health. `https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health`, 2024. [Accessed 1 November 2024].

[40] Wu, D., Zeng, N., Zhai, S., Li, X., and He, J. Australian bushfires (2019–2020): Aerosol optical properties and radiative forcing. *Atmosphere 13*, 6 (2022), 867.

[41] Yee, E. Theory for reconstruction of an unknown number of contaminant sources using probabilistic inference. *Boundary-layer meteorology 127*, 3 (2008), 359–394.

Appendices

Appendix A: Wind Model Implementations

.1 RealWindNearestNeighbour: Full Memory Table with Nearest Spatial Query

This was the initial prototype implementation for integrating real meteorological wind data into the AdvectionGP framework. It loads wind vectors from MERRA-2 files, averages them over selected vertical layers, and stores all spatiotemporal points in a single Pandas DataFrame. A KD-tree is then constructed for fast spatial lookup.

Design

- All wind data points are stored in memory, allowing fast nearest-neighbour lookup via KD-tree.
- Timestamp matching is performed using a brute-force search for the closest time entry.

Observations This approach is straightforward and very fast for small numbers of particles. However, because each time lookup scans the full timestamp with no binning, and each spatial query hits a single KD-tree built over all times, it does not scale well: once you trace thousands of particles over many time-steps, the repeated scans become a bottleneck especially in the forward simulations.

Limitations

- Poor scalability with particle count and simulation time.
- Repeated timestamp lookups are costly.
- The lack of temporal indexing means each `getwind()` call performs an $\mathcal{O}(N)$ search over timestamps.
- The single KD-tree over all times leads to poor locality and cache performance when particles cover many days.

.2 RealWindBinned: Timestamp Binning for Fast Temporal Lookup

To improve temporal lookup speed, this version grouped wind data into fixed 3-hour time bins (matching the MERRA-2 sampling interval). Each timestamp had its own KD-tree for spatial queries.

Design

- Wind vectors are grouped into dictionaries indexed by timestamp.
- KD-tree is built once per time bin.
- During lookup, the nearest time bin is selected, and spatial coordinates are queried.

Benefits

- Faster than the brute-force timestamp search used in `RealWindNearestNeighbour`.
- Scales better for long time durations.

Issues Encountered During forward simulation, particles with timestamps outside the indexed range resulted in NaNs (missing wind vectors), triggering runtime errors in ‘computeConcentration()‘.

Outcome The model showed excellent runtime improvements but was sensitive to incomplete timestamp coverage. This motivated the development of an interpolated version.

.3 RealWindHybrid: Time Interpolation Between Binned Snapshots

uilding upon the limitations of the Binned KD-tree approach, the Hybrid method sought to address missing data issues, this version interpolated wind vectors across adjacent time bins. If a particle’s timestamp lay between two bins, a linear blend of the two nearest wind values was returned.

Design

- Performs temporal interpolation between two timestamps.
- Uses spatial KD-trees from `RealWindBinned` for spatial queries.

Benefits

- Prevents NaNs by ensuring every particle has a wind value.
- Produces smoother particle trajectories.

Drawbacks

- Significantly increases runtime due to double KD-tree queries per point.
- Not feasible for large-scale testing or multi-day simulations.

.4 FastWindGrid: Direct Grid Indexing on Precomputed Wind Cube

This method was developed to maximise runtime performance. Instead of KD-trees, wind data are preloaded into structured 3D NumPy arrays with dimensions `[time, y, x]`. Wind lookup is performed via direct index conversion from particle coordinates.

Design

- Assumes a regular grid in UTM space.
- Projects all lat/lon to UTM once and saves `x_vals`, `y_vals`.
- Wind vectors are stored as `wind_u[time, y, x]` and `wind_v[time, y, x]` arrays.

Performance

- Fastest among all implementations.
- Enables scalable simulation with minimal memory or computational overhead.

Limitations

- Less flexible—requires uniform grid assumption.
- Cannot dynamically expand or interpolate off-grid.

Appendix B: Wind Lookup Method: Computational Complexity

RealWindNearestNeighbour

- **KD-tree spatial query:** $\mathcal{O}(\log M)$, where M is the total number of wind points.
- **Time lookup:** Brute-force search through all timestamps $\Rightarrow \mathcal{O}(M)$ in worst-case.
- **Total per query:** $\mathcal{O}(M + \log M)$.
- **Total per `getwind()`:** $\mathcal{O}(N \cdot (M + \log M)) \approx \mathcal{O}(NM)$ for large M.

RealWindBinned

Computational Complexity

- **Pre-binned timestamps:** Each 3-hour bin contains a smaller subset K.
- **KD-tree spatial query:** $\mathcal{O}(\log K)$ within the bin.
- **Time bin lookup:** Dictionary access in $\mathcal{O}(1)$.
- **Total per query:** $\mathcal{O}(\log K)$.
- **Total per `getwind()`:** $\mathcal{O}(N \cdot \log K)$.

RealWindHybrid

Computational Complexity

- **Interpolated across time:** Finds both lower and upper time bins.
- **Time search:** $\mathcal{O}(T)$, where T is the number of time bins.

- **KD-tree queries:** $2 \cdot \mathcal{O}(\log K)$ (one per time bin).
- **Interpolation:** $\mathcal{O}(1)$.
- **Total per query:** $\mathcal{O}(T + \log K)$.
- **Total per `getwind()`:** $\mathcal{O}(N \cdot (T + \log K)) \approx \mathcal{O}(NT)$.

FastWindGrid

- **Fixed grid lookup:** Coordinates directly mapped to array indices.
- **Access time:** $\mathcal{O}(1)$.
- **Total per query:** $\mathcal{O}(1)$.
- **Total per `getwind()`:** $\mathcal{O}(N)$.

Appendix C: Data Retrieval Challenges

.1 Accessing MERRA-2 Data from NASA Earthdata

The MERRA-2 datasets used in this study are hosted by NASA's Goddard Earth Sciences Data and Information Services Center (GES DISC), which requires authentication through the Earthdata Login system.

Initial attempts were made to retrieve the data programmatically using Earthdata's API and the `earthaccess` Python library. While this approach offers automation, reproducibility, and integration with Jupyter workflows, several challenges arose:

- **Credential Management:** Each session required re-authentication or secure token management, complicating unattended batch downloads.
- **Performance Constraints:** The large file sizes (NetCDF format, 100–300 MB per file) led to slow response times and interrupted downloads during peak hours.
- **Incomplete Retrieval:** Multi-day downloads often failed mid-session due to unstable connections or timeout errors, resulting in incomplete datasets.

Due to these limitations, a manual download strategy was adopted to ensure data completeness and continuity.

.2 Manual Download Workflow

Instead of relying on automated scripts, all required NetCDF files were downloaded individually from the GES DISC data portal. Although this approach lacks scalability, it ensured the following:

- **Complete Dataset Acquisition:** All selected AOD and wind files were verified manually for integrity and completeness.
- **Simplified Integration:** Files could be easily indexed, renamed, and loaded in Jupyter Notebooks without re-authentication.

- **Consistency Across Runs:** Manual control prevented unintentional overwriting or inconsistent sampling that can occur with parallel scripts.

While not ideal for large-scale studies, this approach was suitable for the scope and timeline of this project. Future work could revisit automated retrieval pipelines using tools such as `earthaccess`, `wget`, or `cURL` with persistent login tokens or '.netrc' authentication files.

Appendix : Map Projection Comparison

.0.1 Projection Selection and Evaluation

Representing Earth's three-dimensional curved surface on two dimensional maps is crucial in choosing the accurate map projections. Latitude and longitude defines points on the curve surface of the Earth but doesn't represent distance on a flat surface. The three Projections that I have explored were:

- Albers Equal Area Projection
- Mercator Projection
- UTM

Comparisons of Projections

Accurate map projection is essential for minimizing spatial distortion in pollutant dispersion analysis. Three projections were evaluated for this project:

- **Albers Equal Area Projection:** This projection is widely used for continent-wide studies due to its preservation of area. However, it introduces significant shape and distance distortions near the poles and outside standard parallels, making it less suitable for regional-scale analyses such as Australia [15].

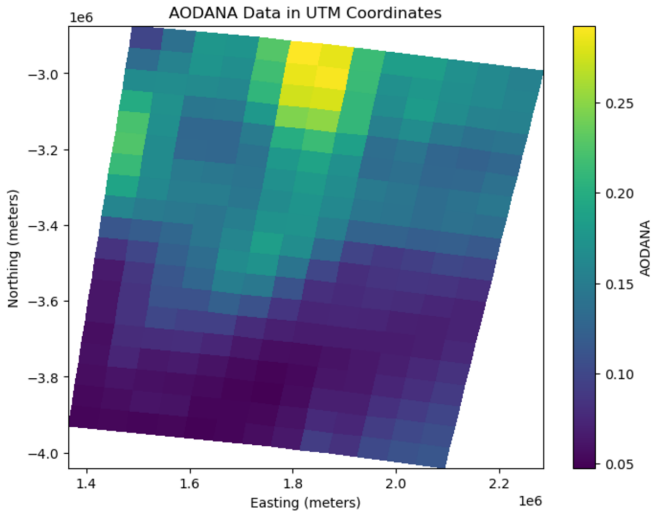

Figure 15: AOD presented using Albers Equal Area Projection.

- **Mercator Projection:** While it preserves compass bearings and is commonly used in navigation, it significantly distorts areas farther from the equator. This limits its accuracy for mid-latitude regions such as Australia [11].

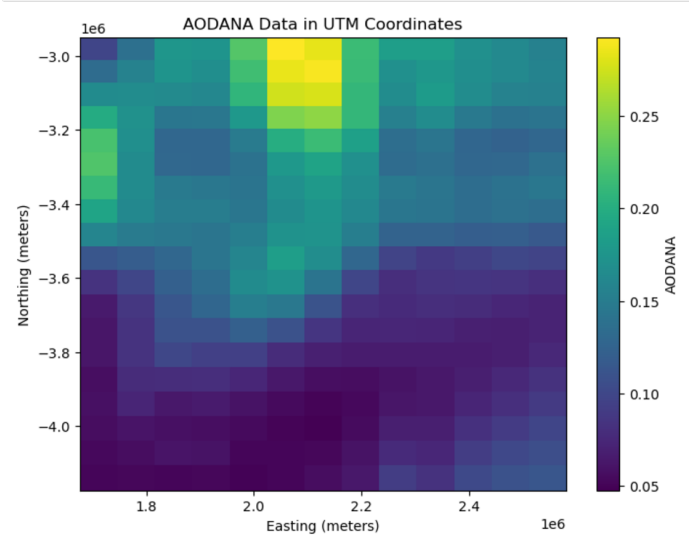

Figure 16: AOD presented using Mercator Projection.

- **Universal Transverse Mercator (UTM) Projection:** UTM maintains local distance and directional accuracy within zones, with minimal distortion (error less than ±1m per 1000m) [9]. Zone 56S, which covers eastern Australia, is especially suited for this study.

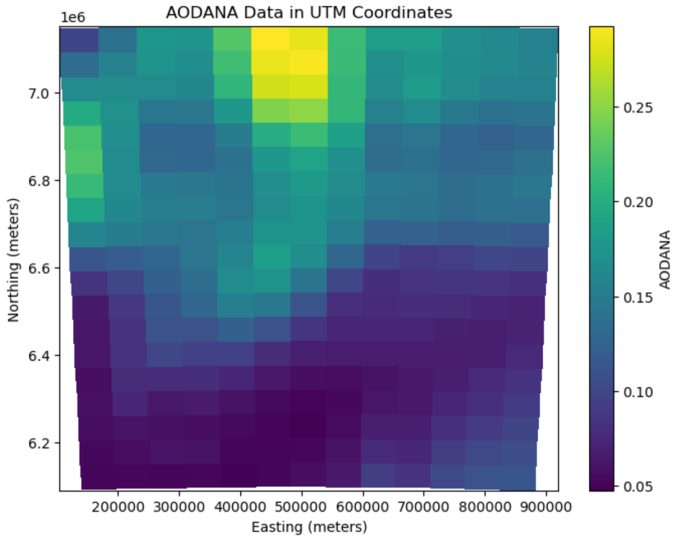

Figure 17: AOD presented using UTM Projection.

A side-by-side comparison of the projections is summarized in Table 1. Based on this evaluation, the UTM projection was selected for its superior accuracy in local-scale analysis and its compatibility with the bushfire-affected regions of New South Wales.

Table: Comparison of Map Projections

Table 1: Comparison of Map Projections for Pollutant Dispersion Analysis

Projection	**Advantages**	**Disadvantages**
Albers Equal Area	Preserves area; ideal for large regions such as continents.	Shape and distance distortion increases away from standard parallels [15].
Mercator Projection	Preserves compass bearings; suitable for navigation.	Significant area distortion away from the equator; not accurate for Australia [11].
UTM Projection	Preserves distance and direction locally; minimal distortion in zones.	Accuracy decreases outside the defined UTM zone [9].

Selected Projection: UTM Zone 56S The UTM Zone 56S projection was identified using the EPSG Geodetic Parameter Dataset [1, 2], which provides comprehensive information on coordinate systems and transformations. This zone was chosen for its ability to preserve spatial accuracy in the eastern regions of Australia most affected by the 2019–20 bushfires. The map in Figure 17 was generated using the EPSG platform's visualisation tool, powered by MapTiler [2].

Appendix E: AOD to $PM_{2.5}$ Conversion Pipeline

E.0.2 Feature Construction

Using spatial KD-tree nearest neighbour search and temporal alignment (e.g., rounding to the nearest hour), the three datasets were matched to form a unified table called `matched_df`. Each row represents a spatiotemporal observation point, with the following columns:

- `AOD`, `latitude`, `longitude`
- `U`, `V`, `wind_speed`
- `temperature(temp)`, `relative humidity(rh)`, `surface pressure(ps)`
- $PM_{2.5}$ (target variable)

This feature-rich table enabled the training of a machine learning model to estimate surface-level $PM_{2.5}$ from satellite and meteorological inputs. Various combinations of features were tested to identify the most meaningful predictors, discussed further in the following sections.

E.0.3 Choice of Features and Model Testing

An initial feature set was constructed, including AOD, latitude, longitude, wind components (U, V), wind speed, temperature(temp), and relative humidity (rh) , surface pressure(ps). A correlation heatmap (Figure 18) was used to understand the relationships between input features and $PM_{2.5}$.

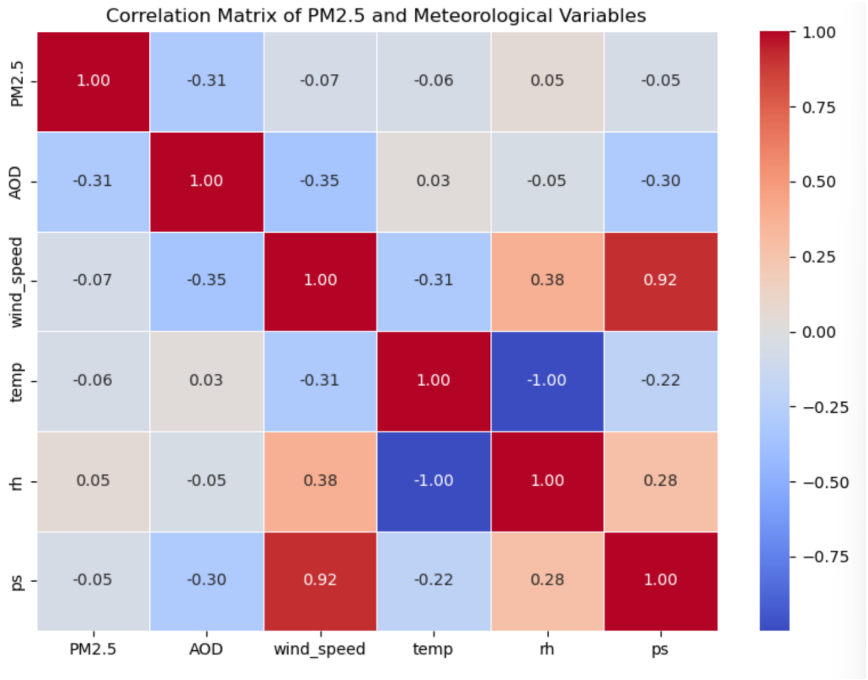

Figure 18: Correlation heatmap between input features and $PM_{2.5}$.

Following this, feature importance from the trained Random Forest model helped refine the feature set (see Figure 19).However, simpler models using only AOD, latitude, and longitude also produced reasonable performance.

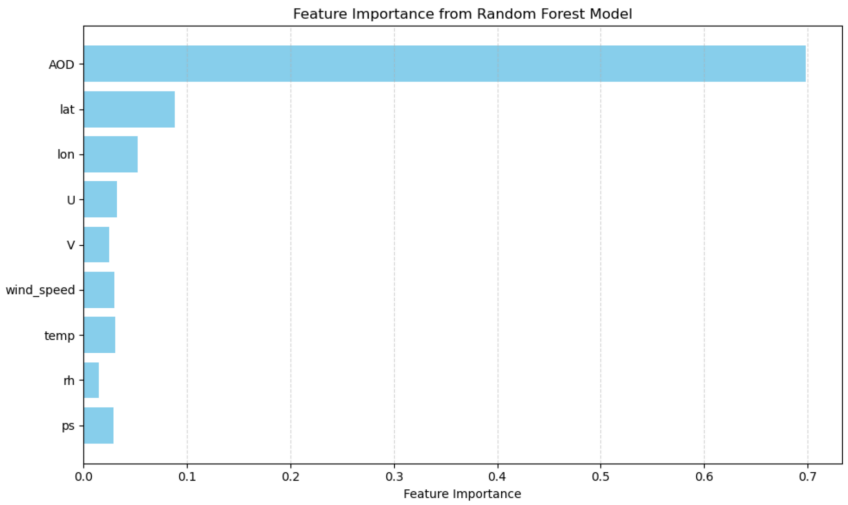

Figure 19: Feature importance values from the Random Forest model.

E.0.4 Model Performance

The final model was evaluated on an 80/20 train-test split, yielding a coefficient of determination (R^2) of 0.69. This indicates a good degree of predictive accuracy and suggests that the model was effective at approximating surface-level $PM_{2.5}$ concentrations from AOD and supporting meteorological variables. Alternative machine learning models such as Gradient Boosting or Neural Networks could further improve predictive accuracy at the cost of higher complexity.

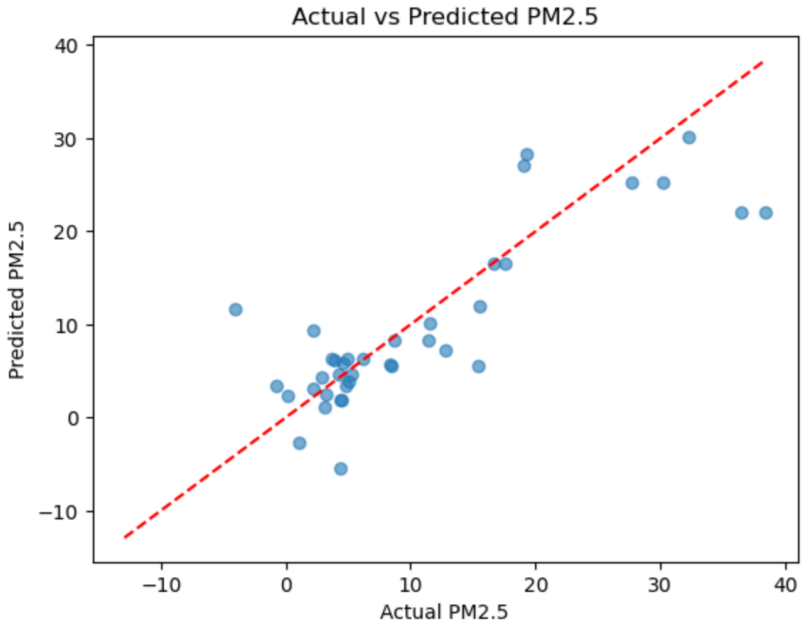

Figure 20: The perfomance of actual vs predicted $PM_{2.5}$

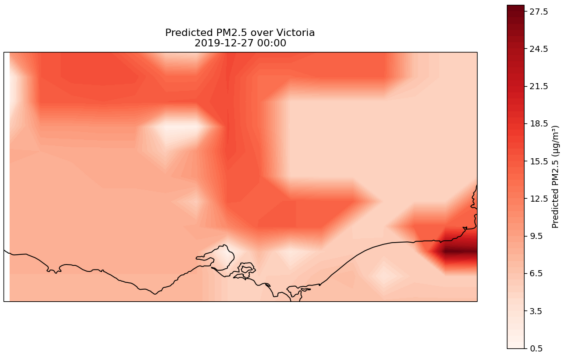

Figure 21: Spatial distribution of inferred $PM_{2.5}$ concentrations across Victoria on 27 Dec 2019

E.0.5 Integration into the Inference Framework

Once the model was trained, predicted $PM_{2.5}$ values were computed across the spatial grid of Victoria. These predictions were then used as the observation vector Y for the source inference model.

This transformation allowed the model to make use of satellite-derived data in a more physically interpretable manner, bringing it closer to actual ground-level pollution exposure. This transformation allowed the model to make use of satellite-derived data in a more physically interpretable manner, bringing it closer to actual ground-level pollution exposure.

Appendix F: Sensitivity Analysis: Effect of Vertical Layer Selection

Initial results revealed that the inferred sources did not always align well with known fire locations. This prompted an investigation into the impact of vertical wind layer selection—an important consideration given that the model operates in two dimensions with a fixed vertical layer. In real atmospheric conditions, wind direction and speed can vary considerably with height, influencing pollutant transport trajectories.

To quantify this impact, source inference was repeated using three vertical wind ranges extracted from the MERRA-2 dataset (Table 2). Previous studies have shown that $PM_{2.5}$ is typically concentrated near the surface, particularly within the first 500 m of the boundary layer [34].

Table 2: Mapping of vertical height ranges to MERRA-2 native levels.

Height Range	**MERRA-2 Levels**	**Approximate Pressure (hPa)**
0–500 m	Levels 72–70	985–955
500–1000 m	Levels 69–67	940–910
Above 1000 m	Levels 66 and below	$\leq$895

Three scenarios were tested:

1. Wind averaged over 0–500 m (Levels 72–70),
2. Wind averaged over 500–1000 m (Levels 69–67),
3. Wind averaged over 0–1000 m (Levels 72–67).

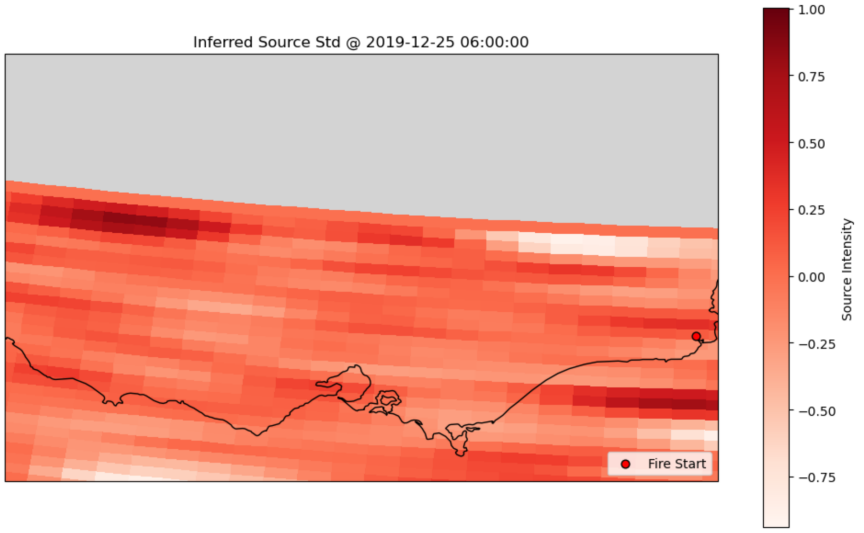

Figure 22: Inferred source distribution using 0–500 m wind.

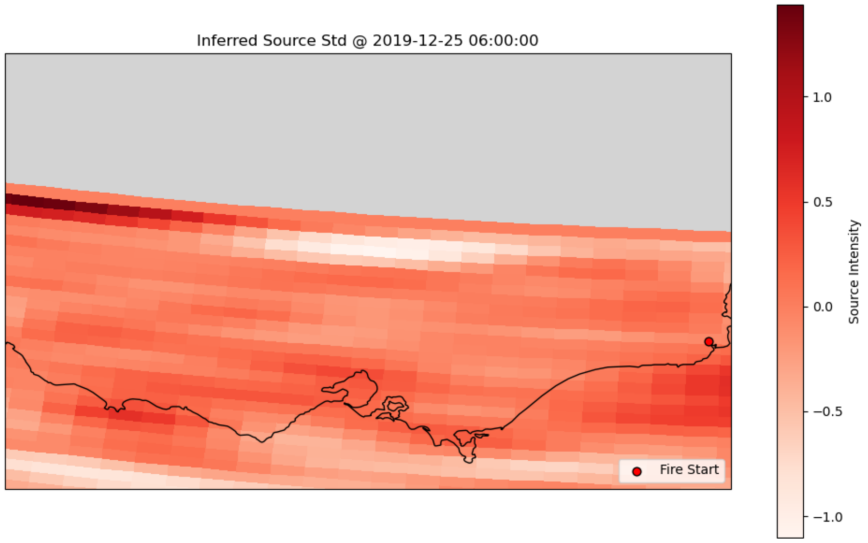

Figure 23: Inferred source distribution using 500–1000 m wind.

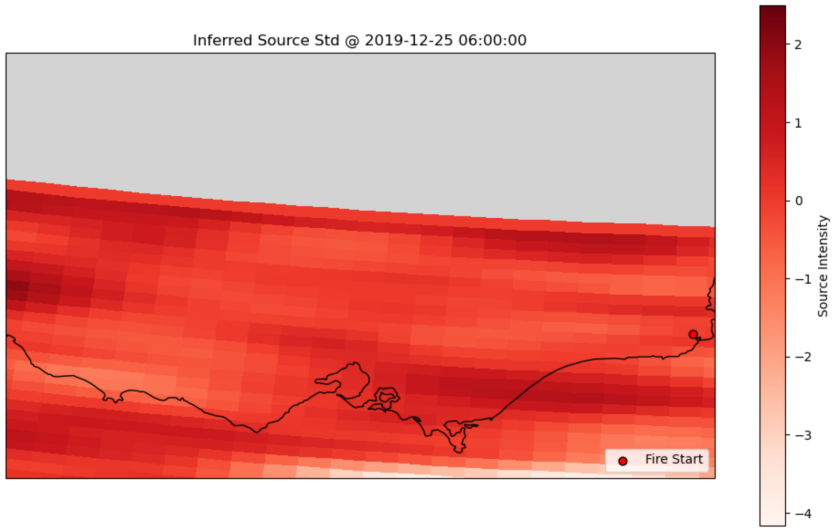

Figure 24: Inferred source distribution using 0–1000 m wind.

Results indicate that:

- **0–500 m wind:** Produces sharper, more localised sources aligned with ignition points but may be more sensitive to noise.
- **500–1000 m wind:** Leads to smoother, more diffuse source maps, indicating higher-altitude, long-range transport.
- **0–1000 m wind:** Provides a balanced view but may blur finer structures due to averaging across layers.

Based on these outcomes, the 0–500 m range was selected for further analysis, as it most closely matched expectations for ground-level fire-driven emissions and improved the fidelity of source attribution.

This experiment highlights one of the key limitations of using a fixed-height, 2D model: source inference is highly sensitive to vertical layer selection. In future, dynamic height modelling or full 3D frameworks may offer more accurate representations of pollution dispersion.