Karlsruhe Institute of Technology

Advection acceleration with Machine Learning Enabled Species Compression

Master's Thesis of

Tai, Pak Yuen

At the Institute for Meteorology and Climate Research
IMKTRO – Troposphere Research

First examiner:	Prof. Dr. Corinna Hoose
Second examiner:	TT-Prof. Dr. Peer Nowack
First advisor:	Dr. Ali Hoshiyaripour
Second advisor:	Dr. Charlotte Debus

03 February 2024 – 07 May 2026

Institut für Meteorologie und Klimaforschung
Hermann-von-Helmholtz-Platz 1
Gebäude 435
76344 Eggenstein-Leopoldshafen

Advection acceleration with Machine Learning Enabled Species Compression (Master's Thesis)

I declare that I have developed and written the enclosed thesis completely by myself. I have not used any other than the aids that I have mentioned. I have marked all parts of the thesis that I have included from referenced literature, either in their original wording or paraphrasing their contents. I have followed the by-laws to implement scientific integrity at KIT.

Karlsruhe, 07 May 2026

. . . . . .

(Tai, Pak Yuen)

Abstract

Prognostic tracer advection in Numerical Weather Prediction scales linearly with increasing number of species, resulting in major bottleneck in Numerical Weather Prediction simulation involving high number of tracer species, e.g. Atmospheric Chemistry simulation where 10s-100s tracers must be advected independently at each advection time step. This thesis investigates whether a physically consistent reduced order representation can serve as a computationally beneficial surrogate for the tracer transport operator in the ICON-ART model. By using mineral dust experiment–a standard test suite in ICON-ART–as a pilot study, 4 embedding models with increasing level of complexity are developed and evaluated in an offline training & evaluation framework: baseline Principle Component Analysis (PCA), Non-negative Matrix Factorisation (NMF), a Gradient based PCA (1 layer Autoencoder with no activation function), and a novel Multi-head Autoencoder that decouples magnitude construction from zero/non-zero classification through a split Softplus and Sigmoid decoder heads, broadly expanding the dynamic range of the output and ensuring non-zero, well defined gradient across the entire interval of output.

PCA achieves the lowest reconstruction error (total normalised RMSE $\sim$ 1% at 4 latent dimensions) with exact idempotency under iteration. However, the mean centred property, unconstrained basis vector rotation and the absence of non-negativity constraint, making it unsuitable for operational deployment, only serving as a theoretical lower bound of degrees of freedom of the manifold. NMF preserves non-negativity by construction, underestimates concentration with high absolute magnitude while overestimates the counterpart, denoting a "drift to centre" property. It suffers from a zero-variable problem, whereby near-zero/zero concentrations exhibited strong variance (ghost mass) for species with high absolute concentration, warranting the need for a zero/non-zero classifier head. Gradient PCA fails completely in lower absolute concentration, where mass concentration dust A, $O(1-6)$ smaller than other species, exhibiting strong variance (up to 70x to the species magnitude), further intensify the need for a classifier head. Conventional Autoencoder examined in the ablation study, collapses to a narrow output band. Close examination of the gradient of the Autoencoder revealed that this behaviour is structurally determined, as vanishing gradient problem occurs at $|h| \approx 8.67$ in float32 precision.

The proposed Multi-head Autoencoder addresses the limitations by employing an additional non-zero head (Sigmoid, interval $(0, 1)$), in addition to the magnitude head (Softplus, interval $(0, \infty)$) , whose element-wise product extends the effective output dynamic range. Gradient analysis demonstrates that the 3 backpropagation pathways operation in complementary regime: the BCE classifier dominates at zero elements, preserving classification accuracy, while the magnitude head (MSE Loss) dominates at non-zero elements, preserving concentration fidelity. The model achieves strong single-pass reconstruction for Dust B-C

(higher absolute magnitude pair, $R^2 = 0.82 - 0.91$, $nRMSE \sim 10 - 14\%$), and optimal manifold preservation (total ACC=0.9986, normalised Wasserstein with IQR = 0.0035). However, the iterative stability limits its practicality at the current state: total mass degrades to 80.7% after 100 offline iterations, possibly driven by the asymptotic Sigmoid gate ($\sigma < 1$ by construction) and unbounded error growth. Error in the FULL MODEL testing dataset sharply collocates with connectively active regions under spatial analysis. This is attributed to the training strategy, where a no-source-sink (convective and other source and sink terms are turned off) dataset are used for dataset training, which excludes the corresponding sub-manifold geometry from the training data, limiting the generalisability to these domains.

A proxy timing study with PyTorch-ComIn-ICON coupling revealed a 21% reduction in advection cost, fall within the expected range, suggesting that the approach is computationally attractive, even after I/O bottleneck in host-machine learning model coupling is considered. The author therefore concludes that, a reduced order tracer embedding for ICON-ART is indeed feasible in principle, however, no model simultaneously satisfies all physical criterion set in the beginning of the assize–non-negativity, mass conservation, manifold preservation and iterative stability–at level sufficient for operational deployment. Architectural improvements, including a steepened Sigmoid gate, auxiliary mass-conservation head and GPU/native Fortran coupling are proposed as directions for future work.

The code and model are publicly available as GitHub repository SuperspeciesICON

Due to size constraint, the training data is only available in Levante.

Disclosure on the use of Artificial Intelligence

The following statement is made in accordance to the regulation regarding the use of generative Artificial Intelligence.

LLMs were used in literature review, analysis, coding, translation and sentence structure and grammar checks. Literature review involves a mix use of with LLMs and human research; mathematically proves where made with assistant with LLMs for verification of result. Codex, copilot was used to assist in code development, code review, debugging and alteration of the format of the plots to be consistent and adhere to high standard. Translation from English to German abstract was done with assistant with LLMs to ensure accurate grammar and sentence structure. LLMs were also used in review phase on this thesis, where the sentence structure and grammatical consistency were examined and corrected.

Keywords: Numerical Weather Prediction, ICON-ART, atmospheric tracer transport, reduced-order modelling, Non-negative Matrix Factorisation, Autoencoder, embedding

Zusammenfassung

Die prognostische Tracer-Advektion in der numerischen Wettervorhersage skaliert linear mit steigender Anzahl an Spezies, was einen erheblichen Engpass in Simulationen mit einer hohen Anzahl an Tracer-Spezies darstellt – etwa in der atmosphärischen Chemie, wo 10 bis 100 Tracer in jedem Advektionszeitschritt unabhängig advehiert werden müssen. Diese Arbeit untersucht, ob eine physikalisch konsistente Reduced-Order-Darstellung als rechnerisch vorteilhafter Ersatz für den Tracer-Transportoperator im ICON-ART-Modell dienen kann. Anhand eines Mineralstaubexperiments – einer Standard-Testkonfiguration in ICON-ART – als Pilotstudie werden vier Embedding-Modelle mit zunehmender Komplexität entwickelt und in einem Offline-Trainings- und Evaluierungsframework bewertet: eine Baseline-Principle Component Analysis (PCA), Non-negative Matrix Factorisation (NMF), eine gradientenbasierte PCA (einschichtiger Autoencoder ohne Aktivierungsfunktion) sowie ein neuartiger Multi-Head Autoencoder, der die Magnitudenrekonstruktion von der Zero/Non-Zero-Klassifikation durch geteilte Softplus- und Sigmoid-Decoderköpfe entkoppelt, wodurch der dynamische Bereich des Outputs erheblich erweitert und ein von Null verschiedener, wohldefinierter Gradient über das gesamte Ausgabeintervall sichergestellt wird.

PCA erzielt den niedrigsten Rekonstruktionsfehler (totaler normalisierter RMSE $\sim 1\,\%$ bei 4 latenten Dimensionen) mit exakter Idempotenz unter Iteration. Allerdings machen die Mittelwertzentrierung, die uneingeschränkte Basisvektorrotation und das Fehlen einer Non-Negativity-Bedingung PCA für den operationellen Einsatz ungeeignet; sie dient lediglich als theoretische Untergrenze der Freiheitsgrade der Mannigfaltigkeit. NMF bewahrt Non-Negativity konstruktionsbedingt, unterschätzt jedoch Konzentrationen mit hoher absoluter Magnitude und überschätzt deren Gegenstück, was eine „Drift zur Mitte“-Eigenschaft kennzeichnet. Zudem leidet NMF unter einem Zero-Variable-Problem, bei dem nahe-null- bzw. Nullkonzentrationen eine starke Varianz (Ghost Mass) für Spezies mit hoher absoluter Konzentration aufweisen, was die Notwendigkeit eines Zero/Non-Zero-Klassifikationskopfes unterstreicht. Die gradientenbasierte PCA versagt vollständig bei niedrigen absoluten Konzentrationen, wobei die Massenkonzentration von Dust A, die um $O(1\text{–}6)$ kleiner als die anderer Spezies ist, eine starke Varianz (bis zum 70-Fachen der Speziesmagnitude) aufweist, was den Bedarf an einem Klassifikationskopf weiter verstärkt. Ein konventioneller Autoencoder, der in der Ablation Study untersucht wurde, kollabiert auf ein schmales Ausgabeband. Eine eingehende Untersuchung des Gradienten zeigt, dass dieses Verhalten strukturell bedingt ist, da das Vanishing-Gradient-Problem bei $|h| \approx 8{,}67$ in float32-Präzision auftritt.

Der vorgeschlagene Multi-Head Autoencoder adressiert diese Limitierungen durch einen zusätzlichen Non-Zero-Kopf (Sigmoid, Intervall $(0, 1)$) neben dem Magnitudenkopf (Softplus,

Intervall $(0, \infty)$), deren elementweises Produkt den effektiven dynamischen Ausgabebereich erweitert. Die Gradientenanalyse zeigt, dass die drei Backpropagation-Pfade in komplementären Regimen operieren: Der BCE-Klassifikator dominiert bei Nullelementen und bewahrt die Klassifikationsgenauigkeit, während der Magnitudenkopf (MSE Loss) bei Nicht-Null-Elementen dominiert und die Konzentrationstreue sicherstellt. Das Modell erzielt eine starke Single-Pass-Rekonstruktion für Dust B–C (das Paar mit höherer absoluter Magnitude, $R^2 = 0{,}82–0{,}91$, $nRMSE \sim 10–14\,\%$) sowie eine optimale Manifold Preservation (totale $ACC = 0{,}9986$, normalisierte Wasserstein-Distanz mit $IQR = 0{,}0035$). Allerdings begrenzt die iterative Stabilität die Praxistauglichkeit im gegenwärtigen Zustand: Die Gesamtmasse degradiert nach 100 Offline-Iterationen auf 80,7 %, was möglicherweise durch das asymptotische Sigmoid-Gate ($\sigma < 1$ konstruktionsbedingt) und unbeschränktes Fehlerwachstum verursacht wird. Der Fehler im Testdatensatz des vollständigen Modells kolokalisiert unter räumlicher Analyse deutlich mit konvektiv aktiven Regionen. Dies wird auf die Trainingsstrategie zurückgeführt, bei der ein No-Source-Sink-Datensatz (konvektive und andere Quell- und Senkenterme sind deaktiviert) für das Training verwendet wurde, wodurch die entsprechende Submannigfaltigkeitsgeometrie aus den Trainingsdaten ausgeschlossen und die Generalisierbarkeit auf diese Domänen eingeschränkt wird.

Eine Proxy-Timing-Studie mit PyTorch-ComIn-ICON-Kopplung ergab eine Reduktion der Advektionskosten um 21 %, was im erwarteten Bereich liegt und darauf hindeutet, dass der Ansatz auch nach Berücksichtigung des I/O-Engpasses bei der Host-ML-Modellkopplung rechnerisch attraktiv ist. Der Autor schlussfolgert daher, dass ein Reduced-Order Tracer Embedding für ICON-ART im Prinzip machbar ist, jedoch kein Modell gleichzeitig alle zu Beginn der Arbeit definierten physikalischen Kriterien – Non-Negativity, Massenerhaltung, Manifold Preservation und iterative Stabilität – auf einem für den operationellen Einsatz ausreichenden Niveau erfüllt. Architekturelle Verbesserungen, darunter ein steileres Sigmoid-Gate, ein zusätzlicher Mass-Conservation-Kopf sowie GPU-/native Fortran-Kopplung, werden als Richtungen für zukünftige Arbeiten vorgeschlagen.

Keywords: Numerical Weather Prediction, ICON-ART, atmospheric tracer transport, reduced-order modelling, Non-negative Matrix Factorisation, Autoencoder, embedding

Contents

List of Figures

List of Tables

Acknowledgement

I would like to take this opportunity to thank all parties involved in this thesis project—the DKRZ, IMKTRO, SCC and ITI—for the resources and time that allowed the project to be completed. Dr. Hoshyaripour (IMK-TRO), Dr. Debus (SCC) and M.Sc. Kieckhefen have supported me throughout the project, and I would like to acknowledge their comments and guidance. I also wish to thank the examiners, Prof. Hoose (IMK-TRO) and TT-Prof. Nowack (ITI), for the time devoted to reviewing this thesis.

Nomenclature

CH CPU Hour

ICON Icosahedral Nonhydrostatic Weather and Climate Model

CNN Convolution Neural Network

LSTM Long-short Term Memory

ART Aerosol and Reactive Trace Gases

NWP Numerical Weather Prediction

SVD Singular Value Decomposition

NMF Non-negative Matrix Factorisation

RMSE Root Mean Square Error

NRMSE Normalised Root Mean Square Error

MAE Mean Absolute Error

NMAE Normalised Mean Absolute Error

ANN Artificial Neural Network

MLP Multilayer Perceptron

NN Neural Network

ReLU Rectified Linear Unit

MSE Mean Square Error

ADAM Adaptive Moment Estimation

ML Machine Learning

CDO Climate Data Operator

FNN Feed-forward Neural Network

CFL Courant–Friedrichs–Lewy

EOF Empirical Orthogonal Function

PCA Principle Component Analysis

1. Introduction

1.1. Motivation and scientific context

Atmospheric composition is governed by transport, mixing and chemical interaction of tracer species [1]. In NWP and Earth System modelling, prognostic tracers—aerosols, reactive trace gases, greenhouse gases and passive tracers—must be advected individually at every time step. Each tracer obeys the same flux-form continuity equation (Equations (2.1) and (2.3)), while advection tendencies are computed independently, as the flux limiter that preserves the monotonicity and non-negativity of the solution depends on the local state of the individual species. This independent treatment is necessary to preserve the physical fidelity of the solution at the expense of the computational cost of the transport module, which scales at least linearly with the number of prognostic species [2, 3].

For a small number of species, for instance 6 species, the computational cost of the total advection operator is about 4% of the total computation time (Table E.2). Modern atmospheric chemistry schemes, however, routinely simulate tens to hundreds of species. In the ICON-ART model [4, 5] used in this thesis, even a comparatively simple dust configuration (DUST_RAD) advects 3 prognostic dust species with mass and number concentrations that are treated individually. Each of these advected species must be processed by the Semi-Lagrangian advection operator and the flux limiter [2, 3, 6–8] at each integration step. The cost of tracer advection quickly becomes a heavy burden when simulating complex interactions in atmospheric aerosol chemistry and its coupling with other components of the model (e.g. aerosol-cloud interaction, radiative forcing).

This scaling problem is not merely an engineering and cost inconvenience. The fidelity of the tracer field has direct consequences for downstream physical processes. Aerosol-radiative interaction [9–11] requires accurate spatial distributions of aerosols to parametrise absorption and scattering; aerosol-cloud interaction, which depends on the local size distribution and number concentration, is a key factor for Cloud Condensation Nuclei (CCN) activation [11–14], which controls droplet activation, cloud albedo and precipitation efficiency. These feedbacks are non-linear: a slight redistribution of the aerosol loading profile can alter the radiative budget, modify boundary layer stability, and change the cloud microphysical properties over large domains. Consequently, an acceleration of the tracer transport module must preserve the spatial structure (manifold geometry), magnitude and physical fidelity (non-negativity and mass conservation) of the advected fields to an extent that the downstream interactions remain intact and physically admissible.

Sturm *et al.* [15] argues that tracer fields advected by the same model are rarely independent: tracer species with a shared emission source, size distribution or removal pathway exhibit strong correlations in spatial and temporal evolution. In the dust configuration acting as a pilot study in this thesis, the 3 size classes are linked by common emission parametrisation and deposition velocities, and the mass and number concentrations of these classes are implicitly linked by their geometric conversion, as discussed in Section 3.1.2. The correlation suggests that the tracer states may occupy a manifold of lower intrinsic dimensionality—a property that, in principle, can be exploited for a reduced-order representation for advection.

Mathematically, a reduced-order representation can be obtained by an embedding operator $\mathcal{P}$ that embeds the source manifold into a lower-dimensional latent space, and its inverse operator (reconstruction) $\mathcal{R}$ recovers the source dimension after the latent vector is transported, such that the advection operator receives only the latent representation. The computational saving is then proportional to the compression ratio minus the extra overhead introduced by the $\mathcal{P}, \mathcal{R}$ operators. It must be stressed that the embedding-transport-reconstruction cycle should not result in cascading errors that grow unboundedly or in unphysical behaviour under iterative operation.

Whether such a model can be made admissible is a non-trivial question, as the tracer manifold is heterogeneous, with properties such as high dynamic range (orders of magnitude difference), sparsity with extensive near-zero domains interspersed with localised maxima, and heavily long-tailed, skewed distributions. These properties prove to be challenging for both linear and non-linear embedding, as we shall see across the entire thesis. A linear factorisation may fail to capture the non-linear relationships between tracers [16], while having a bounded and stationary error. Neural network based solutions such as Autoencoders may produce negative concentrations [17], result in vanishing gradient problems or smoothen fine-scale structure if there are no hard constraints on these conditions. The flux limiter, as a non-linear state-dependent operator, introduces additional complexity, as the embedding must be mathematically admissible: discontinuities and sharp gradients will cause the flux limiter to fall back to low-order flux, further eroding the fine-scale structure of the tracer field.

This thesis therefore attempts to address the stated scientific gap: whether a physically consistent, reduced-order representation can be constructed for the tracer operator such that the cost of advection is meaningfully reduced whilst remaining physically admissible. The non-negativity, mass conservation and manifold geometry should be faithful to the source manifold. The investigation is carried out with the 6-tracer DUST_RAD experiment, chosen as a pilot study for future, more complex chemistry simulations.

1.2. ICON-ART and the tracer transport problem

This thesis examines possible solutions to accelerate advection in the ICON-ART model. ICON is a Numerical Weather Prediction model that solves the non-hydrostatic Navier-Stokes equations, serving as a strong host model for NWP operations, while the ART

		Time consumed (CH)	% of total time taken
51 Tracers	Total	41.06	
	Transport	2.95	7.18%
	Horizontal advection	2.34	5.69%
	Vertical Advection	0.52	1.26%
44 Tracers	Total	40.72	
	Transport	2.66	6.53%
	Horizontal advection	2.12	5.21%
	Vertical Advection	0.45	1.12%

Table 1.1.: Total time taken for each process in the Transport scheme. Two simulations were run based on the vol-aero-rad standard test suite in ART. 51 and 44 tracers were advected individually with the exact same configurations. CH denotes the Compute Hours taken by all MPI processes in total.

module extends ICON with explicit aerosol and reactive trace gas handling capabilities. In the coupled ICON-ART model, each prognostic tracer in ART must be advected individually, meaning that the computational cost of advection scales with the number of species introduced in ART, even if their tendencies or advection properties are strongly correlated.

This thesis attempts to answer the question of whether a physically consistent reduced-order embedding is possible with computational benefit. This has several constraints:

1. The error must be sufficiently small and stable over iterations
2. The species must be physically admissible; in particular, they must be non-negative and mass conserving
3. The coupled reduced-order-host model should have a sufficient computational cost reduction to justify the error introduced by the reduced-order model

Using Vol-aero-rad in the ICON standard test suite as an example (Table 1.1), the standard model advects 51 aerosol and chemical species. By turning off the transport scheme of the nmb tracer, simulation 2 only advects 44 species. Table 1.1 shows the time taken for all the processes in both experiments. Results shows that the transport module consumes 2.95 and 2.66 CH respectively. The average computational cost for each tracer in this configuration is therefore: $2.95/51 = 0.058$ or $2.66/44 = 0.060$, $\pm 3.5\%$ error. This means that the computational cost of an additional species scales linearly with advection. Theoretically, if a 50% compression ratio is achieved, the model will advect 25 species, a $0.058 \times 26 = 1.51CH$ reduction, the total computational saving is therefore 3.53% for the entire simulation.

These requirements make the problem more demanding than conventional embedding tasks. Although the model may achieve low reconstruction error and core manifold geometric preservation, it is likely that extreme values will be the dominant failure mode, for which the machine learning model performs considerably worse in boundary conditions. Furthermore, unphysical values such as negative concentrations lead to strong model drift and mass loss

during iterations; sparse manifolds and magnitude errors remain strong contributors to model error and drift, for the same reason as extreme values.

As such, this thesis assesses the model not only in error statistics, but more importantly in the preservation of the manifold's local and global structure. The author also performs model drift experiments to examine the stability of the solution and physical interpretability. However, it must be stressed that the drift experiment can only act as a reference, as in an online-coupled experiment, the stability of the solution might be improved or worsened by means of implicit damping in other components of the ICON model. This complex system behaviour cannot be assessed in a theoretical sense and must be investigated through online coupling of this module. This goes beyond the scope of this thesis, and the author will discuss the possible coupling solutions in Chapter 6.

1.3. Reduced-order models as a candidate solution

Reduced-order modelling is the natural framework for the proposed problem. The thesis first discusses the baseline Principal Component Analysis (PCA), then proceeds to Non-negative Matrix Factorisation (NMF); both are linear methods that exploit the low-dimensional structure in the tracer set, serving as interpretable baselines. The author then proposes non-linear transforms such as conventional 2-layer Autoencoders, which in principle capture more complex dependencies between tracer manifolds, offering a more flexible latent embedding. The key question is whether the non-linear transformation actually translates into better physical consistency and accuracy.

For aerosol tracers, such improvement is not immediately trivial. Tracer manifolds are highly heterogeneous, skewed and sparse. These properties make the learning difficult, particularly in boundary conditions and peripheral states of the manifold, as the author will demonstrate in Chapter 4. If a particular species is orders of magnitude smaller than other species, or lies in sparse/near-zero domains, a "small" reconstruction error in the sense of the entire model would lead to large relative error for this particular species, and quickly becomes unstable and physically implausible. In this thesis, dust species are chosen in terms of size distribution, which are connected through underlying properties of the aerosol. A reduced-order model should therefore preserve not only the statistical structures, but also the physically meaningful relationships, making the problem even more complex.

This thesis presents the problem in an offline training and evaluation fashion. The author first constructs a controlled training and testing dataset from the ICON-ART standard test suite, then trains and evaluates the models based on the constructed dataset. This allows the modelling choices to be tested systematically before considering any online coupling with the forecast model.

The author then performs model drift, sensitivity, ablation and timing experiments. Model drift will be evaluated offline with the testing dataset, then in a "proxy" online fashion, as a fully coupled model would require extensive modification of the ICON code base, and will

be discussed in Chapter 3. Timing experiments will also be performed in a proxy manner, where the author performs 3 parallel experiments:

1. Native ICON-ART experiment
2. ICON-ART coupled with ML model, where the encode-decode cycle occurs before the advection call, and the reconstructed data is sent back to the host model as a "proxy" approach.
3. ICON-ART experiment with ART module off (no aerosol tracer advected)

Simple subtraction of the timing required by ComIn and the difference in time consumed between advecting 6 dust species and no aerosol species will give an approximate estimate of the computational cost reduction.

1.4. Positioning of this thesis

The literature on reduced representations for atmospheric composition provides useful guidance, but it does not directly answer the present modelling question. Existing studies demonstrate that tracer fields can often be compressed effectively, and that latent-space methods can recover chemically or physically meaningful structure [15]. However, previous work is not directly transferable to the specific ICON-ART dust configuration studied here. In particular, the present thesis considers six tightly related dust tracers, including paired mass and number concentrations, and evaluates reduced-order approaches under criteria that are directly relevant for future model deployment: reconstruction quality, non-negativity, robustness of the manifold reconstruction, iterative drift, and computational plausibility.

The thesis therefore occupies a specific niche between atmospheric model development and machine learning. Its purpose is not merely to obtain the smallest possible reconstruction error on a static dataset. Rather, it aims to determine whether a reduced latent representation can serve as a technically credible surrogate for the transport tendencies of a physically coupled tracer system. This requires comparing simple and complex latent representations, identifying where each approach fails, and paying particular attention to the physical admissibility of the reconstructed tracer fields.

The next section places the present work in relation to similar reduced-order approaches and introduces the baseline comparison used throughout the thesis.

1.4.0.1. Training result from similar approaches

There is no existing baseline for dimensionality reduction in ICON; therefore, Sturm *et al.* [15] is used as a reference (Table 1.2). Although he did not directly provide the normalised RMSE metric in his paper, it is trivial to compute the normalised RMSE by dividing the RMSE by $\bar{y}$, consistent with our definition of normalised RMSE. It must be emphasised that the dataset in Sturm *et al.* [15] is very different from that in this thesis, as it contains

Current	Appr.1	Appr.2	Appr.3	Appr.4
aVOC	0.10232558	0.23255814	0.48837209	0.25581395
bVOC	0.09923664	0.29770992	0.69083969	0.16030534
POA	0.19534050	0.51075269	0.54838710	0.25448029
siSOA	0.32679739	0.56209150	0.61437908	0.37254902
TOA	0.04481865	0.34455959	0.26165803	1.7876e-12
TOM	0.03397516	0.14906832	0.20372671	6.2112e-13

Table 1.2.: Normalised RMSE calculated from RMSE and the mean of the distribution reported in [15]. Approach 1 denotes the NMF/Pseudoinverse; Approach 2 is the classical NMF; Approach 3 denotes an Autoencoder with ReLU activation (acting as non-zero gate); Approach 4 denotes a mass conserving NMF model.

58 species with 4 distinct classes, and separate models were trained for individual species classes. This distinction is far less clear in this thesis, as Dust A, B, and C represent dust species with different size, density and so on, where number and mass concentration (e.g. dusta & numb_dusta) are inter-convertible (Equation (3.2)).

It must be stressed that Sturm *et al.* [15] did not investigate model drift or manifold geometry; therefore, the long-term stability and manifold structure preservation remain questionable, particularly when the model predicts ghost mass as high as 200% of the source species. Geometric indicators were not mentioned in that experiment, which the author takes note of and will address carefully in this thesis.

1.5. Research question

This thesis investigates whether reduced-order representations can be used to compress tracers into their latent representation, then recover the pointwise post-advection tendencies while remaining physically admissible and computationally beneficial. More specifically, it addresses whether a reduced latent representation can reproduce tracer fields with sufficient accuracy, preserve key physical constraints such as non-negativity, and reduce computational cost relative to the native tracer advection formulation.

To answer this question, the thesis pursues four objectives. First, the author defines the relevant modelling framework by introducing the ICON host model, the ART extension, and the tracer-advection components that motivate the reduced-order approach in Chapter 1. Second, an offline experimental setup is constructed that produces suitable datasets for training, validation, and testing under controlled conditions in Chapter 3. Third, reduced-order models are developed and compared, including baseline linear methods and neural-network-based approaches, with particular attention to physical admissibility (Chapters 3 and 4). Fourth, these models are evaluated with respect to reconstruction accuracy, magnitude fidelity, non-negativity preservation, drift under iterative application, generalisability across

simulation conditions, and expected computational benefit (Chapters 4 and 5). Conclusions and outlook for future research are presented in Chapter 6.

2. Theoretical Background

Section 2.1 introduces advection in Numerical Weather Prediction and the ICON-ART modelling framework, followed by a brief discussion of each module involved in this experiment. The author then proceeds to discuss the various latent space embedding architectures, including linear (Sections 2.3 and 2.3.1) and non-linear (Sections 2.3.2 and 2.3.3) methods, including an Autoencoder with multi-head decoder (Section 2.3.3), which is the key novel contribution of this experiment.

2.1. Numerical Simulation

2.1.1. Tracer transport (Advection)

Advection redistributes mass across the atmosphere, altering the atmospheric composition, resulting in changes in chemical properties, alteration of cloud and rain patterns, and changes to the radiative properties of the atmosphere. Each prognostic tracer must be advected independently; the computational burden scales at least linearly, which quickly results in a strong bottleneck in larger simulations. These independent prognostic tracers are a natural target for embedding models. In a comprehensive troposphere-stratosphere chemistry interaction simulation, 1–200 gas-phase tracers and a number of aerosol tracers are simulated and advected. As each species must be advected independently, the transport module scales linearly and quickly becomes an expensive process in the simulation. Sturm *et al.* [15] demonstrated in his experiment that, for organic aerosol simulation in the LOTOS-EUROS model, advection is the single most expensive process, consuming up to 7–80% of the total CPU clock time in chemistry-transport configurations.

The tracer transport module in ICON [19] uses the flux-form Semi-Lagrangian transport scheme [2]. In this framework, the tracer tendency is formulated in conservative form such that both transport accuracy and mass consistency are maintained as far as possible within the numerical discretisation. Unlike pure Eulerian finite-volume schemes, where fluxes are computed at cell interfaces and are hence limited by the Courant–Friedrichs–Lewy (CFL) constraint, the semi-Lagrangian scheme tracks upstream departures such that the flow can span multiple time steps without concern for the stability of the solution [2, 7]. This enables much larger advection time steps without losing numerical stability. This is particularly relevant at high horizontal resolutions where the wind speed approaches the limit of

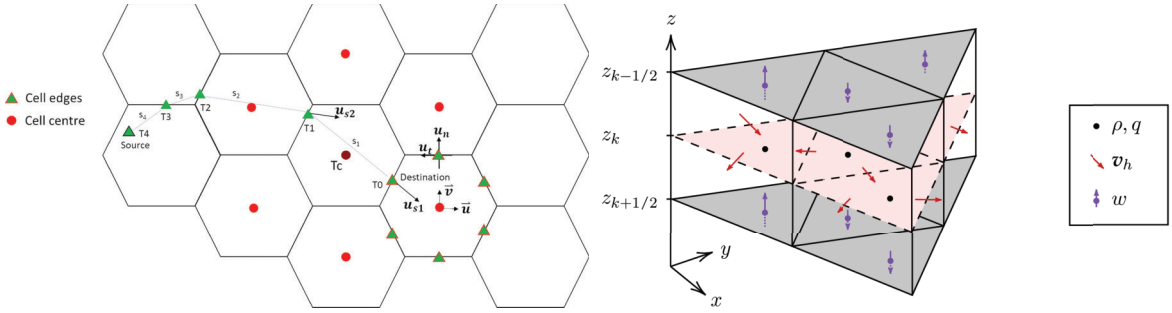

Figure 2.1.: Schematic diagram of the transport scheme on an irregular grid. (left) An air parcel moving through an arbitrary 2D space discretised by a pentagonal grid. The air parcel begins its motion at T4, reaching the final cell through cell wall T0. Adapted from Févotte & Idier [18]. (right) Semi-Lagrangian advection scheme for ICON on a triangular grid. Adapted from Strassmann [19]. Note that the "cell centre p, q" depicted is an inaccurate representation of the ICON grid structure, as values are stored at the circumcentre as cell-averaged quantities rather than point values.

Equation 2.4. The continuity equation for tracer mixing ratio q_k is written schematically as

$$\frac{\partial(\rho q_k)}{\partial t} + \nabla \cdot (\rho q_k \mathbf{v}) = S_k, \tag{2.1}$$

where ρ denotes air density, v the horizontal wind field, and S_k collects all source and sink processes such as emission, deposition or sub-grid-scale parametrisation, as denoted on the R.H.S. of Equation (2.1). The source and sink processes can be expanded as

$$S_k = \left.\frac{\partial \rho_i}{\partial t}\right|_{mix.} + \left.\frac{\partial \rho_i}{\partial t}\right|_{conv.} + \left.\frac{\partial \rho_i}{\partial t}\right|_{scav.} + \left.\frac{\partial \rho_i}{\partial t}\right|_{chem.} + \left.\frac{\partial \rho_i}{\partial t}\right|_{emiss.} + \left.\frac{\partial \rho_i}{\partial t}\right|_{dep.} \tag{2.2}$$

and are handled in the model physics (hydrometeors and other host model tracers) and the ART extension (chemical, dust and aerosol species).

In ICON [4], the equation under triangular transform becomes

$$\frac{\partial \overline{\rho}\hat{q}_k}{\partial t} + \nabla \cdot (\overline{\rho}\hat{q}_k\hat{v}) = -\nabla \cdot (\overline{J}^z_k k + \overline{\rho q"_k v"}) + \overline{\sigma}_k \tag{2.3}$$

In the context of this thesis, the superspecies approach modifies the tracer state before and after advection, while the dynamical core and the model physics remain unaltered.

Flux limiter

The flux limiter is introduced in the flux-form Semi-Lagrangian scheme to preserve monotonicity and suppress spurious oscillations [6, 20, 21]. While high-order transport schemes are substantially more accurate in smooth gradients and desirable for fine-scale structure reconstruction with minimal diffusion, Godunov's theorem demonstrates that any linear, monotonicity-preserving advection scheme is at most first-order accurate [22]. High-order schemes therefore inevitably produce unphysical oscillations (over- and undershoots) near

sharp gradients unless supplemented by a non-linear limiting procedure. Such oscillations are particularly problematic, as negative tracers and "ghost" masses lead to unphysical behaviour, especially under long simulations. The flux limiter ensures the transported tracer field remains a bounded function while retaining the accuracy gains from the high-order transport scheme. The flux limiter is expressed schematically as [21]:

$$F^{\text{lim}} = F^{\text{low}} + \phi \left(F^{\text{high}} - F^{\text{low}} \right), \tag{2.4}$$

Where $\phi \in [0, 1]$ is a coefficient defined from local solution bounds.

The transport update is then a linear combination of a low-order, monotonic flux and a high-order correction flux. The low-order flux term results in a smoother, diffusive gradient, while the high-order term, although returning sharp gradients and reducing numerical smearing, will encounter over- and undershoots in strong gradient regions. The flux limiter is introduced to correct the behaviour of both systems by adjusting the proportion of the high-order flux term according to the local smoothness and boundedness of the tracer field.

The correction is strongest when the solution is locally smooth: the high-order flux can be introduced without the risk of introducing new extrema. In a discontinuous or strong gradient structure, such as initialising a tracer with a point source, ϕ should be kept at a minimum to preserve non-negativity and avoid the Gibbs phenomenon [23].

This is particularly relevant, as ICON computes the mass transfer over the finite departure and arrival region, as opposed to cell interfaces only in Eulerian schemes [3, 24]. The deformed upstream control volume can yield high-accuracy transport, while the risk of unphysical behaviour is also increased if no limiter is applied. The flux limiter prevents unphysical runoff from occurring under iterative update steps, particularly for manifolds with steep gradients such as aerosol and tracer fields.

This however, comes in a high cost.

The flux limiter introduces additional non-linearity to the transport scheme. The ϕ for each grid cell must be calculated at every update step and for every constituent, depending on the neighbouring tracer states [5]. This means that a single, linear mapping cannot be achieved for the entire tracer field, as the computational cost scales with the reconstruction, limiter and conservative remapping operations [25]. As such, advection can create a major bottleneck in Numerical Weather Prediction, particularly when a high number of tracers are advected.

From Figure 2.2, the CPU hours taken for advection in the ART-off configuration are approximately 52% of the reference run. This means that, if the number of species is reduced by half, one could theoretically expect a 25% reduction in total CPU time required for advection, and 2% in total CPU hours. This indicates that advection can become a major contributor to compute power consumption, particularly when many tracers are involved (e.g. in atmospheric chemistry simulations).

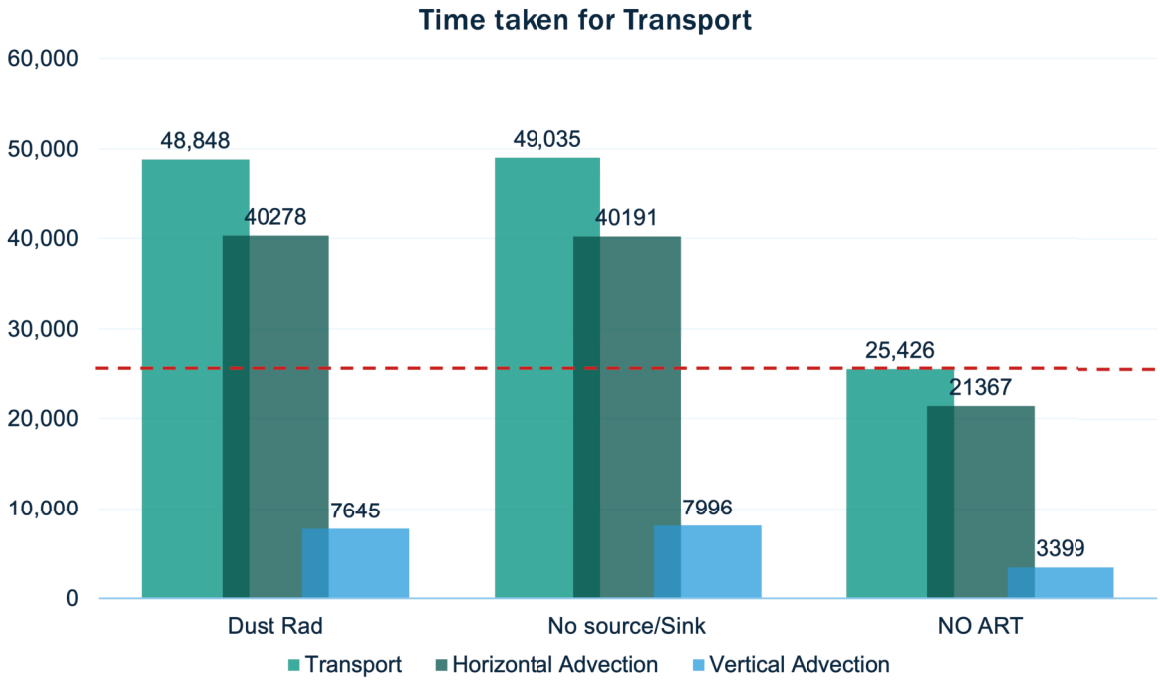

Figure 2.2.: Total transport time (in seconds) for a reference ICON-ART run using the exp.dust_rad configuration from the ICON-NWP test suite. Three configurations are shown: the full model with dust–radiation coupling (Dust Rad), the full model with source/sink terms disabled in the ART source code (No source/Sink), and a baseline run with `lart = .false.` in `run_nml` (NO ART). Each group is decomposed into total transport, horizontal advection, and vertical advection. The red dashed line marks the NO ART transport baseline, illustrating the overhead attributable to tracer advection. Note that the No source/Sink configuration exhibits marginally higher timings than the full model, as disabling sources and sinks is implemented by multiplying concentrations by zero, introducing a small additional computational cost.

In conclusion, tracer advection is a computationally relevant component due to the repeated, independent computation necessary for each prognostic tracer. The Semi-Lagrangian operator is numerically complex compared to a pure Eulerian approach, as trajectory and mass flux are both computed, with the flux limiter introducing an extra level of complexity. For this reason, the embedding of tracers before advection is scientifically relevant for lowering computational cost, particularly in large atmospheric chemistry simulations, provided that the tracer fields remain physically consistent.

2.1.2. The ICON Model

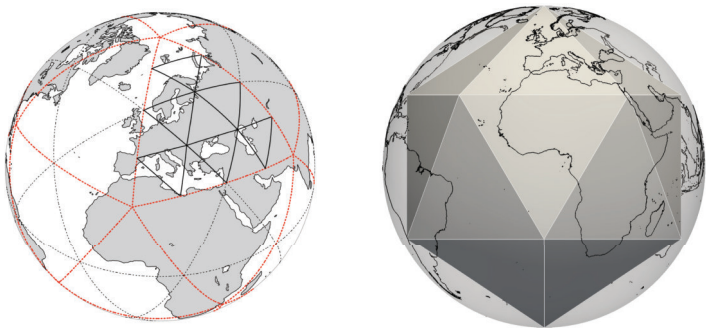

Figure 2.3.: Icosahedron grid and the bisection method adopted in ICON. (left) Illustration of the grid construction procedure. The original spherical icosahedron is shown in red, denoted as R1B00 following the nomenclature described in the text. In this example, the initial division (n=2; black dotted), followed by one edge bisection (k=1) yields an R2B01 grid (solid lines). (right) Graphical representation of the 12 pentagon points. Here, the base icosahedron is rotated exactly like the DWD operational grids. (Adapted from Prill *et al.* [4])

The numerical experiments in this thesis are performed with ICON, a Numerical Weather Prediction (NWP) model developed by the Max Planck Institute for Meteorology and the German Weather Service (DWD) [4]. ICON solves the fully compressible, non-hydrostatic Navier-Stokes equations on unstructured triangular grids, particularly suitable for resolutions ranging from global NWP to large-eddy simulations. The removal of the hydrostatic approximation is essential in km-scale simulation, as the hydrostatic approximation assumes vertical accelerations are negligible compared to gravitational and pressure gradient forcing. These assumptions break down at these scales, as convection can no longer be ignored and must be resolved explicitly [26]. ICON has been increasingly adopted in global, regional and large-eddy simulations, acting as a unified framework for future atmospheric science research [4]. In contrast to conventional Gaussian meshes in classical NWP models such as the ECMWF-IFS [27], ICON employs an unstructured icosahedral (20-faced polyhedron) grid, permitting improved homogeneity and relaxed vertices across its horizontal grid. Conventional NWP models use a variant of the latitude-longitude grid, where singularities at the poles of the sphere deform the rectangular grid cells near the pole, resulting in a uniform grid towards the equator and more trapezium-like grid cells towards the poles. The ratio of the smallest to the largest cell area remains below $1 : 1.2$ across the sphere, compared to $1 : 10^3$ in regular lat-lon grids [26]. This relaxes the CFL condition by shrinking the area ratio of the smallest to the largest grid cell, as the CFL condition is defined as [28]

$$C = \frac{u\Delta t}{\Delta x} \leq 1 \rightarrow u\Delta t \leq \Delta x \tag{2.5}$$

such that the grid size cannot exceed the displacement of the air parcel at a discretised time, and the velocity u is usually defined as the local speed of sound. This means that a finer grid size must be accompanied by a much smaller time step, as shown in Figure 2.4. Consequently, a larger time step can be used for ICON at the same resolution than for a lat-lon grid model.

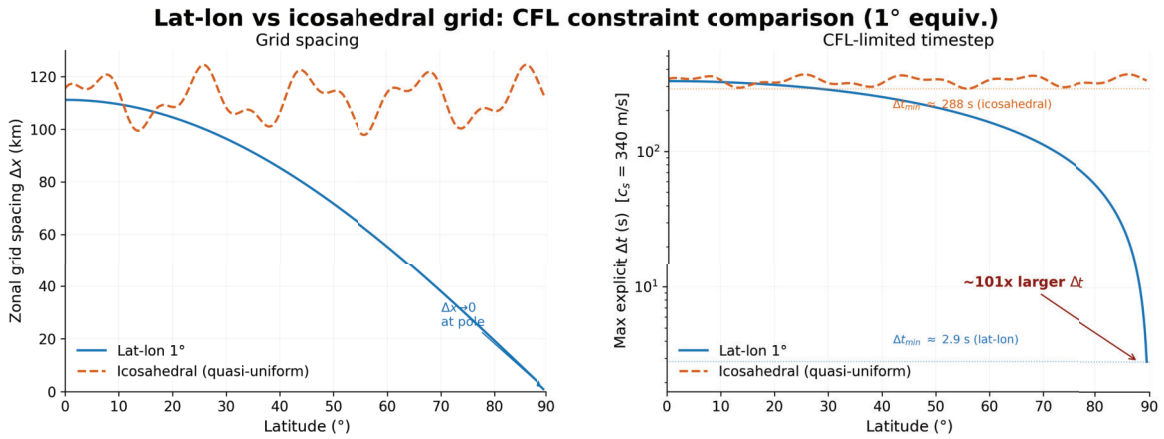

Figure 2.4.: Courant-Friedrichs-Lewy constraint at 1^o equivalent mesh. (left) The grid spacing at 1^o Latitude equivalent resolution. The circumference of circle shrinks as latitude increases, resulting in a decrease in grid spacing. In Icosahedron grid, the average grid spacing remains relatively similar, oscillating as the grid approach and depart the vertex of the base triangle of the Icosahedron. (right) The corresponding time discretisation needed to maintain the same grid resolution. Equation 2.5 gives the relation of Δt & Δx at u=speed of sound.

This has a significant impact on the representativeness of processes at the equator, as the resolution of resolved processes at the equator will be lower compared to the same process at the pole.

In ICON, the horizontal resolution is commonly described by the pair (R, B or n, k), where n denotes the initial refinement of the base triangle (dividing the triangle into n sections, hence n^2) and k the number of subsequent edge bisections (Figure 2.3). The grid cells are then projected onto a sphere, and the area of each cell changes with height to match the spherical geometry of the Earth. The prognostic variables are stored at the circumcentres and edge midpoints under C-type staggering, which has been shown to have favourable dispersion properties for internal-gravity waves in ICON [26]. The total number of horizontal cells is given by Prill *et al.* [4]

$$n_{cells} = 20n^2 4^k \tag{2.6}$$

and the corresponding effective mesh size are approximated by

$$\overline{\Delta x} \approx \frac{5050}{n2^k} \text{ km} \tag{2.7}$$

2.1.3. Aerosols and Reactive Trace Gases module

The ART module is an extension to the ICON model, in which emissions, transport, gas-phase chemistry, and aerosol dynamics are coupled to the host model, such that a better representation of aerosol-cloud and aerosol-radiation interactions can be properly simulated [29]. This coupling between aerosol radiative effects and meteorology is crucial for the accurate representation of semi-direct aerosol effects on boundary layer stability. ART

allows an arbitrary number of prognostic tracers, tracer properties, emission source control, chemical reaction pathways and deposition processes via XML [30]. Aerosol dynamics are represented as a superposition of log-normal modes: Aitken, accumulation and coarse mode, each advecting mass and number concentration as independent prognostic variables. The interaction between these modes and the host model radiation scheme (e.g. ECRAD, Hogan & Bozzo [31]) enables the representation of direct aerosol-radiative forcing. In addition, coupled land use and surface wind are introduced to parametrise mineral dust, sea salt and anthropogenic aerosol emissions [5, 29]. As the atmosphere is composed of numerous number of species, a realistic NWP model can only select a finite number of species that has strong impact to the downstream forecasting skill, therefore, an increasing number of species could represent increasing details of the true atmospheric composition, hence imrpoving the forecasting skill. The high number of tracers carries high computational cost, which this thesis attempts to address by producing a latent representation of tracers introduced in the ART submodule, such that the number of tracer fields being advected is reduced.

2.2. Application of Machine Learning in Atmospheric Sciences

Machine learning has a long history in atmospheric sciences and numerical modelling. Linear methods, such as linear regression for data analysis, Principal Component Analysis and Singular Value Decomposition for spectral and wave analysis (e.g. El Niño-Southern Oscillation [32], North Atlantic Oscillation [33] and Arctic Oscillation [34]), have been widely used alongside remote sensing inversion.

The advancement of GPU technologies, in particular FP32 performance, which is relevant for machine learning model training, has grown exponentially in the past 20 years [35], allowing non-linear models such as Convolution Neural Network (CNN) with increasing depth to be implemented. Deep Neural Networks [36], Long-short Term Memory (LSTM) [37], CNN + LSTM [38], ConvLSTM [39], U-Net [40], Graph Neural Networks [41], and most recently attention-based mechanisms, provide high-fidelity weather prediction without the need for large computer clusters in traditional Numerical Weather Prediction.

Pangu [42], an Earth-specific Transformer architecture, surpasses the operational IFS forecast in numerous variables, such as 2m temperature, 10m wind, and 500 hPa specific humidity. Most excitingly, tropical cyclone trajectories, a long-standing problem for operational forecasting, are accurately predicted by Pangu, even for cases where the operational IFS failed completely. This has substantially revolutionised model development, with deep learning methods increasingly incorporated into operational forecasts. ECMWF-AIFS [43] and AICON [44] were introduced recently, featuring shifted window attention with latitude-longitude and unstructured grids respectively.

Hybrid numerical-deep learning models are also being introduced, as a balance between physical admissibility and the performance improvement brought by deep learning methods. Grundner *et al.* [45] demonstrated that deep learning based cloud cover parametrisation can

be trained on coarse-grained storm-resolving ICON simulation data, producing accurate sub-grid scale estimates while remaining interpretable through SHAP-based feature attribution. Parametrisation schemes, which estimate sub-grid processes that are not explicitly resolved, are the natural target for such emulation. By performing Large Eddy Simulation, a reduced representation (parametrisation) can be trained for coarser resolution simulations, resulting in vast computational advantage over conventional parametrisation schemes.

The increasing adoption of machine learning in atmospheric modelling reflects a broader paradigm shift, where data driven methods complement physics-based simulations to achieve improved accuracy or computational efficiency. However, while most research has focused on either full model replacement through foundation models or parametrisation emulation through neural network surrogates, relatively few studies have investigated the use of machine learning for accelerating individual numerical operators within the host model. Tracer transport, which must independently advect each prognostic species, represents a computationally expensive module that scales linearly with the number of tracers and is therefore particularly amenable to dimensionality reduction. This thesis explores both linear and non-linear embedding methods to compress the tracer state before advection and reconstruct it afterwards, a strategy that retains the physical dynamical core while reducing the computational burden of the transport operator.

2.3. Latent Space Transformation

This section discusses the fundamentals of a reduced order representation of the tracer state. The author also discusses the various methods to be pursued in the later sections.

Dimensionality reduction has a long history in atmospheric sciences: Empirical Orthogonal Function (EOF) [46], or Principle Component Analysis (PCA), decompose a spatial-temporal dataset into spatial (Empirical Orthogonal Function (EOF)) and temporal (Principal Components) modes, or mathematically, the rotation to the direction of maximum variance. These modes are a pinnacle of modern weather research, revealing essential spatial patterns that prove impactful to socio-economic development. Recently, Sturm *et al.* [15] employed a "superspecies" approach to the advection operator, in which chemically related organic aerosol species are linearly embedded into a set of latent "tracers" for advection, then reconstructed after the tracers are advected. Their work on the LOTOS-EUROS model demonstrated that Non-negative Matrix Factorisation (NMF) can achieve up to 75% compression rate with reconstruction error below 5%, achieving a 30–50% speed-up in the advection operator. This thesis attempts to extend the approach to ICON-ART and, in addition to the linear approach, introduces methods to exploit non-linear embedding approaches, proposing a novel multi-headed Autoencoder that aims to capture non-linear relationships between the tracer species.

The objective of the dimensionality reduction approach is to embed the higher-dimensional tracer state $\mathbf{V}$ into its lower-dimensional latent representation

$$V \in \mathbb{R}_{+}^{m \times n} \mapsto H \in \mathbb{R}_{+}^{r \times n}$$

where $r < m$ and r denotes the latent dimension. The latent representation is obtained through a projection (embedding) denoted by $\mathcal{P}$,

$$\mathbf{H} = \mathcal{P}(\mathbf{V}),$$

and the original tracer vector is approximated by reconstruction (embedding inversion) denoted by $\mathcal{R}$,

$$\widehat{\mathbf{V}} = \mathcal{R}(\mathbf{H}).$$

Ideally, the transform $(\mathcal{P}, \mathcal{R})$ satisfies

$$\widehat{\mathbf{V}} \approx \mathbf{V}$$

with sufficiently small reconstruction error.

The 2 classes of transformation proposed in this thesis are:

1. **Linear methods**, represented by non-negative matrix factorisation (NMF);
2. **Non-linear methods**, represented by a feed-forward Autoencoder.

The key constraint here is that the reconstructed matrix should remain numerically stable under iteration and should not introduce unphysical negative tracer values in the reconstructed state.

Linear Inversion

For a linear map $\mathbf{A}$, a formal inverse is not generally available when the system is ill-conditioned (rectangular, rank deficient). In such cases, the inverse of the matrix can be obtained via the Moore-Penrose pseudoinverse [47], which provides a least-squares inverse in the sense of minimum-norm reconstruction. Although the pseudoinverse exists for any matrix, the numerical sensitivity of ill-conditioned matrices, where singular values are small, is severe. This means that under repeated reconstruction using SVD, an ill-conditioned inverse amplifies errors and degrades numerical stability rapidly, which is not ideal for our application.

If the Singular Value decomposition (SVD) of $\mathbf{A}$ is

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}}$$

then the pseudoinverse is given by

$$\mathbf{A}^{+} = \mathbf{V}\Sigma^{+}\mathbf{U}^{\mathrm{T}} \mid \mathbf{U} \in \mathbb{R}^{m\times m}, \Sigma^{+} \in \mathbb{R}^{m\times n}, \mathbf{V} \in \mathbb{R}^{n\times n} \tag{2.8}$$

where Σ^{+} is obtained by taking the reciprocal of all non-zero singular values and transposing the diagonal structure.

$$\Sigma_{ij}^{-1} = \begin{cases} 1/\sigma_i & \text{if } \sigma_i \neq 0 \\ 0 & \text{if } \sigma_i = 0 \end{cases} \tag{2.9}$$

This is the mathematical foundation of PCA in an ill-conditioned system. In PCA, the orthogonal basis vectors are first rotated to the direction of maximum variance, then to the direction of maximum remaining variance orthogonal to the principal component. In this case, maximum variance can be extracted along each successive direction.

However, PCA is mean-centred, and will drift significantly as the manifold shifts in the manifold space. Furthermore, linear PCA would be computationally intensive due to the multiple linear operators required, particularly for a large dataset such as NWP simulations.

This motivates the use of a regularised, constrained form of factorisation over a purely unconstrained inverse. In particular, for tracer species that are explicitly non-negative and often sparse, an additive, parts-based decomposition is more natural than PCA's subtractive decomposition [48]. The natural candidate for such a method is Non-negative Matrix Factorisation (NMF).

Gradient based model optimises weights based on an objective function, often time with AdamW [49] or AdaGrad [50]. The weights are initialised randomly, then the training data is piped through the initialised weights of the network, compared to the ground truth by a loss function, then the back propagation is computed by evaluating the gradient of the networks, updating the weights and biases. Each iteration of all training data are termed "epoch", with monotonically decreasing prediction-ground truth error, until reaching the optima ("convergence").

2.3.1. Non-Negative Matrix Factorisation

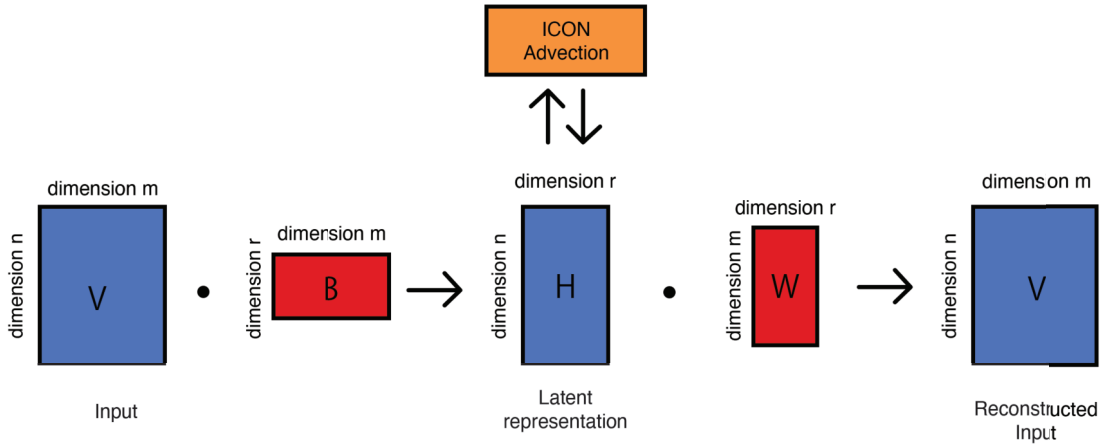

Figure 2.5.: Schematic of Non-negative Matrix Factorisation. The source matrix **V** is projected onto its latent subspace **H** ("superspecies") using the projection matrix **B**. The compressed superspecies are passed to the ICON advection operator. After advection, the advected superspecies are reconstructed to the original space $\mathbf{V}_{\text{reconstructed}}$ using the reconstruction matrix **W**.

Non-negative Matrix Factorisation (NMF) was introduced by Lee & Seung [48] as a parts-based representation learning method, originally applied to facial image decomposition,

where the non-negativity constraint produces components that correspond to recognisable parts rather than the signal-mixed eigenvectors of Principle Component Analysis (PCA). The method was then formalised with multiplicative updates (Equation (2.15), Seung, Lee, *et al.* [51]), showing that the update results in monotonically decreasing Euclidean distance and generalised Kullback-Leibler divergence (Equation 2.14) between the input and target. This is relevant for atmospheric sciences, as many parameters are physically constrained to be non-negative, and negative values denote unphysical behaviour.

Non-negative Matrix Factorisation (NMF) approximates the non-negative tracer vector $\mathbf{V}$ by the product of its non-negative latent representation $\mathbf{H}$ and its inverse embedding vector $\mathbf{W}$,

$$\mathbf{V} \approx \mathbf{WH}, \tag{2.10}$$

with

$$\mathbf{W} \in \mathbb{R}_+^{m \times r}, \qquad \mathbf{H} \in \mathbb{R}_+^{r \times n}. \tag{2.11}$$

The explicit non-negativity in NMF makes it particularly relevant for physical systems, as any reconstruction remains non-negative, which is a physical constraint for many systems. Unlike PCA, NMF yields additive combinations of the basis vectors, where each row of the basis matrix $\mathbf{W}$ denotes the linear relationship of each input channel (in this case tracers) with the particular latent channel ("superspecies"). A key trade-off, however, is that NMF does not yield a unique solution: different initialisations can converge to different local minima of the objective function, and the non-convexity of the joint optimisation over $(\mathbf{W}, \mathbf{H})$ means that only monotonically decreasing optima, rather than global optimality, are guaranteed by the multiplicative update rules [18, 51]. This is similar to the behaviour of Neural Networks' non-convex loss surfaces and initialisation-sensitive properties [52], which will be discussed in later sections.

	Dimension 1	Dimension 2	Matrix
V	m	n	Input
H	r	n	Compressed "Superspecies"
B	r	m	Projection from input to superspecies
W	m	r	Projection (reconstruction) from superspecies to input space

Table 2.1.: Dimensions and roles of the matrices in the NMF decomposition $V \approx WH$**.** Here m denotes the number of input species, n the number of samples, and r the number of superspecies (latent components).

The factorisation is obtained by solving the β-divergence equation

$$\min(d_\beta(\mathbf{x} \,\|\, \mathbf{y})), \tag{2.12}$$

where d_β denotes the β-divergence between the data and its reconstruction [53]. β-divergence is a distance-like function that measures the dissimilarity between each element of the two manifolds (x, y). In the context of NMF, it measures the dissimilarity between the input

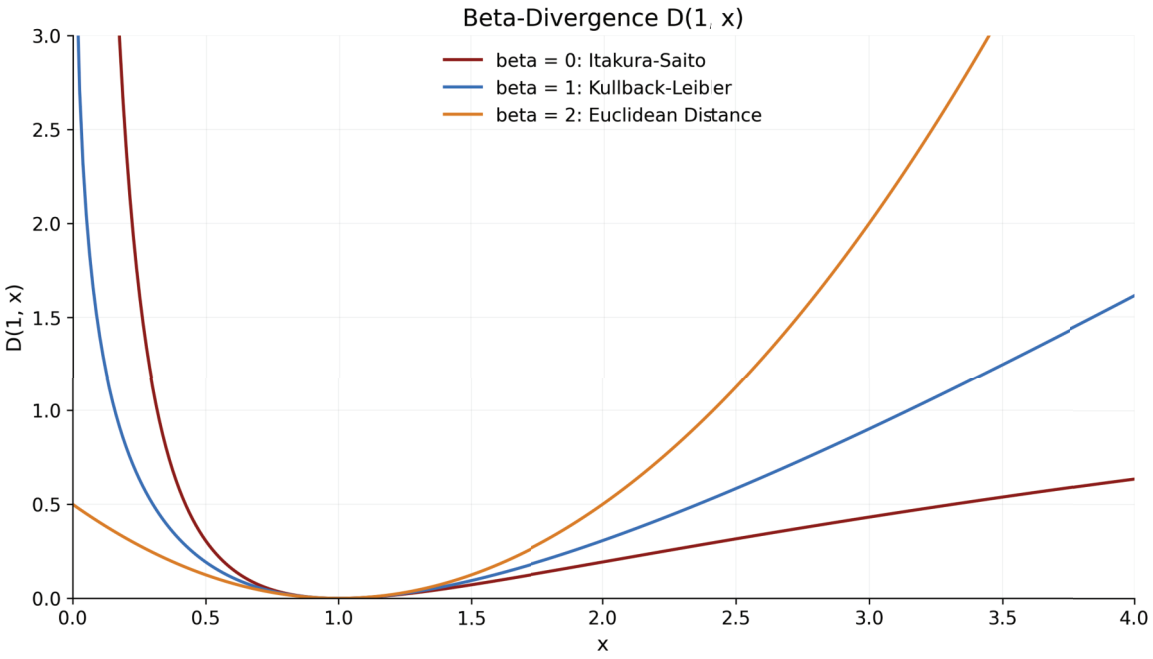

Figure 2.6.: Beta-divergence $D_\beta(1, x)$ as a function of x for three special cases: Itakura–Saito ($\beta = 0$), Kullback–Leibler ($\beta = 1$), and Euclidean distance ($\beta = 2$). All three divergences share a global minimum of zero at $x = 1$. The Itakura–Saito divergence penalises underestimation most heavily near zero, while the Euclidean distance grows quadratically for large x, making the choice of β critical when the input data spans several orders of magnitude, as is the case for aerosol concentrations.

V and the product of its latent representation and the reconstruction matrix **WH**. The function

$$d_\beta(V, WH) = \sum_{ij} D_\beta(V_{ij} || (WH)_{ij}) \tag{2.13}$$

is written formally as

$$d_\beta(x \mid y) \overset{\text{def}}{=} \begin{cases} \dfrac{1}{\beta(\beta-1)}\left(x^\beta + (\beta-1)y^\beta - \beta x\, y^{\beta-1}\right), & \beta \in \mathbb{R} \setminus \{0, 1\}, \\ \frac{1}{2}(x-y)^2 & \text{for } \beta = 2 \text{ (Euclidean distance)}, \\ x \log\left(\dfrac{x}{y}\right) - x + y, & \text{for } \beta = 1 \text{ (KL-Divergence)}, \\ \dfrac{x}{y} - \log\left(\dfrac{x}{y}\right) - 1, & \text{for } \beta = 0 \text{ (IS-Divergence)}. \end{cases} \tag{2.14}$$

where the function can be generalised as common loss functions at particular values of β, including the squared Euclidean distance ($\beta = 2$) and the Kullback–Leibler divergence ($\beta = 1$) [53].

The obtained y value in Equation (2.14) is further factorised into $\mathbf{W}, \mathbf{H}$ via the classical Heuristic update

$$\begin{cases} H \leftarrow H \odot \frac{(W^T V) \odot (W^T WH)^{\beta-2}}{(W^T WH)^{\beta-1}}, \\ W \leftarrow W \odot \frac{(VH^T) \odot (WWH)^{\beta-2}}{(WWH)^{\beta-1}}. \end{cases} \tag{2.15}$$

Under this strategy, $\mathbf{H}$ is first held fixed, and the beta divergence optimises the $\mathbf{W}$ matrix with respect to $\mathbf{V}$, and vice versa.

Training Strategy

This thesis adopts a two-step approach for the NMF algorithm. The tracer vector $\mathbf{V}$ is first factorised into $\mathbf{W}, \mathbf{H}$ under Equation (2.10). The reconstruction matrix $\mathbf{W}$ and latent representation $\mathbf{H}$ are updated for the current epoch,

$$\underset{\mathbf{W},\mathbf{H}}{\operatorname{argmin}} ||\mathbf{V} - \mathbf{WH}|| \quad | \quad \mathbf{V} \in \mathbb{R}_{+}^{\mathbf{m}\times\mathbf{n}} \tag{2.16}$$

then the embedding

$$\mathbf{B} \in \mathbb{R}_{+}^{\mathbf{r}\times\mathbf{m}}$$

is computed by setting $\mathbf{H}, \mathbf{V}$ constant for the current epoch,

$$\underset{\mathbf{B}}{\operatorname{argmin}} ||\mathbf{H} - \mathbf{BV}|| \quad s.t. \quad \mathbf{B} \in \mathbb{R}_{++}^{\mathbf{r}\times\mathbf{m}} \tag{2.17}$$

This strategy is analogous to an Autoencoder, such that the encoder/embedding would be computed before the advection operator, and the inverse embedding/decoder reconstructs the state space after the latent vector is processed by the advection operator.

2.3.2. Non-linear Transformation

The following model extends beyond the linear latent space projection by introducing non-linear embedding and reconstruction. Non-linearity is prominent in atmospheric tracers. For instance, the concentration of secondary gas-phase aerosol depends non-linearly on the precursor gas and photochemical conditions [1]. Linear embedding, by construction, cannot capture such dependencies and therefore requires a higher number of degrees of freedom in the latent representation, while non-linear, Neural Network based Autoencoders [16, 54] provide a flexible backbone for learning non-linear manifolds from data.

A simple, conventional toy model is trained to assess the limitations and generalisability of non-linear methods. The author subsequently introduces a custom, multi-head decoder that addresses the issues identified in linear embeddings and the non-linear toy models.

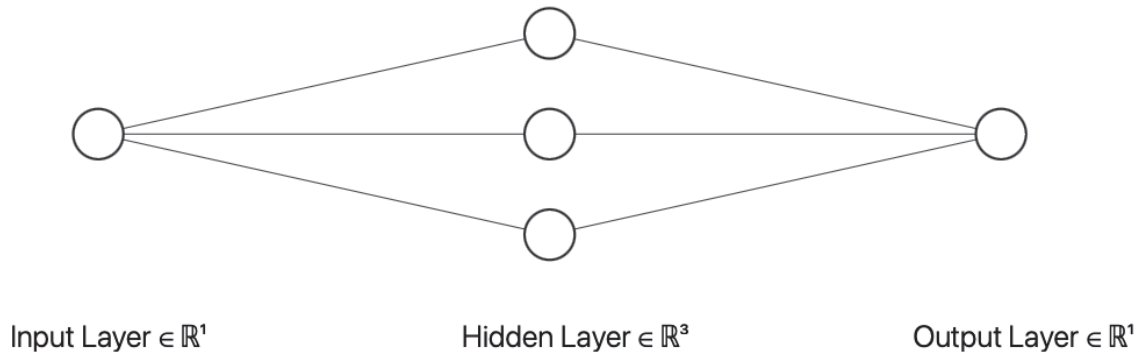

Figure 2.7.: Structure of an idealised single hidden layer feed-forward network with 3 neurons. The input layer receives the feature vector $\mathbf{x}$, which is mapped through trainable weights θ_{ij} and biases ϕ_j to the hidden layer. An activation function $a[\cdot]$ is applied element-wise at the hidden layer before the output is computed via a second affine transformation (Equation 2.19).

Feed-forward Neural Network

A Feed-forward Neural Network (FNN) can be seen as a parametric map

$$\mathbf{y} = f[\mathbf{x}, \theta_{ij}, \phi_j] \tag{2.18}$$

where $\mathbf{x}$ denotes the input, and θ, ϕ denote the trainable weights and biases respectively. Subscripts i, j denote the n^{th} node in the current layer and the m^{th} layer in a FNN respectively [55]. Each layer applies an affine transform followed by a non-linear activation function (e.g. ReLU, Sigmoid). Formally, a single hidden layer FNN can be written as

$$\mathbf{y} = \phi_2 + \theta_{i2}\, a[\theta_{i1}\mathbf{x} + \phi_1] \tag{2.19}$$

where $a[\bullet]$ denotes the activation function. The non-linearity introduced by the activation function enables the network to represent non-linear maps that cannot be expressed by linear projection. A Feed-forward Neural Network (FNN) with no activation is equivalent to linear regression, as the composition of affine layers collapses into a single affine transformation regardless of depth, as demonstrated in Goodfellow [56]. The choice of activation function therefore has profound importance for training dynamics and the mathematical behaviour of the Neural Network, as we shall see in Chapters 4 and 5. It has been mathematically proven by Cybenko [57] and Hornik *et al.* [58] that a multi-layered FNN with Sigmoid activation function in a univariate mapping can be viewed as a universal function approximator. More precisely, the Universal Approximation Theorem [57, 58] states that a single-layer FNN with a sufficient number of neurons and a non-polynomial activation function can approximate any continuous function on a compact subset of $\mathbb{R}^n$ to arbitrary precision.

Autoencoder architecture

The Autoencoder architecture was introduced by Le Cun & Fogelman-Soulié [59] as a compressed internal representation through back-propagation, and deepened by Bourlard & Kamp [60], establishing the formal equivalence between linear Autoencoders and PCA, while Hinton & Zemel [61] subsequently formalised the Autoencoder framework. A single-layer Autoencoder with no activation functions is equivalent to PCA, as it recovers the PCA subspace (up to rotation), and the introduction of activation functions enables the model to learn curved manifolds in the source space. Kramer [16] explicitly extended the Autoencoder as a generalised, non-linear extension of PCA, for which Hinton & Salakhutdinov [54] later proved that deep Autoencoders with multiple hidden layers can produce substantially lower reconstruction error than PCA for complex high-dimensional datasets.

The Autoencoder is employed in this thesis as an alternative to NMF. An autoencoder consists of an encoder and decoder [16, 17], where the encoder

$$\mathbf{h} = f_{enc}(\mathbf{v}) \tag{2.20}$$

embeds the input tensor $\mathbf{x} \in \mathbb{R}^m_{\geq 0}$ into its lower-dimensional representation $\mathbf{h} \in \mathbb{R}^r$, and the decoder

$$\widehat{\mathbf{v}} = f_{dec}(\mathbf{h}) \tag{2.21}$$

reconstructs the state space from its latent representation. It is a class of self-supervised FNN that is optimised by minimising the distance between the input and reconstructed state space, often using the β-divergence algorithm [18].

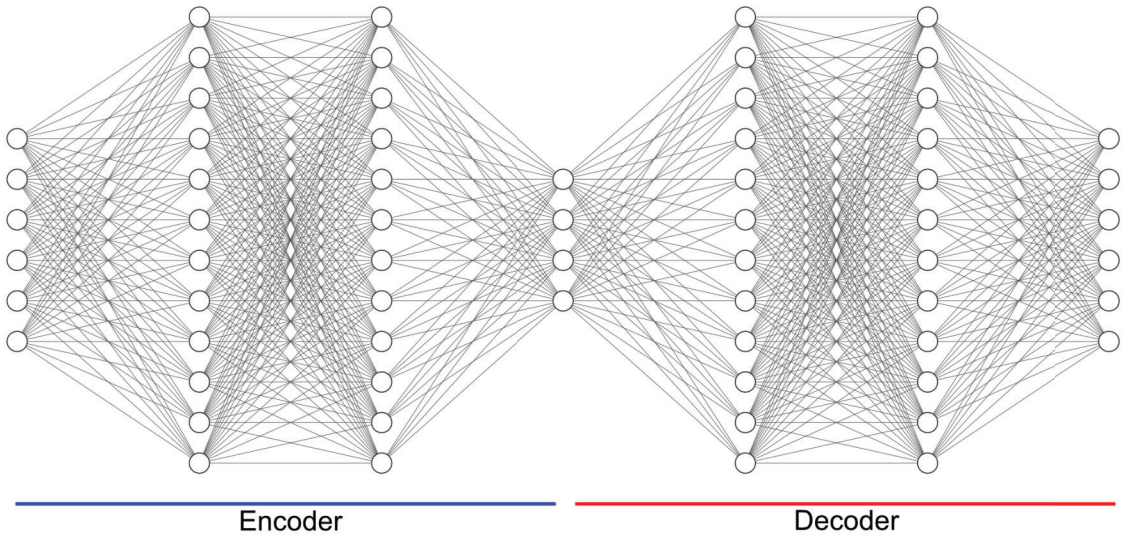

Figure 2.8.: Structure of the conventional two-layer Autoencoder. The encoder (blue) maps the 6-dimensional tracer input through a 12-neuron hidden layer with *Tanh* activation to a latent representation of dimension $r = 4$ via a *Softplus* activation. The decoder (red) mirrors this structure, reconstructing $\hat{x}$ from the latent vector through a 12-neuron hidden layer (*Tanh*) followed by a *Softplus* output layer to enforce non-negativity.

The architecture used here is a fully connected multilayer perceptron with an encoder-decoder structure [62]. The encoder embeds the 6-dimensional tracer vector into a latent representation with dimension r, and the decoder maps the latent space back to the original state space. Hidden layer widths are chosen with cross-validation, with the aim of being small and efficient while allowing non-linear mixing between tracers.

Following the footsteps of Sturm *et al.* [15] et al., the author begins by reproducing the 2-layer deep Tanh→ReLU Autoencoder with 12 neurons while making some mathematical improvements to the Autoencoder structure. The Autoencoder's bottleneck output is set to 4 to conceptually allow 2 latent representations per channel, preventing the network from saturating the degrees of freedom of the dataset. Although ReLU is the conventional choice for non-negative output, it is important to note that non-negativity is achieved by hard-cropping all negative inputs to 0, which in this case will result in mass losses, as we shall see in Section 4.4. Therefore, Softplus [63] is introduced to preserve non-negativity while preventing mass loss at the activation gate or negative values that are introduced by other alternatives such as GELU. Softplus is formulated as

$$\text{Softplus}(x) = log(1 + e^x) \tag{2.22}$$

It is chosen for its non-negativity constraint, which ensures that tracers remain physical. Both the Encoder and Decoder employs a Softplus output node, meaning that non-negativity is explicitly preserved in latent and reconstructed species. However, non-negative activations do not by themselves guarantee mass conservation in the reconstructed state space, such that mass conservation must be handled separately. It must be stressed that Softplus and Tanh also face other numerical challenges, and the author will discuss the possible solutions and improvements to the Autoencoder in Chapter 6.

Given the very small input dimensionality, a shallower network is chosen as the primary architecture. This choice is based on Cybenko [57] and Hornik *et al.* [58], where a 2-layer deep MLP can be seen as a sufficient function approximator in specific settings and functions, as described in Section 2.3.2. Grinsztajn *et al.* [64] argues that while deeper architectures provide theoretical benefits for better generalisability, it does not immediately follow that all datasets would benefit from deep neural networks. Empirical benchmarks support this: a shallow network is desirable for low-dimensional, tabular-style datasets.

A more complex, multi-head decoder is introduced after experimenting with the toy model with a 2-layer deep NN.

2.3.3. Multi Head Autoencoder

The multi-head decoder is inspired by multi-head Neural Networks (e.g. [65–67]). For instance, Zou & Karniadakis [66] showed that multi-head architectures in the L-HYDRA model improve the approximation of partial differential equations by specialising various heads on different solution components or spatial domains. The concept was originally proposed by Caruana [68], in which multiple parallel Dense layers (i.e. heads) were used to

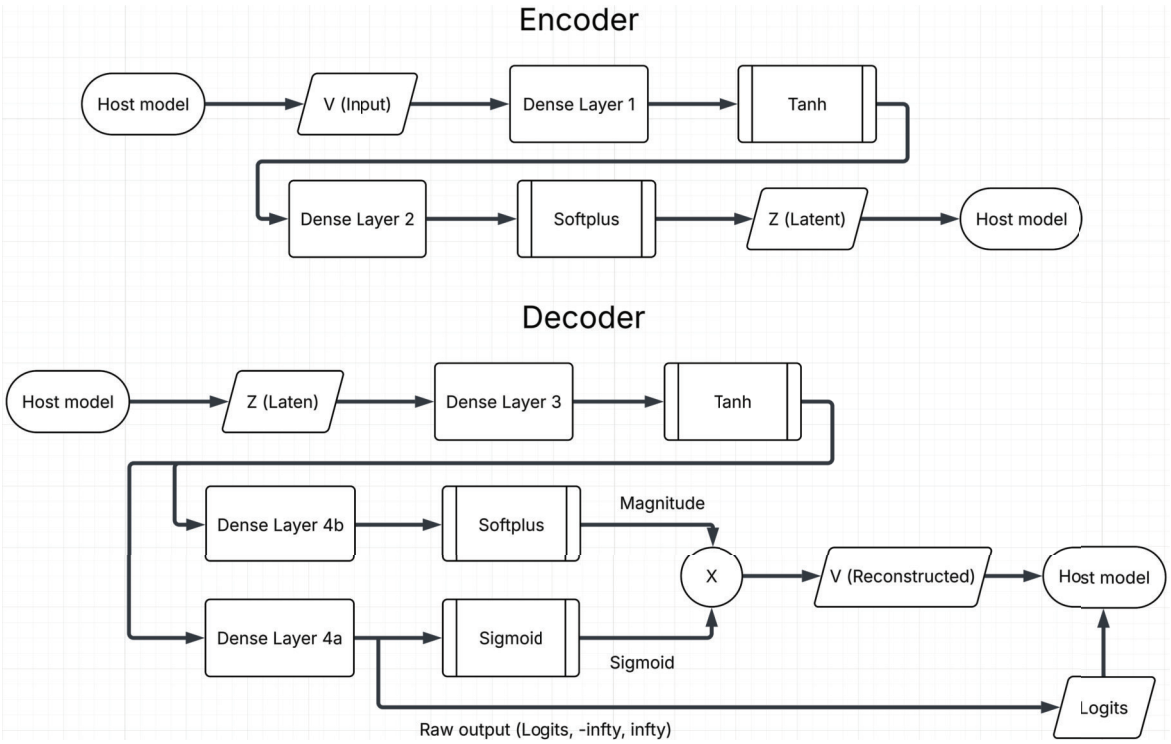

Figure 2.9.: Architecture of the proposed multi-head Autoencoder. The encoder maps input x through two dense layers with *Tanh* and *Softplus* activations to a latent representation z. The decoder reconstructs $\hat{x}$ from z via a shared dense layer followed by two parallel heads: a magnitude head (*Softplus* activation) producing non-negative concentration values, and a non-zero head (*Sigmoid* activation) producing mask probabilities, whose element-wise product yields the final reconstruction $\hat{x}$. Mask logits are retained separately for loss computation.

capture different feature subspaces. Each part of the network is tuned to handle specific tasks, allowing the network to handle a wide range of problems. The dataset in this experiment, as discussed, contains a wide dynamic range. Conventional Autoencoders and loss functions might result in negative values if no explicit constraint is applied. However, explicit constraints such as the ReLU activation function, which crops out negative values, might cause the optimiser to favour negative value outputs in the linear layer and then crop all values to 0. This is because the fastest path to a local minimum in the loss surface would implicitly be connected to local clusters in the state space, if no explicit limit on such behaviour is applied. The multi-head decoder is designed to address this issue by preserving both the magnitude and the non-zero data structure in the reconstructed state space.

The encoder of this network is designed as 2 consecutive Dense layers with Tanh and Softplus activations respectively. The input data is first expanded into a 12-neuron (2× input) layer, fed into the Tanh activation function. The data then passes through a bottleneck layer into its latent representation (4 in this experiment), finally transformed with a Softplus activation function, resulting in a 66% compression ratio.

The author proposes a multi-head decoder (Figure 2.9), where the network splits into 2 heads following the first layer in the decoder. The data is first transformed with a 12-neuron

layer with Tanh activation function. The data is then passed to a magnitude head and a non-zero head.

1. Magnitude head: Data in the magnitude head is first bottlenecked into the size of the original state space (6), then passed through a Softplus activation function. The output (magnitude) is passed to an element-wise multiplication operator to be combined with the output of the non-zero head, resulting in the reconstructed state space $\widehat{V}$. This ensures the non-negativity of the output, constraining unphysical behaviour as Softplus has an output interval of $(0, \infty)$.
2. Non-zero head: Data in the non-zero head is also bottlenecked with the same size, after which a skip connection is passed to the final output as the Mask Logits variable $(-\infty, \infty)$ to be used in the Binary Cross Entropy loss function, discussed below. The data is then passed to a Sigmoid activation function, returning a $(0, 1)$ masked probability of a given element being non-zero. This allows the network to prevent the model from suppressing all tracers to 0, hence addressing the zero-variance problem discussed in Section 4.2.

The skip connection is made because the author employs the *BCEWithLogitsLoss* loss function from Torch [69], where an inherent Sigmoid activation function is applied before the Binary Cross Entropy is computed. Blanchard *et al.* [70] and Nielsen & Sun [71] argue that the log-sum-exp trick improves numerical stability over a separate Sigmoid activation followed by the BCE loss function, as discussed in Section 2.3.4.

The output of the decoder is then the element-wise multiplication of the 2 heads:

$$\widehat{V} = m(z) \cdot \sigma(z) \tag{2.23}$$

where m, σ denote the magnitude and non-zero head respectively.

2.3.4. Multi-objective loss function

A custom loss function combining Mean Squared Error and Binary Cross Entropy [69] is applied in the context of the modified multi-head structure.

The two loss functions act as a balance between the non-zero binary head and the magnitude head. The total loss is computed as a weighted sum of the two components:

$$\mathcal{L}_{MultiHead} = \alpha_1 \mathcal{L}_{WeightedMSE} + \alpha_2 \mathcal{L}_{BCEWithLogits} \tag{2.24}$$

The parameter α_1, α_2 can be viewed as a regularisation term to tune the strength of the Binary Cross Entropy to the Weighted MSE loss.

Weighted Mean Squared Error

In response to the introduction of BCE loss, the loss function for the magnitude head should prioritise magnitude error of the non-zero elements in the source manifold. The author introduced a non-zero weight to the conventional MSELoss, where the non-zero elements are scaled η times:

$$w = \eta \cdot \xi + 1 \tag{2.25}$$

The ξ denotes a boolean matrix of the condition $x > 0$. The Weighted MSE Loss is then defined:

$$\mathcal{L}_{WeightedMSELoss} = w \cdot \frac{1}{m} \sum_{j=1}^{m} (\mathbf{v}_j - \widehat{\mathbf{v}}_\mathbf{j})^2 \tag{2.26}$$

Binary Cross Entropy with Logits

The Binary Cross Entropy is a loss function used for binary classification. The function measures the dissimilarity between the probability distribution of the prediction $\widehat{v}$ and the boolean ground truth with a mapping from $y = [0, 1], x = \{0, 1\} \rightarrow [0, +\infty)$, where 0 denotes a perfect prediction. The probability distribution of a binary classification can therefore be expressed as a Bernoulli probability distribution [72]:

$$P(y \mid x) = p^y (1 - p)^{1-y} \tag{2.27}$$

It is trivial to see that this is a reduced form of the Binomial distribution [72] with C_1^1. Therefore, the negative log-likelihood of this distribution becomes:

$$\mathcal{L}(p, y) = -\log P(y \mid x) = -\log(p^y (1 - p)^{1-y}) \tag{2.28}$$

After moving the exponential term to the front, Binary cross entropy for a matrix is therefore:

$$\mathcal{L}_{BCELoss} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \tag{2.29}$$

y_i denotes the boolean ground truth at i^{th} index, and p_i denotes the prediction at i^{th} index.

BCE Loss of True and False label can be interpreted as follows:

$$\begin{cases} \mathcal{L}_{True} = -\log p \mid y = 1, \\ \mathcal{L}_{False} = -\log(1 - p) \mid y = 0. \end{cases} \tag{2.30}$$

when $y = 1$, the $-\log p$ is minimum, when $p \rightarrow 1$, vice versa.

This has profound importance to the dynamic range and stability of the solution, as the introduction of a 0/non-zero classifier, allows the MSE Loss, for which approaches to 0 at second-order speed Figure 4.9, making it problematic in predicting near-zero values, this proves to be a major problem in the ablation study in Section 4.4.2.

The Log-sum-exp trick

As described previously, this thesis utilises the *BCEWithLogitsLoss*, which is more numerically stable thanks to the log-sum-exp trick [70, 71]. The loss function takes the raw logits ($(-\infty, \infty)$) directly from the non-zero head.

The output is first transformed to $(0, 1)$ via Sigmoid function. Using

$$\sigma(z) = \frac{1}{1 + e^{-z}}, \qquad 1 - \sigma(z) = \frac{e^{-z}}{1 + e^{-z}}, \tag{2.31}$$

it is trivial that for each element in a matrix, the Binary Cross Entropy of the ground truth and the raw logits from a linear layer becomes:

$$\mathcal{L}(z, y) = -y \log(\frac{1}{1 + e^{-z}}) - (1 - y) \log(\frac{e^{-z}}{1 + e^{-z}}). \tag{2.32}$$

Expressing the fraction in terms of exponents, and moving them down:

$$\mathcal{L}(z, y) = y \log(1 + e^{-z}) + (1 - y)(z + \log(1 + e^{-z}). \tag{2.33}$$

Simplifying the terms:

$$\mathcal{L}(z, y) = (1 - y)z + \log(1 + e^{-z}) = z - yz + \log(1 + e^{-z}). \tag{2.34}$$

Given the identity,

$$\log(e^a + e^b) = m + \log(e^{a-m} + e^{b-m}) \qquad | \qquad m = max(a, b) \tag{2.35}$$

By this form, the largest exponent is skewed to 1, and the other exponent confined to < 1. The final BCE Loss function with raw logits as input can be rewritten i.r.t. Equation 2.33 into:

$$\mathcal{L}(z, y) = max(z, 0) - yz + \log(1 + e^{-|z|}), \qquad Q.E.D. \tag{2.36}$$

BCE Loss in this form ensures that no numerical runaway would occur across the entire range of $z \in \mathbb{R}$, resulting in a numerically stable solution.

2.3.5. Optimiser

The Autoencoder is optimised with AdamW as the gradient descent algorithm. Adam optimises the parameters by combining the 1^{st} and 2^{nd} moment estimates of the gradient to obtain an adaptive learning rate for each parameter. Adam is chosen as it converges robustly for smaller networks and requires little tuning compared to other algorithms such as AdaGrad and SGD with Nesterov momentum [73, 74].

Conventional training strategies for a FNN involve defining a fixed set of parameters for the optimiser and the loss function, in pursuit of a global minimum on the cost function

manifold. These parameters are termed "hyperparameters", hence one of the most common and familiar processes: "hyperparameter tuning" [55].

Learning rate scheduling is desirable in many cases, as the initialisation of the parameters in a Neural Network is randomised, leading to a probability that the initialised matrix is located in a local minimum on the loss surface. If the initial learning rate γ is too low, the optimiser might not have enough "momentum", which is determined partially by the learning rate, to free itself from the local minimum. This is also true for other local minima that are not the desired stopping point for the Neural Network.

However, as the training approaches a specific local minimum of interest, and the optimiser approaches horizontally towards the minimum, it will overshoot and might even free itself from the basin, which is no longer desirable. To combat this issue, many gradient descent algorithms, including Adam [73], AdaGrad [50], and Stochastic Gradient Descent [75], were proposed for better convergence in Neural Network training.

AdamW

Name	Symabol	Description
Learning rate	γ	Step size of the optimiser to train the neurons in the backpropogation
Betas	(β_1, β_2)	Scaling factor for computing the gradient and its squared product such that: $m_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1) g_t$ $v_t \leftarrow \beta_2 v_{t-1} + (1-\beta_2) g_t^2$
Weight decay	λ	Similar, but regularisation is stronger compared to L2-regularisation term [49]

Table 2.2.: Hyperparameters of the AdamW optimiser, where g_t is the gradient at step t, m_t and v_t are the first- and second-moment estimates scaled by β_1 and β_2 respectively, λ is the decoupled weight decay regularisation strength [49], and ϱ and φ govern the learning rate schedule.

AdamW is a weight-decoupled variant of Adam [49, 73, 74] in which the adaptive moment updates are retained, while the shrinkage of parameters is decoupled as an explicit weight decay step. In Adam [73], the first and second raw moments of the stochastic gradient $g_t = \nabla_\theta \mathcal{L}(\theta_t)$ are estimated by exponential averages:

$$\begin{cases} m_t &= \beta_1 m_{t-1} + (1-\beta_1) g_t, \\ v_t &= \beta_2 v_{t-1} + (1-\beta_2) g_t^2, \end{cases} \tag{2.37}$$

The bias corrected estimates are then

$$\begin{cases} \hat{m}_t & = \frac{m_t}{1-\beta_1^t}, \\ \hat{v}_t & = \frac{v_t}{1-\beta_2^t}. \end{cases} \tag{2.38}$$

The Adam update step:

$$\theta_{t+1} = \theta_t - \gamma \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{2.39}$$

where ϵ denotes the stability constant (especially important operationally) to avoid division by 0 if $\hat{v}_t = 0$. The benefit of Adam lies in the fact that each parameter has an individual step size determined by a running average of the gradient variance. This results in a computationally efficient yet robust optimiser for high-dimensional and noisy problems [73].

The key in AdamW is that the weight decay is no longer coupled to the L2 penalty of the loss gradient ($g_t \leftarrow g_t + \lambda\theta_{t-1}$). The decay is instead applied directly after each adaptive gradient step. The update is then

$$\theta_{t+1} = \theta_t - \gamma \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} - \gamma\lambda\theta_t \tag{2.40}$$

This reformulation is important, as L2 regularisation is not equivalent to true weight decay in adaptive optimisers. Loshchilov & Hutter [49] showed that the coupling of the regularisation term to the adaptive pre-conditioner entangles the choice of learning rate and regularisation strength, whereas decoupled regularisation yields cleaner optimisation behaviour and often leads to improved generalisation.

Learning rate scheduling

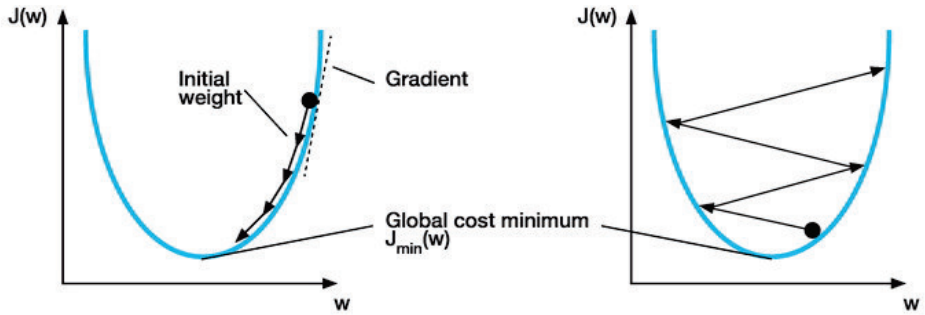

Figure 2.10.: Schematic of proper learning rate in a 2^{nd} order loss surface. Excessive learning rate will result in the optimiser overshooting the minima of the loss surface, whereas proper learning rate allows a smooth, monotonically descending loss towards the minima. Adapted from Ceneda [76]

Learning rate scheduling is therefore proposed as a method of fine-tuning the learning rate such that better convergence can be achieved: high momentum descent to escape

local minima in the beginning, and lower learning rate approaching the basin to prevent overshooting. Loshchilov & Hutter [49] argues that Adam, and its extension AdamW, benefit substantially from learning rate scheduling.

This thesis adopts the *ReduceLROnPlateau* method, where the learning rate is adjusted by

$$\gamma_{new} = \gamma_{original} \times \varrho \tag{2.41}$$

yielding a stepwise descent if the validation loss has stagnated for more than φ epochs.

Hyperparameter tuning

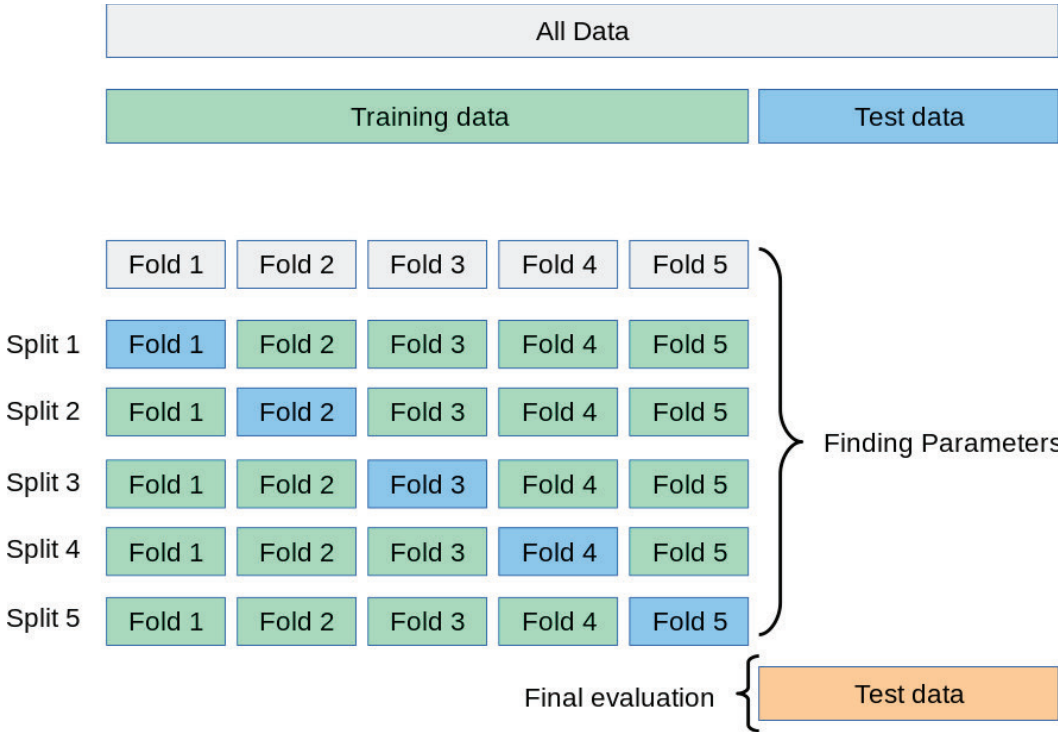

Figure 2.11.: Illustration of 5-fold cross-validation. The training data is partitioned into five equally sized folds; in each of the five iterations (rows), one fold (blue) serves as the validation set while the remaining four (green) are used for training. The held-out test set (orange) remains unseen throughout all cross-validation iterations and is used solely for final model evaluation. Adapted from [77].

Hyperparameter tuning is performed to identify a parameter configuration that yields stable optimisation and good generalisation. In contrast to the trainable weights and biases of the FNN, hyperparameters are prescribed before training and control the optimisation dynamics, model complexity, and regularisation strength. In this thesis, the principal hyperparameters include the learning rate γ, the AdamW coefficients (β_1, β_2), the weight decay λ, the learning rate scheduling factor ϱ, the scheduler patience φ, the batch size, and the latent-space dimension. Since these choices strongly influence both convergence and reconstruction quality, a systematic tuning strategy is required.

	Beta	L1-ratio	Alpha	rank
NMF	Shpae of the map x->d(x), See 2.4.0.1	Ratio of L1-> L2, L2 regularisation=0 and L1 regularisation=1	Strength of regularisation	Number of superspecies
	Neuron size	Bottleneck layer		
Autoencoder	Size of the hidden layer, it determines the depth of the non-linear relationship. Neuron size ≫ feature size means that the model is overparameterised	Size of the projection, determine the size of latent dimension, determines the degree of freedom of the embedding		

Table 2.3.: Hyperparameters of the NMF and Autoencoder models. For NMF: β controls the shape of the beta-divergence loss (Section 2.4.0.1), L1-ratio interpolates between L1 (= 1) and L2 (= 0) regularisation, α sets the regularisation strength, and rank defines the number of superspecies. For the Autoencoder: bottleneck layer size determines the dimensionality of the latent representation and the neuron size determines the ability of the model to capture non-linear relationship

To this end, the hyper-parameters are selected using k-fold Cross Validation [78] with grid search. The dataset is partitioned into k subsets, where in each iteration one subset is retained for validation and the remaining $k - 1$ subsets are used for training. This process is repeated until each subset has served once as the validation set, and the validation performance is averaged across all folds. In this way, the selected hyper-parameter set is less dependent on a single train-validation split and provides a more robust estimate of the expected model performance. The final configuration is chosen based on the best average validation loss, while also considering convergence stability and the absence of pronounced overfitting. The resultant hyperparameters for each model are provided in Appendix A.

Early stopping

To prevent overfitting, early stopping is adopted to improve model generalisability. Overfitting occurs when the model "memorises" the training set, and most importantly the idiosyncratic noise in the training data rather than the underlying, generalisable structure. This typically manifests as a diverging training-validation curve, where the training error dives further, reaching even unity accuracy, while the validation loss, denoting the gen-

eralisability of the model, diverges uncontrollably [55]. Early stopping acts as an implicit regularisation by halting training before the model has fully converged to the training data, so that the complexity of the learned function is well contained to a reasonable degree. Goodfellow [56] argues that early stopping and L2 regularisation for quadratic loss surfaces are formally equivalent. In this thesis, early stopping is triggered when the validation loss starts to diverge from the training loss, or the validation loss has stagnated for over 50 epochs. The training script saves an intermediate checkpoint every 50 epochs.

3. Methods

This chapter discusses the numerical weather prediction model, experimental setup, dataset construction, reduced-order approaches, training strategies and evaluation metrics used in this thesis. The methodological objective is to assess whether a reduced-order representation of the 6 ICON-ART dust species can reproduce transport tendencies with sufficient accuracy while preserving physical admissibility and resulting in a meaningful computational advantage.

Sections 3.1 and 3.1.1 discusses the numerical simulation, alterations to the ICON code base and data generation setup. The author then proceeds to discuss the evaluation metrics (Section 3.2) used in this thesis.

3.1. Experiment setup

This section discusses the controlled numerical experiments used to generate the training and evaluation data, and the different setups employed. The author limits the scope of the thesis to the analysis of tracer advection and embedding in a scientifically interpretable setup, not a fully operational compression strategy for online training.

3.1.1. Idealised Simulation

The training and testing datasets are generated from idealised simulations from the ICON test suite. One no-source-sink experiment is performed to isolate the transport behaviour of tracer fields under controlled conditions. This makes it possible to study the reconstruction accuracy and physical admissibility without immediately conflating with the full source, sink and feedback processes. The author conducts experiments on the following 3 test cases:

The Dust-rad experiment is a standard test case in the ICON test suite (exp.nwp_dust). The simulation uses a R02B06 grid (n_{cell} = 32780 at each height level, equation 2.6) and 90 vertical levels. The effective mesh size of the simulation is therefore 39.5 km (Equation 2.7).

The model is initialised at 22 June 2019 00UTC, with a 6-minute time step. 15 time steps are used to spin up the model, starting from 22:30UTC of the previous day. The simulation has in total 15 spin-up time steps and 480 forecast time steps, ending at 24 June 2019 00UTC.

	Experiment	Description
Full-physics	dust-rad	All source-sink and ART module coupled
No-source-sink	dust-rad	Source-sink processes (emission, deposition, convection and turbulent diffusion) turned off in the source code/XML configuration
ART off	dust-rad	ART module is turned off, the model only advects hydrometeors

Table 3.1.: Summary of the three ICON-ART simulation configurations used in this study. All experiments use the `dust-rad` setup. The *Full-physics* run serves as the reference with all source–sink processes and the ART module fully coupled. The *No-source-sink* run deactivates emission, deposition, convection, and turbulent diffusion, isolating the advective tendency. The *ART off* run disables the ART module entirely, limiting advection to hydrometers only.

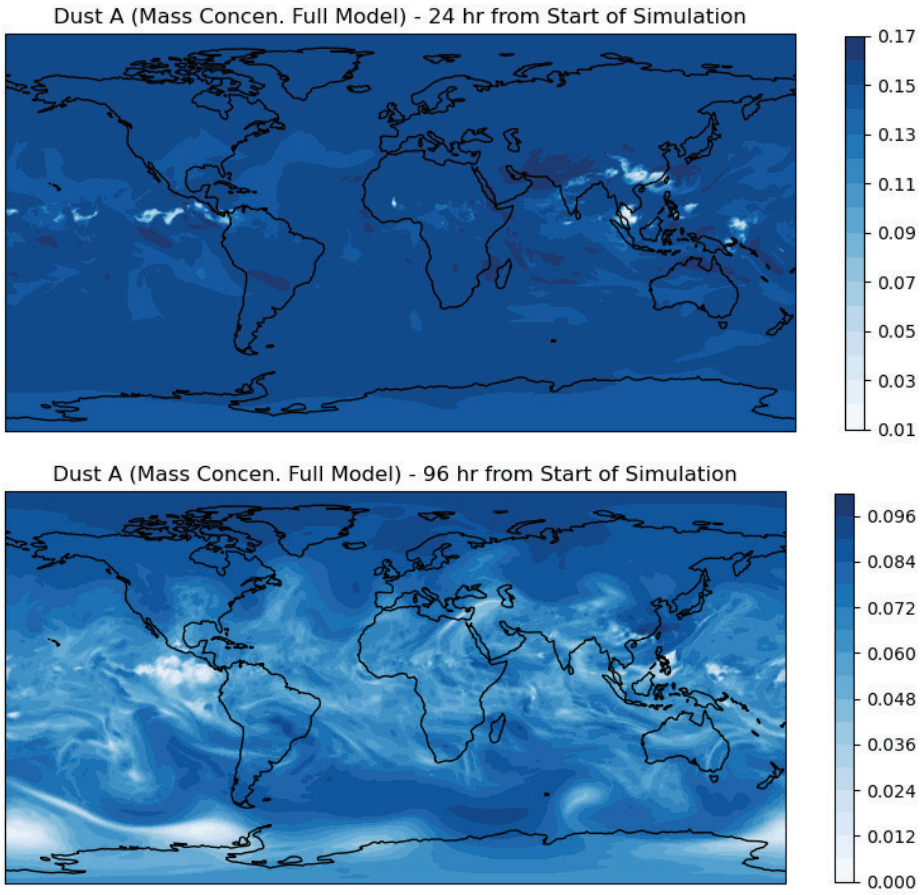

Figure 3.1.: Column-integrated mass concentration of Dust A in the full-physics reference simulation at (top) 24 hr and (bottom) 96 hr after initialisation. The simulation uses the R02B06 grid ($\overline{\Delta x} \approx 39.5$ km) with 90 vertical levels, initialised at 22 June 2019 00 UTC.

3af6059

Some changes have been made to the source code for the no-source-sink ICON binary, such that the tracer update for dry and wet deposition is turned **OFF**. This is because these processes cannot be turned off by namelist options.

ART source code modification for No-source-sink condition

A no-source-sink configuration is introduced to isolate the tracer transport from emission, deposition and other processes. The controlled setup is methodologically meaningful as it provides a clearer target for the embedding model to emulate than the fully coupled tendency field.

Several changes are made to the source code. The author did not remove specific processes directly from the ART source code as this might result in unforeseen bugs, e.g. some other routine referencing the same function. Instead, a $0 \cdot \Delta Mass$ is applied to the update step. Therefore, the time taken to run the wet and dry deposition routines is slightly longer compared to the unaltered source code. Appendix B contains the modified part of the code base.

The Convection and Turbulent Diffusion of aerosol are turned off by modifying the XML file of the tracer:

XML Code

```
<tracers>
  <aerosol id="dust">
    <moment type="int">3</moment>
    <mode type="char">dusta,dustb,dustc</mode>
    <...>
    <iconv type="int">0</iconv>
    <iturb type="int">0</iturb>
  </aerosol>
  <aerosol id="numb">
    <moment type="int">0</moment>
    <mode type="char">dusta,dustb,dustc</mode>
    <...>
    <iconv type="int">0</iconv>
    <iturb type="int">0</iturb>
  </aerosol>
</tracers>
```

Simulation result

The reference simulation is initiated using the conditions above. The simulation is allowed to run freely for 2 days at 6-minute intervals. The synoptic fields of the two scenarios are identical (Figure 3.2). Therefore, the advection results should be identical given that the concentration profiles are identical.

ΔT_{max}: 0.0	ΔT_{min}: 0.0	$\overline{\Delta T}$: 0.0

Table 3.2.: Temperature difference between the full-model and no-source-sink configurations. Maximum, minimum and mean temperature differences across the entire domain are compared. The identical values confirm that disabling source and sink processes does not affect the meteorological fields.

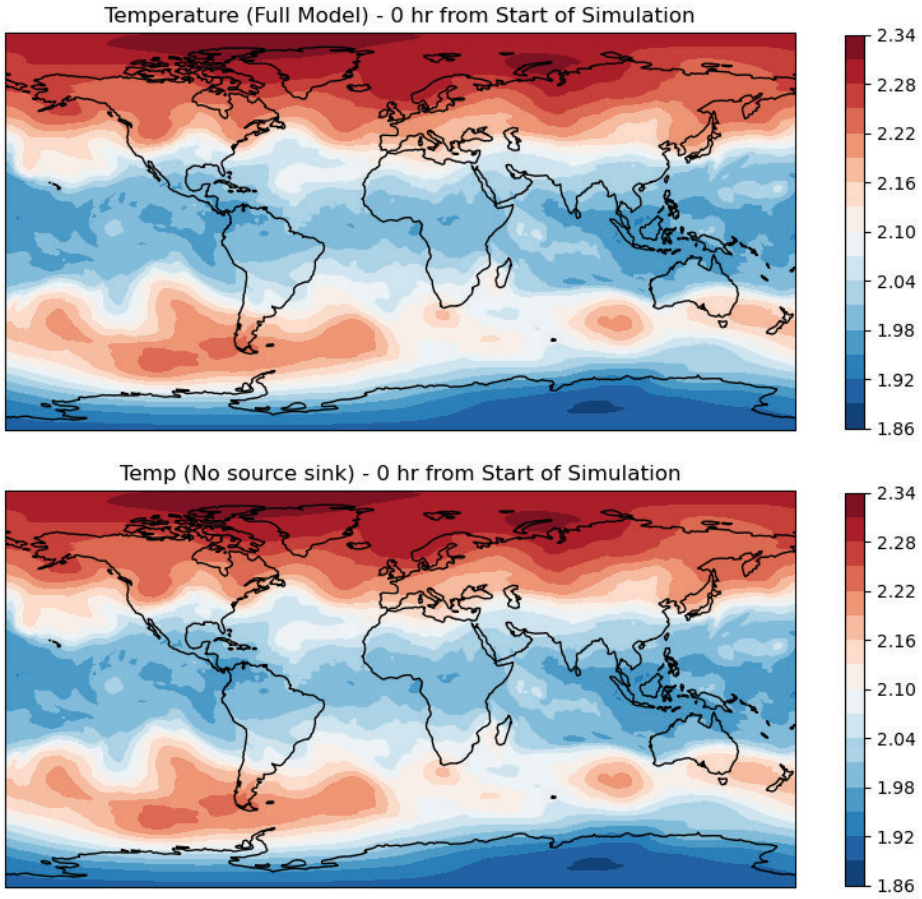

Figure 3.2.: Global Temperature field at height level 40 at 0 hr from the start of simulation. The simulation is interpolated from R02B06 ICON-grid to a Latitude-Longitude grid using CDO. (a) Full Model configuration; (b) Source Sink off scenario.

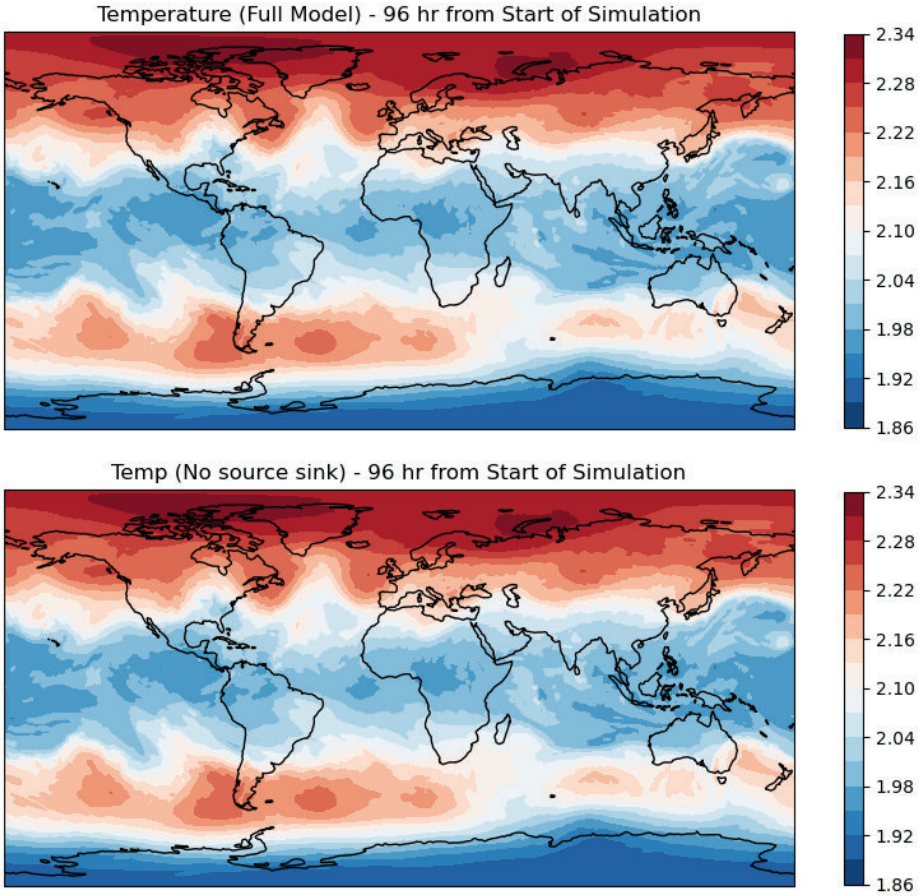

Figure 3.3.: Global Temperature field at height level 40 at 96 hr from the start of simulation. The simulation is interpolated from R02B06 ICON-grid to a Latitude-Longitude grid using CDO. (a) Full Model configuration; (b) Source Sink off scenario.

After source/sink processes are disabled in the modified ICON, the total mass is well conserved. Table 3.2 shows the fractional change in mass between time step 480 and 0. The total change in mass in the full-model simulation ranges from −10% for Dust A to −50% for Dust C. This means that up to half of the total mass injected initially has dissipated. This is because Dust C's diameter (**d_gn**) is $O(1)$ larger than Dust A, and since volume scales as r^3, the deposition and sedimentation are much faster for Dust C.

In the modified ICON model, the Δmass is $O(3)$ smaller compared to the full model, indicating that this strategy has successfully isolated the aerosol from source/sink conditions. However, it is also worth noting that, as sedimentation, convection, turbulent diffusion and deposition are turned off, the vertical mass flux is minimal compared to the full-model scenario. Figure 3.4 shows the zonal mean mass concentration of Dust A. The full model shows a well-mixed mass concentration, while the NSS scenario shows a concentration distribution that is only centred at the height level where the aerosols are initialised. The difference is perhaps due to the 15 spin-up time steps (reanalysis) that are run before the first output time step (+0h), where the full model is allowed to mix freely.

At the end of the simulation (+96h), the aerosol in the full model is mixed polewards, as visible in Figure 3.4. In NSS, only small amounts of aerosol are advected downwards compared to the full model. This indicates that vertical transport processes other than advection are turned off, as turbulence and convection are the major contributors to vertical mixing.

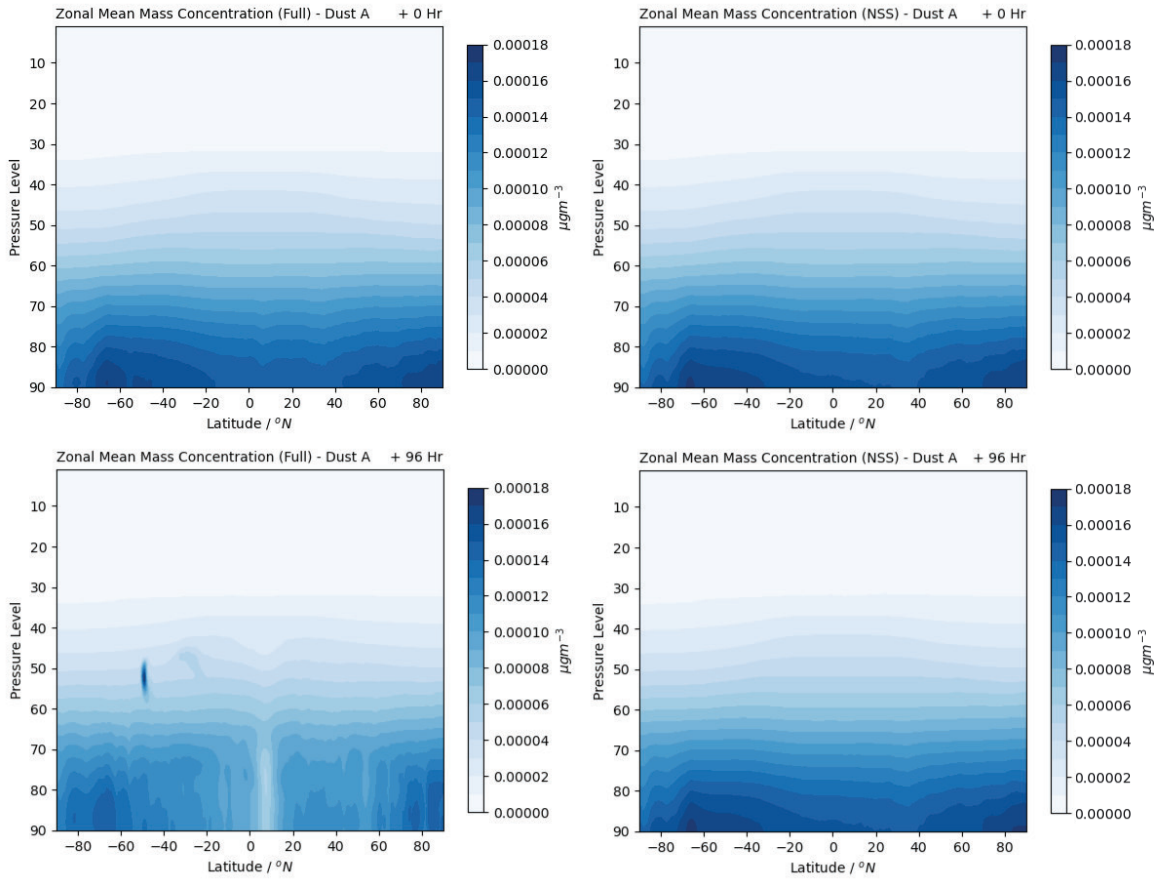

Figure 3.4.: Zonal-mean mass concentration of Dust A at the start of the simulation (top) and 96 hr after initialisation (bottom). The left column shows the full-model configuration and the right column the no-source-sink configuration. The vertical axis denotes the model height level.

Simulation configuration

This section discusses the computing cluster, I/O configurations and training data storage. The author mainly discusses how the training data is extracted from several entry points in the NWP simulation steps, and addresses issues in extracting data from a parallelised model.

The simulation and training are done on the Levante Platform of the German Climate Computing Centre (DKRZ), where both GPU and compute node are used. The Levante system has the following configuration:

Node	Configuration	Feature
GPU	AMD 7763 CPU + Nvidia A100 (40/80GB)	Nvidia Tensor core GPU suitable for ML training
compute	AMD 7763 CPU (128 Core) per node	256-1024GB Memory

Table 3.3.: Configuration of the DKRZ Levante platform. Two node types are used: GPU nodes for Autoencoder training and compute nodes for NMF training and ICON-ART simulations.

GPU nodes are used for Autoencoder training and compute nodes are mostly used for NMF training. Further GPU deployment of the machine learning model is feasible, however, the host model (CPU) – machine learning model (GPU) I/O bottleneck may be severe, and must be evaluated carefully.

The dataset is extracted using the ComIn interface [79]; the raw, unaltered Fortran pointer has a dimension of (t, jg, hlev, jc) for each species. This is the raw pointer from each of the MPI processes. As the model is parallelised, I/O of the raw pointer would lead to different MPI threads overwriting the same output path.

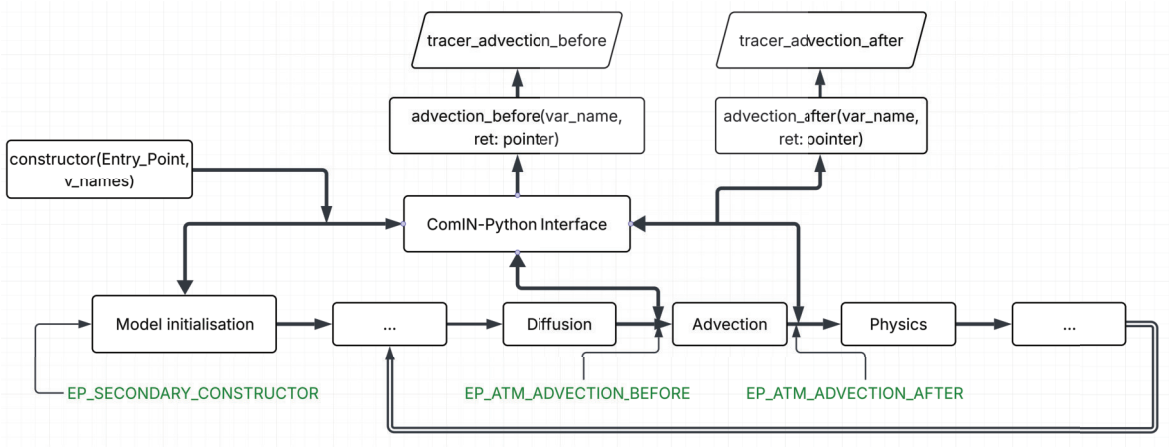

Figure 3.5.: Flowchart of training data retrieval with the ComIn interface. Variable exposures are requested in the secondary constructor before model initialisation, specifying the entry point and variable names. At the designated entry point, a callback function receives the pointer to the exposed internal variable via `np.asarray(var)`. The tracer data are stored in Python global memory for post-processing at the end of the simulation. All communication between the Python functions and ICON proceeds through the ComIn-Python interface.

To address this issue, a unique, random UUID is generated by the *uuid.uuid*4() routine at the *EP.SECONDARY_CONSTRUCTOR*, i.e. the initialisation phase of the simulation. The first 17 digits of the UUID are then assigned as the unique identifier of each thread. Raw pointers at each time step are passed to a *numpy.ndarray*, stacked along a new axis t as the simulation iterates. The resultant arrays are stored on the cluster with unique names for further processing such as random shuffling. Figure 3.5 shows the schematic flowchart of the training, validation and testing dataset preparation. All communications and data exchange are done through the ComIn-Python interface in an independent namelist *comin_nml*.

To support model development and evaluation, three simulation scenarios are used: a full-model configuration, a no-source-sink configuration, and an ART-off reference simulation (Table 3.1). These scenarios serve different purposes in the workflow. The no-source-sink simulation provides a controlled setting for learning and testing transport-dominated behaviour, the full-model simulation is used to assess robustness under more realistic conditions, and the ART-off simulation provides a computational reference for estimating the tracer-related overhead. The dataset is then partitioned along the time axis into training, validation, and test subsets, with shuffling applied only after the time split is performed to avoid data leakage resulting in overfitting.

3.1.2. Dataset preparation

Training, validation and test

This section discusses how raw output data are transformed into a training-ready dataset for training, validation and testing. This ensures the embedding and its inverse are trained on a clearly defined target without the risk of leakage and with better generalisability.

Both the no-source-sink and full-model datasets are split along the time axis, with the first 300 time steps used for training and validation, and the last 100 time steps used for final evaluation of the model performance. As the dataset is highly correlated spatially and temporally, a full random shuffle of the dataset would inevitably lead to data leakage. Therefore, the training, validation and testing datasets are shuffled randomly after the data are split along the time axis.

Finally, the first 300 time steps are split into 200:100 for training and validation, then a random shuffling of 5 Mil:3 Mil samples are done for the 2 datasets respectively. The last 100 time step is then shuffled randomly for 2 Million samples, yielding the testing dataset. This strategy segregates the training, validation and testing dataset by its temporal axis, preventing data leakage during random shuffling. Final evaluation (Section 5) will be performed on the testing dataset to examine the generalisability of the model.

Output variables for emulation

The dataset consists of prognostic aerosol tracers extracted from the ICON-ART dust experiment. In the present test case, 6 tracers are considered: three mass concentrations and three corresponding number concentrations for the dust size classes A, B and C.

The final, training ready dataset is stored as a *torch.tensor* after passing the randomly shuffled *np.ndarray* as $torch.tensor(array)$. Let m denote the number of tracer variables and n the number of samples. One sample corresponds to a single grid-cell state (or, more generally, a single spatial sample at a given output time), such that the tracer data matrix is defined as

$$\mathbf{V} \in \mathbb{R}_{\geq 0}^{m \times n}. \tag{3.1}$$

Each column of $\mathbf{V}$ represents the full tracer state for one sample, and each row contains the spatial distribution of one tracer across all samples.

Because mass and number concentrations differ substantially in magnitude and statistical distribution, the raw dataset is strongly imbalanced and heavily long tailed. This has direct implications for optimisation, because unscaled least-squares losses tend to be dominated by variables with the largest absolute magnitude. For this reason, the error analysis in this thesis relies on normalised metrics in addition to absolute metrics.

Relation between mass and number concentration

For an idealised mono-disperse particle population with particle diameter d_p and particle density ρ_p, the conversion between mass concentration C_m and number concentration C_n can be written as

$$C_n = \frac{C_m}{\rho_p \frac{\pi}{6} d_p^3}. \tag{3.2}$$

Equation (3.2) is shown here only to illustrate the physical link between the two quantities. In the operational aerosol module, however, the relationship between mass and number concentration may be more complex because it depends on the assumed size distribution, particle density, and modal or sectional representation. Accordingly, the two tracer types are treated as distinct prognostic variables in the reduced-order modelling framework.

3.2. Performance metric

To compare across different methods used in this thesis, the reconstructed state spaces are evaluated with error metrics independent of the training loss. This allows a consistent comparison across different models.

Let $y_i, \widehat{y_i}$ denotes each elements of the input space and reconstructed space with i being the element in each sample where $i \in \mathbb{Z} \bigcap [1, n]$.

3.2.1. Error statistics

Root mean square error

The Root Mean Square Error (RMSE) is defined as

$$RMSE = \sqrt{\frac{1}{n} \sum^{n} (y_i - \hat{y}_i)^2} \tag{3.3}$$

Root Mean Square Error (RMSE) is sensitive to extreme values as the $(\bullet)^2$ operator scales quadratically with initial error. Therefore it highlights cases for which the embedding fails strongly in localised regions.

Normalised root mean square error

Tracers, particularly mass and number concentrations, have very different magnitudes; the RMSE is therefore also reported in normalised form, such that errors scale in the same manner across different species. The normalised RMSE [80] is defined as

$$\mathrm{nRMSE} = \frac{\mathrm{RMSE}}{\overline{y}}, \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \tag{3.4}$$

Mean absolute error

The mean absolute error Mean Absolute Error (MAE) is defined by

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{3.5}$$

Compared to RMSE, MAE is less sensitive to outliers.

Normalised Mean absolute error

Analogous to Equation (3.5), the normalised mean absolute error is defined as

$$nMAE = \frac{MAE}{\overline{y}} \tag{3.6}$$

Bias

To quantify systematic over- or underestimation, the mean bias is defined as

$$\mathrm{Bias} = \frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - y_i)\,. \tag{3.7}$$

These metrics together provide a systematic assessment of the quality of the embedding and its inverse transform across tracer species, latent representations and experimental configurations.

Skewness

To assess whether the reconstructed distribution preserves the asymmetry of the original tracer field, the skewness is considered. Skewness quantifies the degree of asymmetry of a distribution about its mean by evaluating the 3^{rd} standardised moment. For a sample of size n, it is defined as

$$\text{Skewness} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \bar{y}}{\sigma_y}\right)^3, \tag{3.8}$$

where $\bar{y}$ denotes the sample mean and σ_y the sample standard deviation. A positive skewness indicates a distribution with a longer right tail, whereas a negative skewness indicates a longer left tail. A value close to zero suggests that the distribution is approximately symmetric.

In the context of tracer reconstruction, skewness is a useful complement to error-based metrics such as RMSE or MAE, since it evaluates whether the reconstructed field retains the distributional asymmetry of the original data.

Kurtosis

In addition to asymmetry, kurtosis measures the extent to which a distribution is peaked around its mean and the degree to which it exhibits heavy tails. It is defined as the 4^{th} standardised moment:

$$\text{Kurtosis} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \bar{y}}{\sigma_y}\right)^4. \tag{3.9}$$

Excess kurtosis is also used, which is obtained by subtracting 3 from the kurtosis such that a Gaussian distribution has zero excess kurtosis.

Kurtosis provides a measure of whether the reconstruction preserves the occurrence of concentrated peaks and rare extreme values. A lower kurtosis in the reconstructed field may indicate that the model smooths out sharp tracer plumes, whereas a higher kurtosis may suggest the artificial amplification of local extremes.

Together, skewness and kurtosis complement the previously introduced metrics by extending the evaluation from pointwise reconstruction accuracy to the statistical structure of the tracer distribution. They therefore provide additional information on whether the latent embedding and inverse transformation preserve not only the mean magnitude of the tracer fields, but also their asymmetry and tail characteristics across species and experimental configurations.

3.2.2. Manifold metrics

To quantitatively analyse the quality of the reconstruction, the author proceeds to assess the distance between the source and reconstructed manifolds using 3 metrics: Anomaly

Correlation Coefficient (ACC), Multi-scale Structural Similarity Index Measure (MS-SSIM, or SSIM) and Wasserstein distance.

Anomaly Correlation Coefficient

The anomaly correlation coefficient quantifies the linear relationship between the spatial pattern of the source and target space, independent of bias and amplitude, for the ground truth y and the target $\widehat{y}$, the ACC is defined

$$\mathrm{ACC} = \frac{\sum_{i=1}^{n} (y_i - \bar{y}) \left(\hat{y}_i - \bar{\hat{y}}\right)}{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2 \sum_{i=1}^{n} \left(\hat{y}_i - \bar{\hat{y}}\right)^2}}, \tag{3.10}$$

The ACC is bounded in $[-1, 1]$ where -1 and 1 denote perfect negative and positive pattern correspondence. ACC is insensitive to uniform scaling or offsets, and is therefore comparable across data of various dynamic ranges, especially in Numerical Weather Prediction [81].

Normalised Wasserstein Distance

Wasserstein-1 Distance [82] measures the distributional fidelity of the reconstructed field. For a 1-D distribution with cumulative distribution functions (CDF) F & $\widehat{F}$, the Wasserstein Distance is defined as:

$$W_1\left(F, \hat{F}\right) = \int_{-\infty}^{\infty} \left|F(t) - \hat{F}(t)\right| \, \mathrm{d}t \, . \tag{3.11}$$

W is not scale invariant and unbounded, rescaling is necessary in order for the measure to be compared across species at different orders of magnitude. The distribution is then rescaled by the interquartile range (IQR), capturing the spread of 50% of the sample. The interval of the rescaled Wasserstein distance is still bounded in $[0, \infty)$, with 0 indicating a perfect geometric reconstruction, larger values denoting greater distributional shift.

4. Result

This chapter discusses the training results of the baseline PCA, non-negative matrix factorisation and various Autoencoder structures. The results demonstrate that, without proper non-zero and non-negativity constraints, the model performs particularly poorly in boundary conditions, e.g. in data close to 0. High variance was observed in model predictions near 0, which is attributed to the lack of constraint on the zero/non-zero classification. This pattern is observed in both linear (non-negative matrix factorisation) and non-linear (Autoencoder with and without activation function) methods.

This observation sparks the need of a multi-head constrain, where the magnitude head would predict the absolute magnitude of the model output, while the non-zero head predicts the probability of each species at a particular cell being 0. The detailed discussion on the architecture can be found in Section 2.3.3. All hyperparameters of each model are provided in Appendix A.

4.1. Baseline - Linear PCA

	4 Latent dimension	3 Latent dimension	2 Latent dimension
		Normalised RMSE	
Species 1	0.1893 %	0.1830 %	1.0119 %
Species 2	1.1370 %	1.2834 %	12.4639 %
Species 3	0.6212 %	3.0002 %	15.6078 %
Species 4	0.1892 %	0.1830 %	1.0263 %
Species 5	1.1352 %	1.2802 %	12.5513 %
Species 6	0.6171 %	3.0048 %	15.4729 %
Total	1.0207 %	2.1888 %	14.2321 %

Table 4.1.: Normalised RMSE of the baseline linear PCA at three latent dimensions. The PCA is computed via Singular Value Decomposition, where the dataset is factorised into $\mathbf{U}, \Sigma, \mathbf{V}$ and the number of retained components determines the effective compression ratio. This model serves as the theoretical lower bound for reconstruction error across all embedding approaches.

Table 4.1 describes the normalised RMSE of the reconstructed state space of the PCA. The PCA is fitted by the training dataset and the reserved component V^T is the "latent dimension".

The fitted PCA is then used to project the testing dataset to the latent space, then perform an inverse transform, written in matrix form:

$$\mathcal{R}(\mathcal{P}(X)) = \mu + (X - \mu)V_k V_k^T \tag{4.1}$$

The PCA, although performed extremely well up to 3 latent dimension, is an unconstrained basis vector rotation to the direction of the maximum variance, there are no physical interpretability or non-negativity preservation in such method. Nevertheless, PCA is very useful to evaluate the intrinsic degrees of freedom of the data, that is, the minimum required "sub-space" to represent the data. The result from 2 Latent dimension drift substantially from that of 3/4 latent dimension, indicating that the 2 latent dimension is insufficient to represent the data, indicating that the latent dimension should be chosen as >2.

4.2. Non-negative Matrix Factorisation

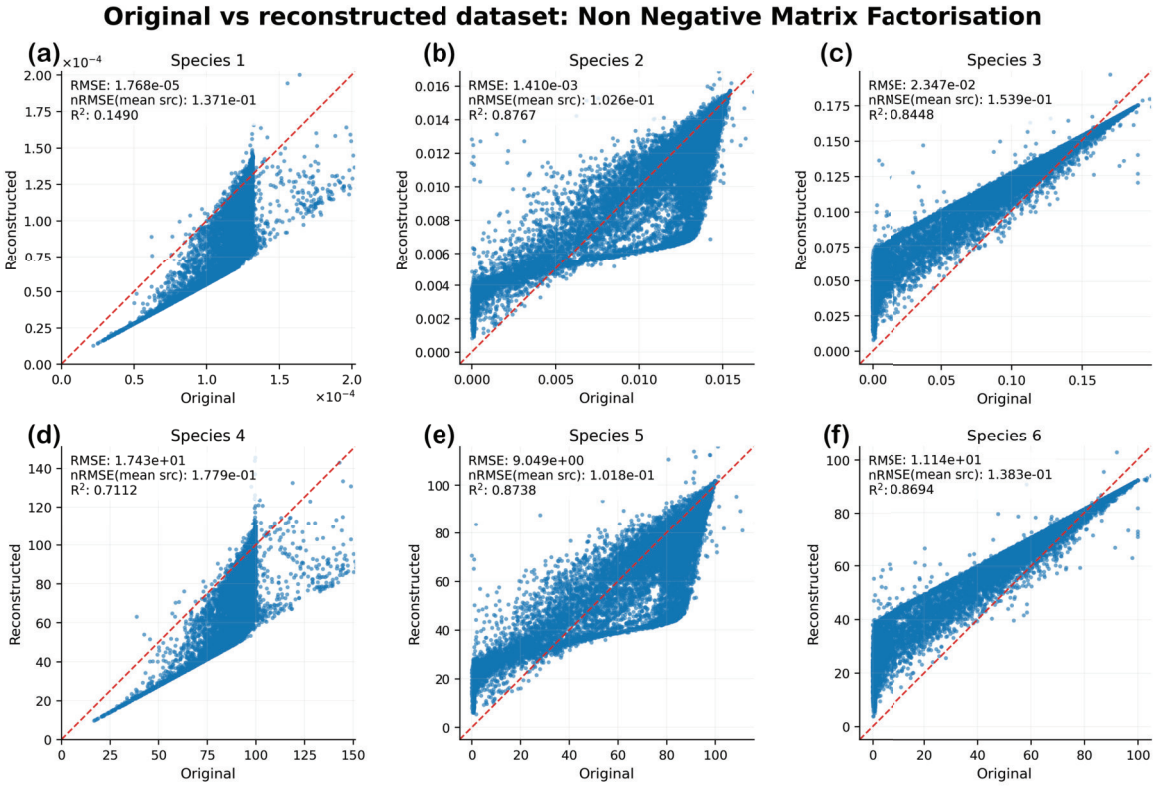

Figure 4.1.: Scatter plot of Non-negative Matrix Factorisation prediction and ground truth. The red dashed line denotes perfect 1:1 agreement. RMSE, nRMSE (normalised by source mean), and R^2 are shown per panel. NMF yields good reconstruction for Species 2, 4, 5, and 6 (R^2 = 0.71–0.88, nRMSE ~ 0.10–0.18), with Species 1 remaining the most challenging (R^2 = 0.15). Compared to the multi-head Autoencoder, NMF substantially improves Species 4 (R^2 = 0.71 vs. 0.06) while performing comparably for the remaining species.

Figure 4.1 shows a cluster of scatter plot of the original-reconstructed evaluation set of each species. The scatter plot was generated by first oversample a random candidate set, then fill the first part of the candidate with the highest magnitude pair, then the remaining candidates were filled randomly by *numpy.random.Generator().choice()*. This not only ensures the scatter plot routine not be oversaturated by the full dataset (up to 20GB per species), while preserving the general structure of the high magnitude domain. This strategy will be used to generate all subsequent scatter plots.

Number and mass concentration exhibit a very similar data structure for each dust species. Dust A shows a systematic underestimation across the entire spectrum of values, with a few samples being outliers, spiking above the slope-1 line at ~ 100 (number) and ~ 1.3 (mass). The estimator seems to exhibit a triangular structure, where an implicit underestimation limiter is found at $slope \approx 0.6$. For instance, in number concentration dust A (Figure 4.1 c), the prediction value are essentially confined within the area within the lines:

$$\begin{cases} y = x, \\ y = 0.6x - 0.05. \end{cases} \tag{4.2}$$

This resulted in an increasingly widening variance.

In Dust B (Figure 4.1 b,e), the model exhibits no apparent systematic bias in prediction. The variance of prediction seems to be more random compared to Dust A, while the centroid and linear fit of the scatter plot lie almost parallel to the slope-1 line, meaning that the model successfully reconstructed the general structure of the dataset while struggling to address the variance of the prediction. A slight under- and overestimation are visible in the lower and upper halves of the spectrum respectively.

Dust C (Figure 4.1 c,f) shows a systematic overestimation that is inversely proportional to the absolute magnitude of the source concentration. A similar implicit overestimation limiter is also observed, where the variance of prediction decay with increasing source magnitude, confined within the slope-1 and line:

$$\begin{cases} y = 0.5x + 40, \\ y = x. \end{cases} \tag{4.3}$$

Here a pattern can be observed: for each group of species, the transition from under- to overestimation seems to be associated with the increasing order of magnitude in the mean of the distribution. This means that the model might have learnt a somewhat normal/-Gaussian distribution, where the model prediction leans towards the region of maximum frequency/centroid of the distribution.

In addition, a strong variance is concurrently observed in 4.1 b-c, e-f, suggesting a high variance at $x = 0$. This indicates the model is experiencing difficulty distinguishing 0-elements from non-zero element, suggesting a need for $\{0, \backslash 0\}$ binary classifier to be introduced in the loss function. This could potentially address the zero variance problem by constraining such behaviour.

The non-negativity is well preserved, as fundamentally the **Non-negative** matrix factorisation explicitly preserve non-negativity in it's multiplicative update step in Equation 2.15.

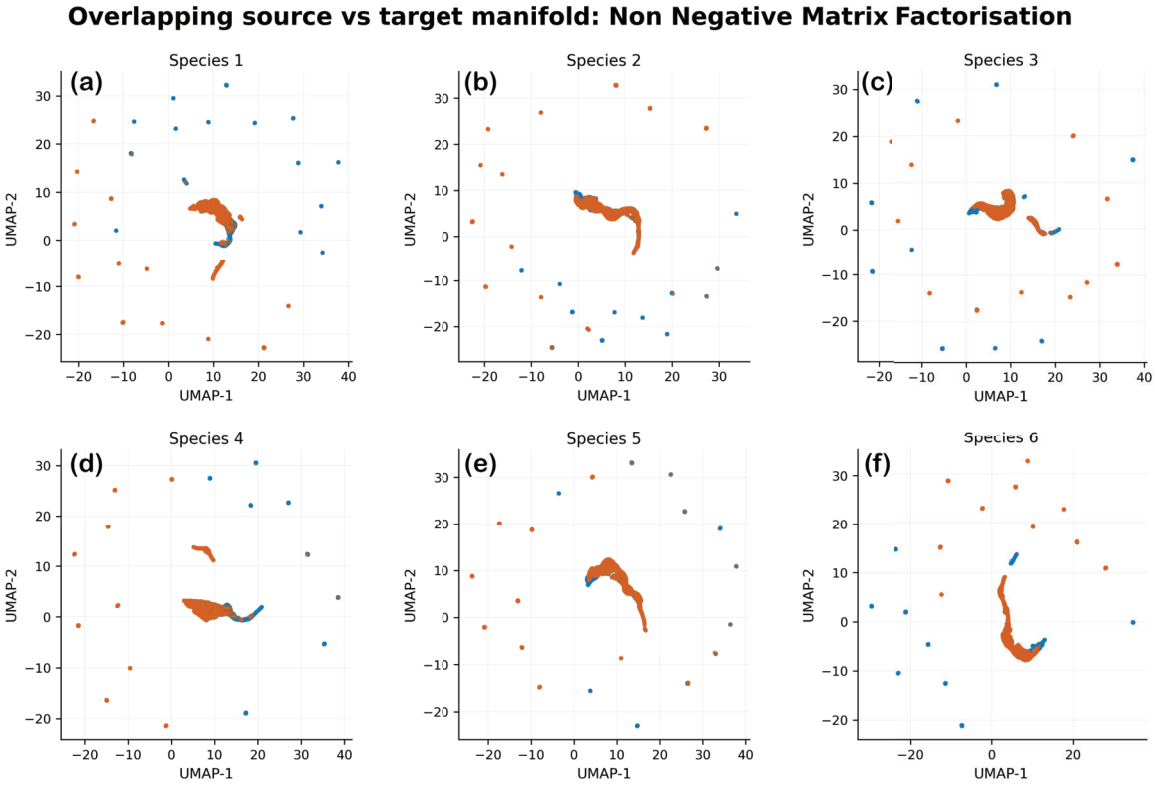

Figure 4.2.: Stacked UMAP projection of the source and reconstructed state space using Non-negative Matrix Factorisation. Each panel shows the two-dimensional UMAP embedding for one species: (a-c) mass concentration and (d-f) number concentration of Dust A, B and C respectively. The source manifold (blue) and NMF reconstruction (orange) are overlaid to visualise geometric agreement. Good overlap indicates faithful manifold preservation, while discrepancies highlight regions where the reconstruction departs from the source distribution.

Figure 4.2 shows the stacked UMAP projection of the source and reconstructed state space. Good agreement is observed in the central, high-density curved branch, indicating that the NMF reconstructs the general distribution of the manifold well. However, it is clear that the low frequency, scattered extreme values were poorly represented. This agrees with the previous observation in Figure 4.1, that the reconstruction tends to project to a somewhat normal distribution centred in dust B. Such tendency is also visible in the centre branch, with the reconstructed state space sits on the centre of the main branch, while failing to reconstruct the tails of the centre branch. Best agreement can be found in dust B (Figure 4.2b,e), agreeing to the observation in figure 4.1 b,e. Some outliers were captured in dust B, as observed in the lower right quadrant in Figure 4.2 b,e, meaning that the dust B has the best reconstruction quality.

Figure 4.3 shows a composite histogram of the UMAP distance and element wise RMSE between the source and reconstructed manifold. Ideally, the the RMSE and UMAP should

stay close to 0, with the source and target UMAP projection overlaying each other. Note that for UMAP, the $\widetilde{UMAP} \ll \overline{UMAP}$, indicating that masses are concentrated around 0-5, with outliers forming a heavy long tail, reaching 50-60.

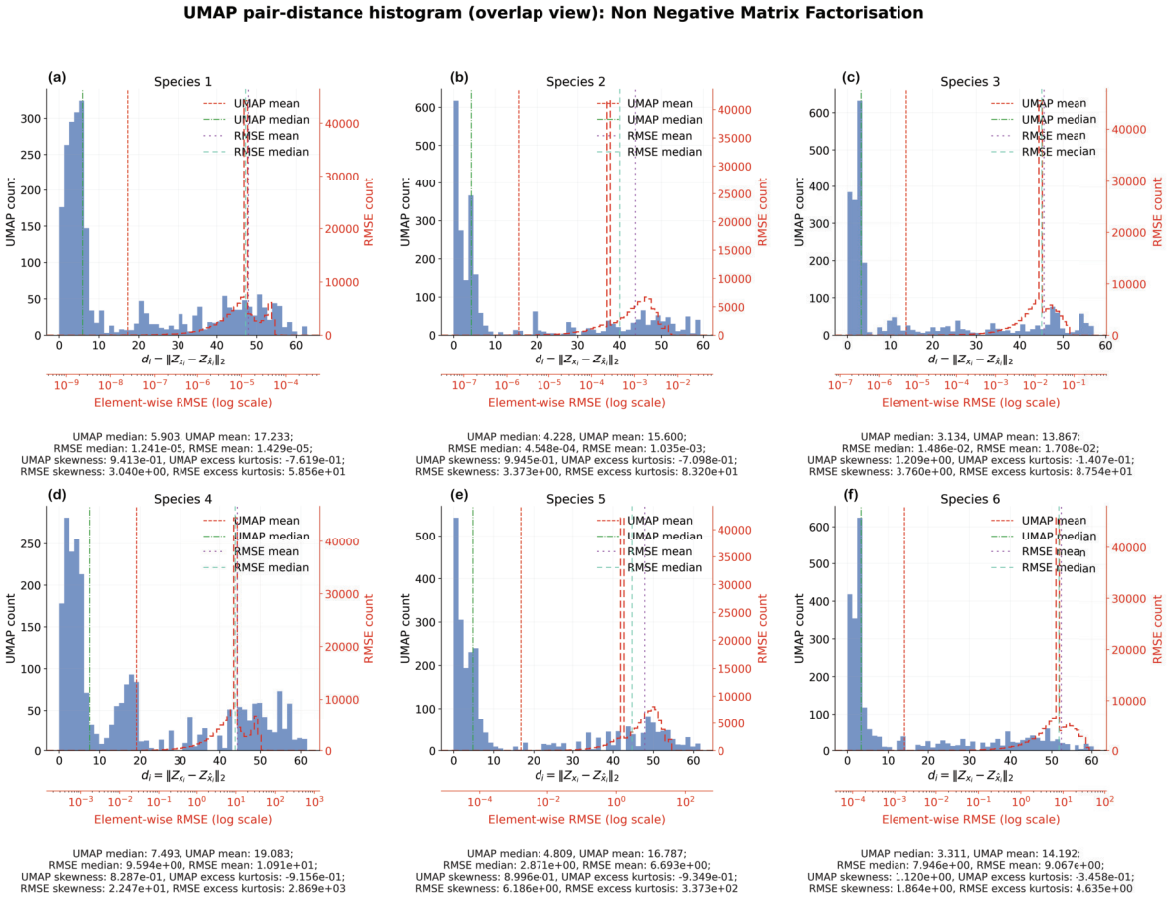

Figure 4.3.: Composite histogram of UMAP distance and element-wise RMSE for the NMF reconstruction. The left y-axis denotes the UMAP distance frequency between the source and reconstructed projections, and the right y-axis denotes the element-wise RMSE frequency (log-scaled). Panels (a-c) show mass concentration and (d-f) number concentration of Dust A, B and C respectively. Vertical dashed lines indicate the median ($\tilde{\cdot}$) and mean ($\bar{\cdot}$) of each distribution.

The RMSE histogram (Figure 4.3 d-f) indicates that the model has relatively good agreement on the magnitude side. The median and mean are relatively close together, indicating that outliers do not have a strong influence on the distribution. The difference in order of magnitude of the RMSE are attributed to the difference in order of magnitude of the actual species concentration rather than the fundamental difference in reconstruction fidelity. When scaled by the characteristics of each species, the reconstruction error is comparable across all 6 species (Figure 4.1 indicates that the normalised RMSE for all species are at the order of magnitude of 10^{-1}).

This agrees with the previous observation, that the model reconstructs the manifold structure reasonably well, while lacking the ability to reconstruct the peripheral states in the manifold

(edge cases). The poor performance may partly contributes to the observed 0 variance, which must be addressed in further development.

In conclusion, the current NMF model reconstructs the dominant structure of the manifold relatively well; the absolute magnitude is also relatively well reconstructed. However, it lacks the ability to capture extreme values and boundary states in the manifold. The most prominent issue observed is the zero-variance problem, where the model predicts non-existent species concentrations, possibly due to a lack of binary constraint (zero/non-zero). This will further result in "ghost" mass, particularly when the model is coupled with a host model and run iteratively.

4.3. Gradient based PCA – Autoencoder with no activation function

As established by Bourlard & Kamp [60], a single-layer Autoencoder with no activation functions recovers the PCA subspace. This thesis therefore experiments whether gradient-based optimisation can recover the same reconstruction quality as the closed-form SVD solution in baseline PCA. It further isolates the effect of optimisation from the effect of non-linearity, which will be introduced in later sections.

The most notable features lie in the mass concentration of Dust A and B (Figure 4.4 a-b, Species 1-2), where Species 1 exhibits a catastrophic failure, with the reconstructed values collapsing into a narrow vertical band, while the predicted values span $\{-6 \times 10^{-3}, +5 \times 10^{-3}\}$. This is most likely a scaling issue, as the range of input data is at $< 10^{-3}$, and the reconstructed data range is 10× the input range, compressing the x-axis to almost 0. The model produces physically meaningless negative concentrations, which is a direct consequence of the absence of an activation function. Without non-negativity constraints, the linear map has the interval $(-\infty, \infty)$. The negative R^2 indicates the model is actually worse than a constant mean predictor, meaning that the linear Autoencoder has failed to learn any meaningful structure at this order of magnitude.

Mass concentration Dust B (Species 2) exhibits a moderate degradation. The model exhibits a systematic underestimation, where the centroid of the distribution lies slightly below the slope-1 line. A discernible broadening of the variance is observable across the full range of the source space, where the variance in the zero and near-zero domain increases dramatically—a "zero-variance" problem. This indicates the inability of the model to distinguish zero values from non-zero elements, resulting in large variance in this domain.

The result is surprising, in contrast to the baseline PCA (Table 4.1), which achieves 0.19% nRMSE for mass concentration Dust A. Although Bourlard & Kamp [60] proved that a single-layer linear Autoencoder with no activation function is mathematically equivalent to a truncated PCA, both being linear projections, the optimisation differs substantially. PCA obtains the global optimal rank-k approximation from a closed form SVD, which decomposes the covariance structure simultaneously, assigning each Principal Component

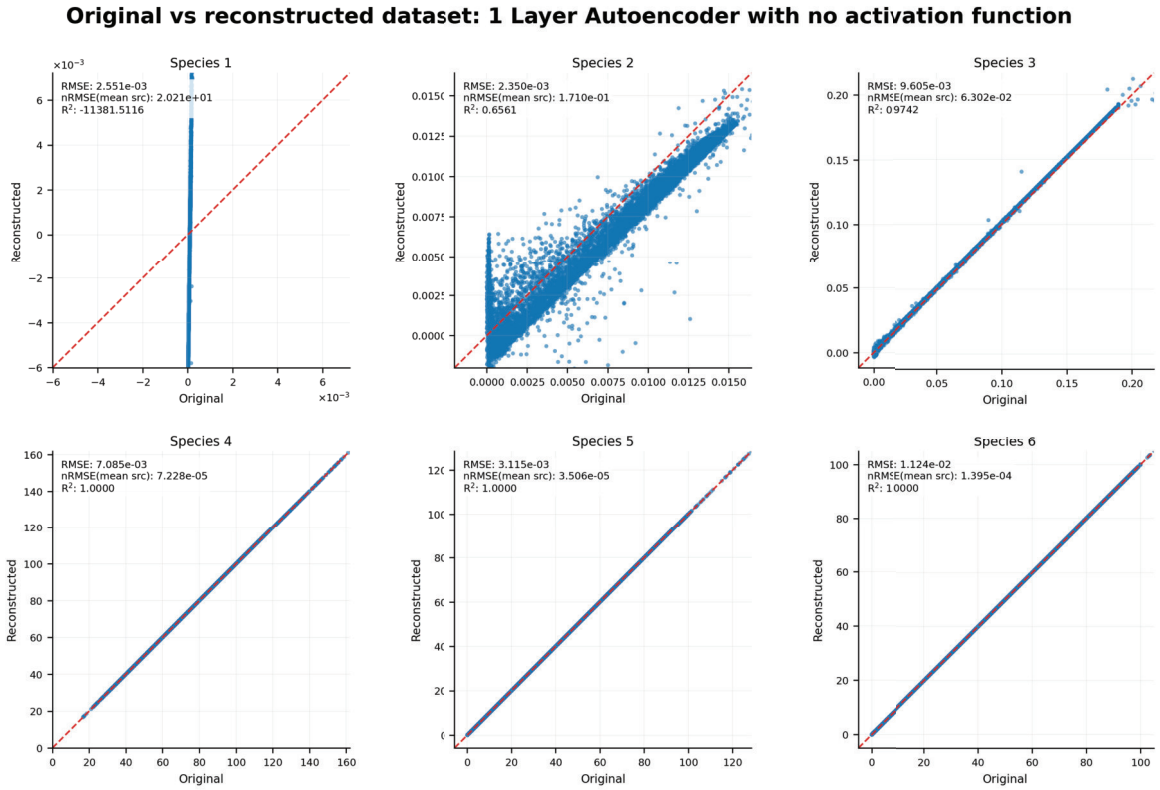

Figure 4.4.: Scatter plot of the original and reconstructed state space using a 1-Layer Autoencoder with no activation function. The red dashed line denotes perfect 1:1 agreement. RMSE, nRMSE (normalised by source mean), and R^2 are shown per panel. Reconstruction is skilful for Species 3-6, while Species 1 exhibit catastrophic failure.

to the direction of maximum residual variance irrespective of the absolute magnitude. In this projection, only the 2 least important components are discarded, retaining a good representation of all species.

The Autoencoder with no activation function, in contrast, optimises with gradient descent on the raw MSELoss. This systematically underestimates the importance of lower concentration values as the RMSE is not scaled correspondingly. The gradient signal of low absolute magnitude species, e.g. Species 1-2 (Figure 4.4 a,b), has a lower order of magnitude than the noise of the signal in other species, effectively silencing these species.

The UMAP projection (Figure 4.5) supports this claim, the centre branch of Species 3-6 (c-f) as well represented while the peripheral states are poorly represented. Species 1-2 (a-b) are a complete failure, where neither the centre branch nor the peripheral states are well represented.

In conclusion, the single-layer linear Autoencoder demonstrates that the theoretical equivalence between linear Autoencoders and PCA [60] does not translate to practical equivalence when optimised via gradient descent on heterogeneous, multi-scale data. The gradient dominance of high-magnitude species, the non-negativity violation, and the zero-variance

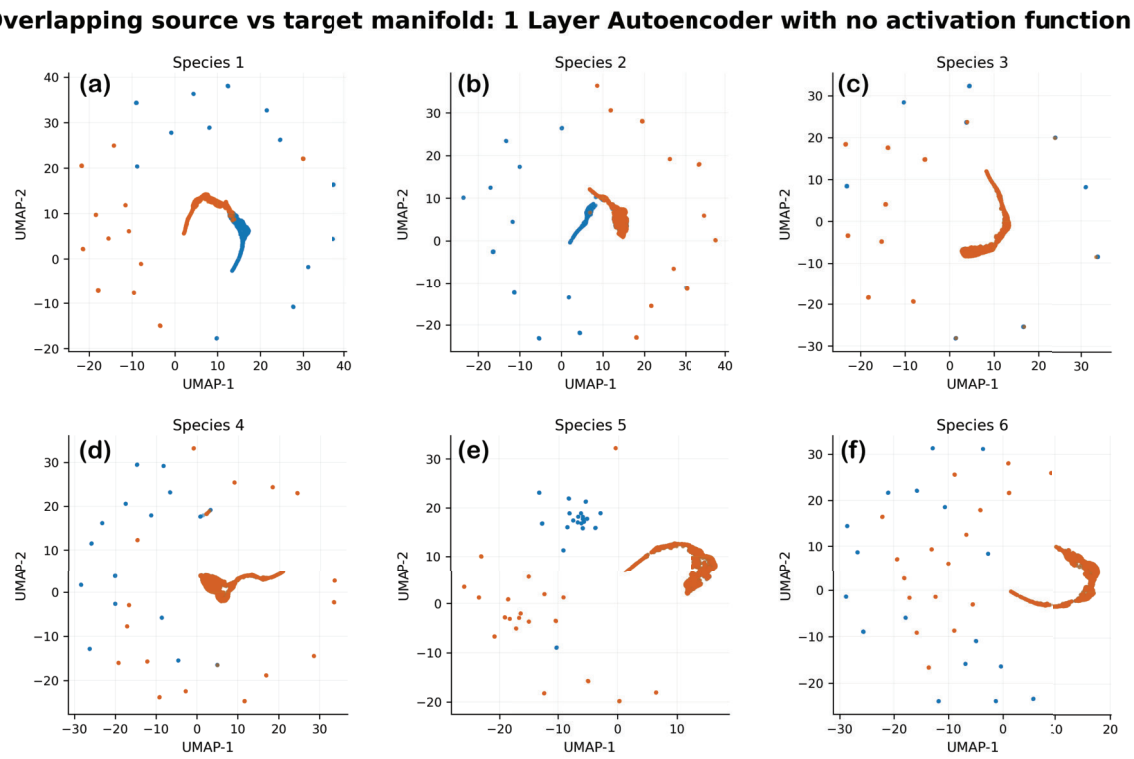

Figure 4.5.: Stacked UMAP projection of the source and reconstructed state space using a single-layer Autoencoder with no activation function. Each panel shows the two-dimensional UMAP embedding for one species: (a-c) mass concentration and (d-f) number concentration of Dust A, B and C respectively. The source manifold (blue) and Autoencoder reconstruction (orange) are overlaid. Species 1–2 (a-b) exhibit catastrophic misalignment, confirming that the gradient-dominated optimisation failed to learn any meaningful structure for low-magnitude species, while Species 3–6 (c–f) show reasonable overlap in the central branch.

problem observed in the NMF (Section 4.2) are all present and amplified. This motivates the introduction of non-linear activation functions and the multi-head architecture discussed in the following section, which address these limitations by introducing non-negativity constraints, a binary classifier for the zero/non-zero distinction, and a weighted loss function that rebalances the gradient contributions across species.

4.4. Non-linear transformation

With the limitations of the previous model in mind, we move our attention to a non-conventional multi-head loss function enabled Autoencoder. The detailed discussion of the network architecture, reasoning and custom loss function can be found in Section 2.3.3. This leads to the proposal of the subsequent multi-head Autoencoder. An ablation study is performed to assess the numerical performance of the raw magnitude head without the binary confinement of the non-zero head. The results show that the raw magnitude head is

architecturally problematic and leads to a systematic failure of the latent and reconstructed representation, resulting in the vanishing gradient problem.

4.4.1. Multi-head Autoencoder

The Autoencoder achieved convergence (Figure D.1) with regularisation and Binary constraint from *WeightedMSELoss* and *BinaryCrossEntropyLossWithLogits*. The hyperparameter for these functions, and the final *MultiHeadLoss* can be found in Table A.1. The main constraint is applied to the *WeightedMSELoss*, where the non-zero elements in the source manifold are weighted 10x in the MSELoss calculation, this allows the Dense layer in the magnitude head to filter the dataset. This allows a much higher quality result without the influence of the undesired noise, skewing the distribution to outputting mostly/wholly 0 data.

The construction of the final, *MultiHeadLoss* involves a simple addition, as discussed in Section 2.3.4. Here the *a3* is chosen as 0.8 from hyper parameter tuning, indicating that strong constraint from the binary head is necessary to counteract the strong weighting from the non-zero value *WeightedMSELoss*.

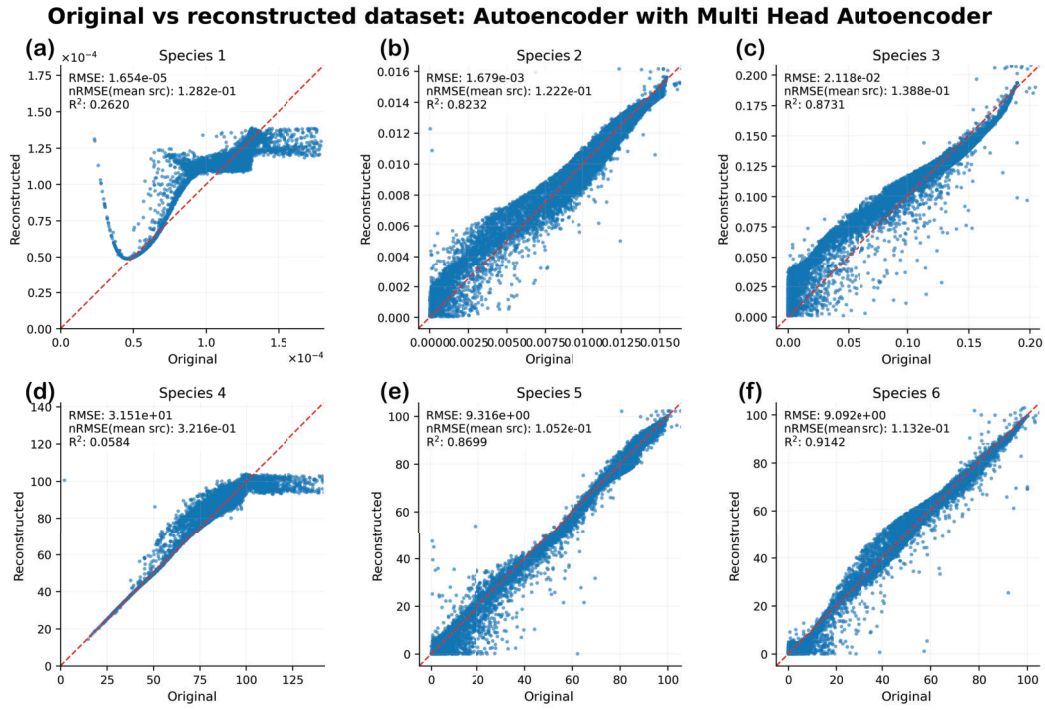

Figure 4.6.: Scatter plot of the original and reconstructed state space using the Autoencoder with Multi Head classification. The red dashed line denotes perfect 1:1 agreement. RMSE, nRMSE (normalised by source mean), and R^2 are shown per panel. Reconstruction is skilful for Species 2, 3, 5, and 6 (R^2 = 0.82–0.91, nRMSE ~ 0.10–0.14), while Species 1 and 4 are poorly captured (R^2 = 0.26 and 0.06, respectively), the latter exhibiting clear saturation at high concentrations.

Figure 4.6 provided a summary of the reconstruction quality of the trained model on the test dataset. The result indicates a good construction quality, particularly in both mass and number concentration of dust b & c (4.6b-c, e-f). The centroid of these species lies very close to the slope-1 (red dotted) line, signifying the model's inclination of performing better in species with a higher order of magnitude, which is no independent phenomenon. This phenomenon is reported in previous sections, particularly in the Non-negative Matrix Factorisation. The author theorises that the Neural Network tends to be optimised to the species with the higher order of magnitude in each side of the distribution, as it would have a high impact on the *MSELoss*, and negative values, or simply 0 values are observed in mass concentration dust a & b in that model.

Dust A shows a different story to the other species, it systematically underestimates the high value domain of the distribution, while the medium values shows a systematic overestimation, stronger that the other 2 distributions. It is worth noting that, mass concentration dust a, the distribution with the lowest order of magnitude in all distribution, exhibits a strong overestimation in the tail of the distribution. The author theorises that this might be an interference between the BCE and MSE Loss.

This indicates that the multi-head approach has successfully achieved a robust reconstruction of the target dataset, with some exceptions in the extreme low values of the distribution. This indicates that the model, as a proof-of-concept for further adaptation in atmospheric chemistry and operational forecasts, can already be safely extended to a wider range of aerosol/chemical species.

As the author discussed in Section 3, this experiment is limited to 6 species, which might not be entirely ideal for a machine learning model; a small channel dataset might very easily lead to overfitting in a small model. The multi-head approach heavily over-parametrises the dataset, and has potentially allowed the model to reveal the hidden structure in the source manifold. Therefore, the parameter space (depth/width of the network) should scale with the number of species, e.g. in atmospheric chemistry simulations.

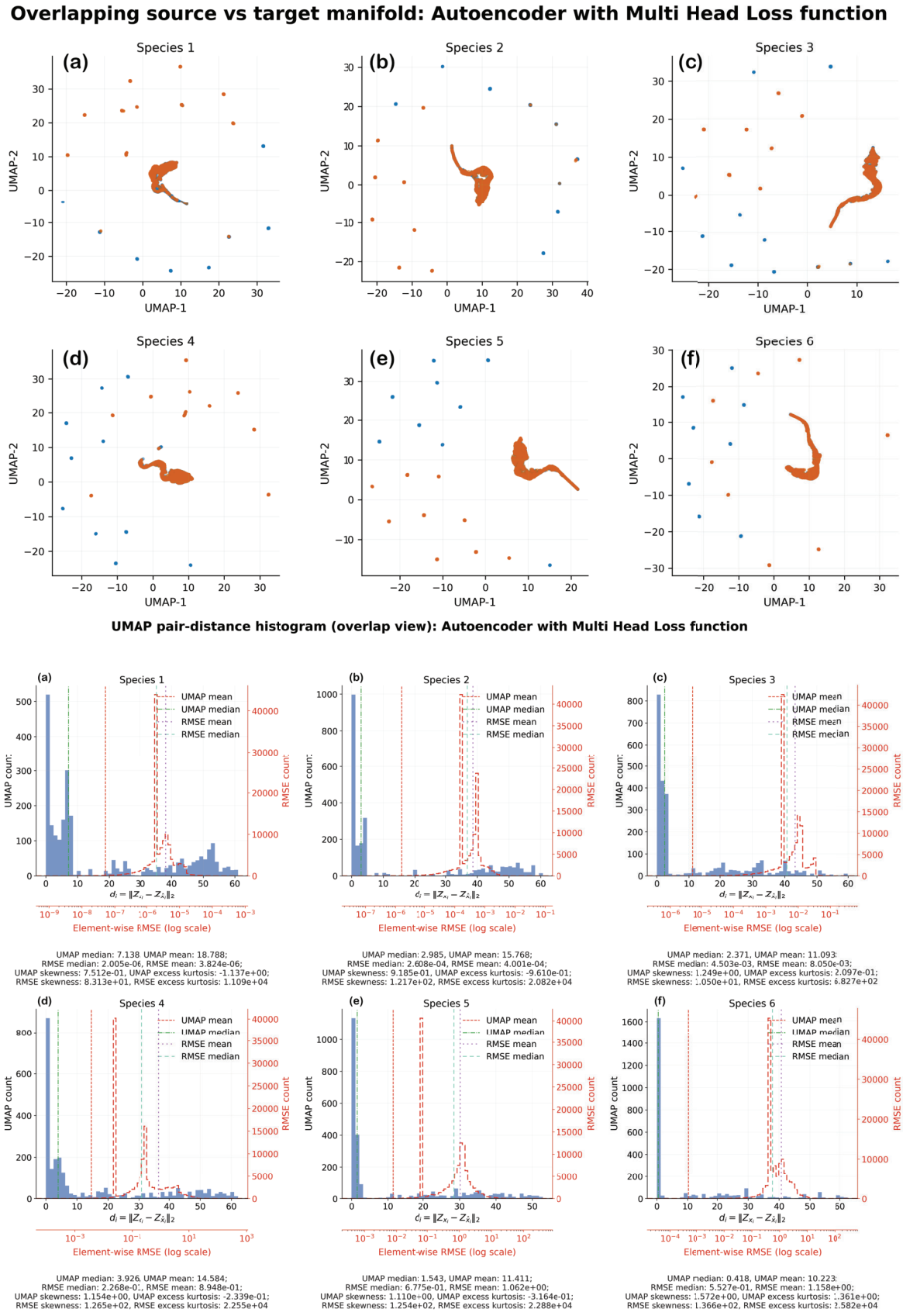

Figure 4.7.: Manifold reconstruction quality of the multi-head Autoencoder. (top) Stacked UMAP projection of the source (blue) and reconstructed (orange) state space for each species: (a-c) mass concentration and (d-f) number concentration of Dust A, B and C respectively. (bottom) Composite histogram of UMAP distance (left y-axis) and element-wise RMSE (right y-axis, log-scaled) between source and reconstructed manifolds for the same species.

4.4.2. Ablation study: Softplus experiment

An ablation study is introduced to assess the raw performance of the Softplus magnitude head without the binary constraint introduced by the non-zero head. All model structures are kept constant except for the removal of the non-zero head. Figure 4.8 shows the scatter plot of individual species of the reconstructed manifold and ground truth.

The baseline Softplus head exhibited a catastrophic failure mode across all 6 species. Scatter plots of the reconstructed dataset and ground truth revealed that the decoder output collapses to a near-constant value for each species, irrespective of the input dynamic range. The log-scaled visualisation (Figure 4.8 g-l) confirmed that the constant output is universal for all species: the output occupies a narrow band within $[0.3, 1.3]$, while the original data spans 6 orders of magnitude. The apparent cluster of mass concentration of Dust A-C (Species 1-3) is only a manifestation of the fact that these species occupy $O(-4)$-$O(0)$, and are better represented in log space while appearing as a "cluster" in linear space. In both cases, the decoder mapped the entire input manifold to a narrow range, which can be understood by a conceptual analysis of the output interval of the 2-layer Autoencoder.

The Tanh activation has an output interval of $(-1, 1)$ and the subsequent mapping $log(1+e^x)$ (Equation 2.22) maps the input to $(log(1+e^{-1}), log(1+e^1)) \mapsto (0.31, 1.31)$, exactly matching the range of data observed. The decoder is therefore structurally incapable of representing values outside of this interval.

Introducing affine parameter can potentially expand the band at the expense of the accuracy of the solution. As the output interval is artificially mapped from $(0.31, 1, 31) \mapsto (0, \infty)$, a minor error at each layer cascade exponentially by the same mapping, resulting in large error, particularly in negative and large positive orders of magnitude output domain. The choice of W will the monotonically determine the output range of the tuned Autoencoder, meaning that the model will be confined to a predefined interval, defeating the purpose of the 2^{nd} mapping to the interval $(0, \infty)$

Moreover, the gradient precision becomes a fatal issue when the output range is expanded. Let W denotes the affine parameter, then the equation of the decoder thus become $\widehat{V} = W \cdot \text{softplus}(\ tanh(h)) + r$, The derivative of the forward pass of the decoder thus become:

$$\frac{\partial \widehat{V}}{\partial h} = W \cdot \sigma(tanh(h)) \cdot (1 - tanh^2(h)) \tag{4.4}$$

As GPUs are optimised for float 32 [83], and that float 64 will result in a significant performance impact on the model, the following discussion will proceed with float 32 in mind.

For float 32, the numerical precision is $2^{-23} = 1.19 * 10^{-7}$, meaning that any gradient smaller than this value will become indistinguishable. As $tanh$ is monotonically increasing, the gradient $d\,tanh(h)/dh = sech^2(h)$, at $d\,tanh(h)/dh = 1.19 * 10^{-7} = sech^2(h)$, the inverse of $sech^2(h)$ is $\ln(\frac{1+\sqrt{1-y}}{\sqrt{y}}) = \ln(\frac{1+\sqrt{1-1.19*10^{-7}}}{\sqrt{1.19*10^{-7}}}) = 8.665$, beyond which any update to the gradient will be lost due to rounding. Here the network hit the vanishing gradient problem,

and therefore the input h must be confined within a small value to maintain non zero gradient. This will further damage the credibility in h, which is the latent representation, or fundamentally the encoder. The channel capacity of this network is architecturally confined and the saturated domain of Tanh and Softplus will confine the available interval for normal output substantially, resulting in the problem observed in Figure 4.8.

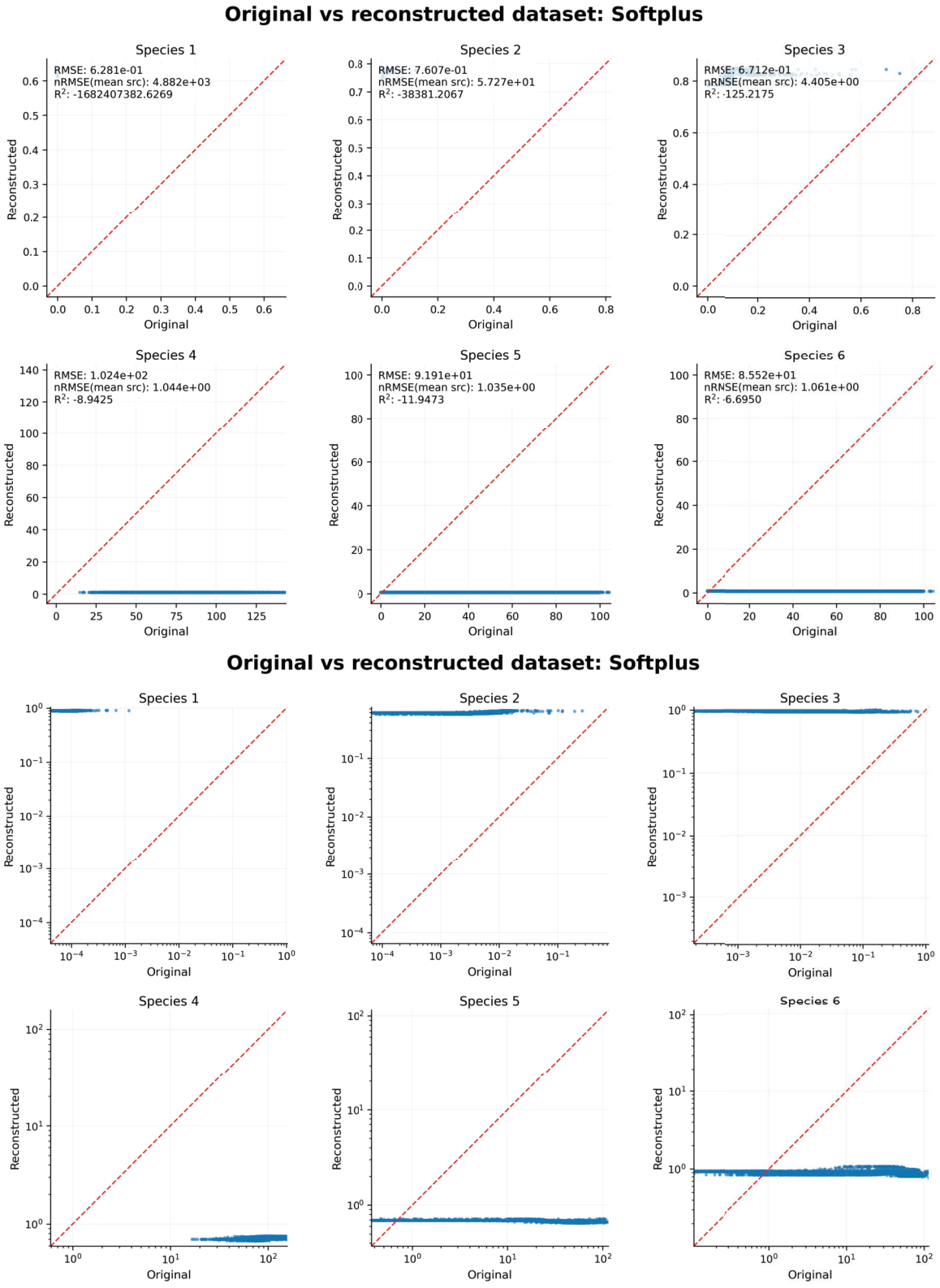

Figure 4.8.: Scatter plot of the Softplus Autoencoder reconstruction and ground truth. (top, a-f) Linear-scaled plot and (bottom, g-l) log-scaled plot for each species. The log-scaled representation reveals that the decoder output collapses to a narrow band within $[0.3, 1.3]$, the composite output interval of $\log(1 + e^{\tanh(\cdot)})$, confirming that the architecture is structurally incapable of representing the full dynamic range of the tracer field.

The proposed multi-head architecture has several benefits, including the ability to approximate true 0 value by the BCE classifier without saturating the gradient.

A custom loss function of *weightedMSELoss* (Equation 2.26) and *BCEwithLogitLoss* (Equation 2.36) was introduced to improve the vanishing gradient problem, and enhance the model's ability to classify 0 and non-zero element.

The multi-head architecture allows a much wider dynamic range of the output. The Softplus activation function has 2 domains, a linear domain where $e^x \gg 1$ such that the log is dominated by the exponential of x, resulting in a linear like behaviour, while the non linear domain has small or negative x, returning a log like behaviour, quickly gliding to 0. The linear domain is the desirable domain for this application, and the introduction of the sigmoid head resulted in a "scaling factor" which can produce values close to zero. This extends the dynamic range of the output $\widehat{V} = m(h) \cdot \sigma(h)$ as arbitrarily large and small values are represented in different heads.

The signal to noise ratio (non-zero/zero magnitude) of the gradient of the model is improved. The gradient of model through the loss function i.r.t. the dense layer before the *tanh* activation can be obtained by chain rule:

$$\frac{\partial \mathcal{L}_{\text{multi-head}}}{\partial h_{tanh}} = \text{sech}^2(h_{tanh}) \cdot (G_{\text{MSELoss, magnitude head}} + G_{\text{MSELoss, non-zero head}} + \alpha_3 G_{BCE}) \quad (4.5)$$

The formulation is consistent to Equation 4.4, replacing the Softplus gradient with combined gradient of the 3 trajectory. The trajectories each have a gradient

$$\begin{cases} \text{Pathway 1} : G_{\text{MSELoss, non-zero head}} = -2w_{MSELoss}(x - pm) \cdot m \cdot p(1-p) \cdot \theta \\ \text{Pathway 2} : G_{\text{MSELoss, magnitude head}} = -2w_{MSELoss}(x - pm) \cdot p \cdot \sigma(a) \cdot \theta \\ \text{Pathway 3} : G_{\text{BCE}} = \alpha_3(p - y) \cdot \theta \end{cases} \quad (4.6)$$

Note that the p is the prediction of the non-zero head, $\sigma(a)$ is the gradient of the Softplus and x & y being the ground truth of the magnitude and the non-zero binary label (*support_true* variable). The 2 extrema

Zero element ($x = 0, y = 0, w_{MSELoss} = 1$)

$$\begin{cases} G_{\text{MSELoss, non-zero head}} = -2(0 - pm) \cdot m \cdot p(1-p) \cdot \theta = 2p^2(1-p)m^2\theta \\ G_{\text{MSELoss, magnitude head}} = -2(0 - pm) \cdot p \cdot \sigma(a) \cdot \theta = 2p^2 m\sigma(a)\theta \\ G_{\text{BCE}} = \alpha_3 p \cdot \theta \end{cases} \quad (4.7)$$

as the model approaches to convergence, the $p \to 0$, then all G approaches asymptotically 0 at different rate. The gradient of the BCE (pathway 3) approaches to 0 proportional to p, meaning that at small p, signal from the BCE classifier dominates the gradient calculation.

Non-zero element ($x \gg 0, y = 1, w_{MSELoss} = 10$)

$$\begin{cases} G_{\text{MSELoss, non-zero head}} = -20(x - pm) \cdot m \cdot p(1-p) \cdot \theta \\ G_{\text{MSELoss, magnitude head}} = -20(x - pm) \cdot p \cdot \sigma(a) \cdot \theta \\ G_{\text{BCE}} = \alpha_3(p-1) \cdot \theta \end{cases} \tag{4.8}$$

At larger than 0, the first term approaches to 0, as the probability of the element being non-zero approaches to 1; and the 3^{rd} term also approaches to 0 for the same reason. The remaining head is then the MSELoss that pass through the magnitude head.

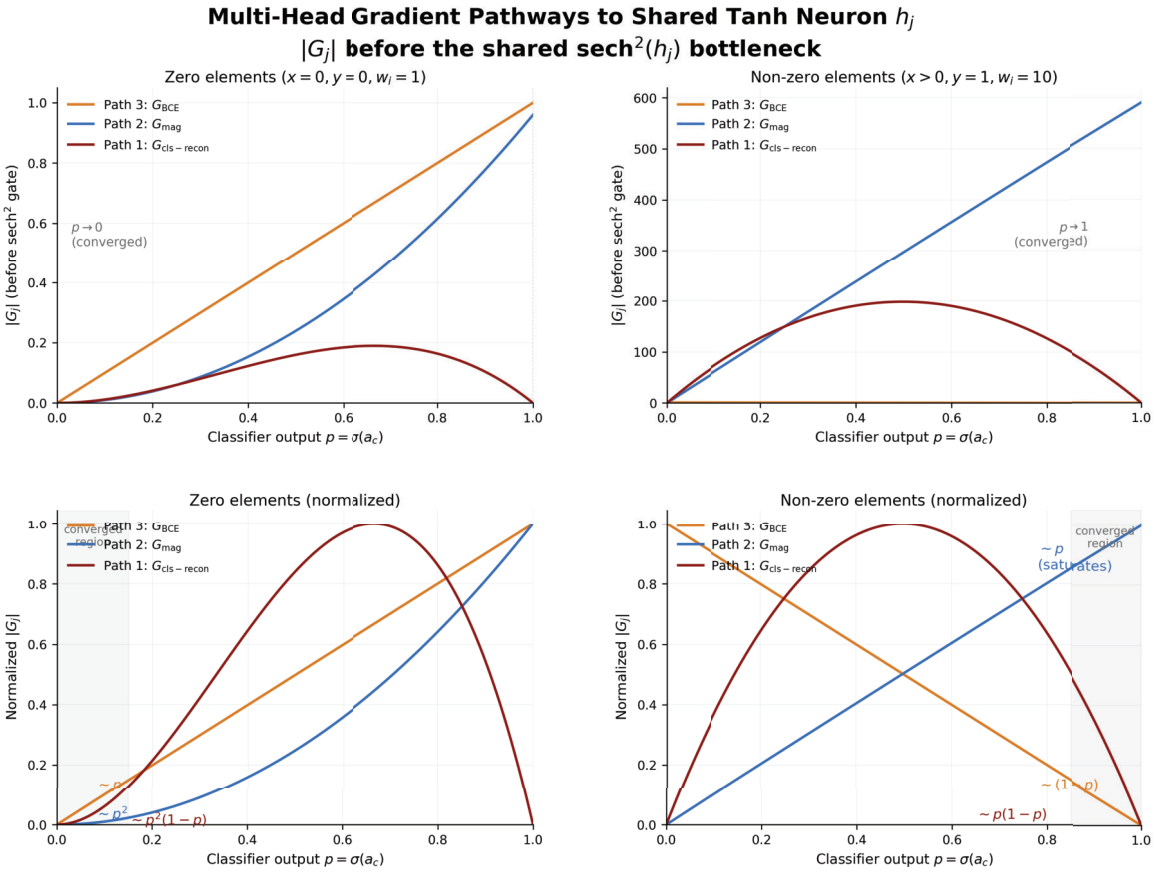

Figure 4.9.: Gradient magnitude of the three backpropagation pathways with respect to the dense layer preceding Tanh. Pathway 1: MSELoss gradient through the non-zero head; Pathway 2: MSELoss gradient through the magnitude head; Pathway 3: BCELoss gradient. The horizontal axis shows the non-zero head output $p \in (0, 1)$. At the zero-element extremum ($p \to 0$), the BCE gradient (Pathway 3) dominates, ensuring classification accuracy; at the non-zero extremum ($p \to 1$), the magnitude head gradient (Pathway 2) dominates, preserving concentration fidelity.

Note that in both cases, the decay rate of the MSELoss through the non-zero head decay at the fastest rate at both extrema, meaning that it only contributes to the gradient descend in the transient domain where model are not close to convergence. Figure 4.9 shows the relationship of the of the output of the non-zero head and the gradient at *tanh*. For zero

element, pathway 1 decays as the classifier output reaches 0 or 1, meaning that this pathway will diminish as the model reaches convergence, consistent to the physical intuition where the reconstruction gradient at the classifier gate will be minimal, as the model is certain on the classifier value, leaving the pathway 2 & 3 competing at these boundary conditions.

At the 2 extrema, it is easy to see that, at zero element, the BCE decays at the slowest rate, meaning that the contribution of the BCE classifier is the strongest, maintaining the accuracy; whilst at the non-zero elements, the gradient of the MSELoss (pathway 3) standout as the classifier output approaches to 1, meaning that the magnitude head's signal dominates. This allows the network to learn different behaviour at different number domain, avoiding the gradient to vanish at both extremes, resulting in a narrow band output as in Figure 4.8.

This has profound meaning for the network's representativeness. The pathway 2 & 3 operates independently, in non-zero elements, pathway 2 reaches up to 600, while pathway 3 peaks only at 1, artificially shutting off the classifier head, allocating the representational capacity of the *tanh* to the magnitude structure of the dust fields. This turns out to be extremely effective in reconstructing the geometry and magnitude of the dust species field. However, it must be stressed that this only represents the ability of the field to reconstruct the input manifold in one forward pass, the stability of the solution in multiple forward pass must be evaluated in a separate model drift study.

4.5. Timing study

This section performs a timing study on the full model, the full model with machine learning model embedding-reconstruction cycle, and the ART-off configuration. By simple addition and subtraction, a proxy for the computational advantage of PyTorch-ComIn coupled embedding can be identified. Note that the machine learning model access the data in *EP_ATM_ADVECTION_BEFORE*, the pointer is fed into the machine learning model via $model(x.T)$ call. The processed field is then fed back the the host model, as a encode-advect-decode cycle will require extensive modification to the advection module, which is beyond the reasonable scope of this thesis.

The net time taken for advection and ComIn callbacks of the 3 advected species plus ComIn callback is 29.82 + 3.5 = 33.32, giving a 21% computational benefit in terms of the transport module. The number is calculated from the time taken for the transport scheme in the host model versus the host model with ComIn callback that modifies the tracer field with the coupled multi-head Autoencoder. Number concentration is not advected in the 3^{rd} setup, advecting only 3 species. This allows an understanding of the saving in computational time for a reduced-order representation.

In this case, the Python-based ComIn callback script did not result in a major bottleneck. This is because $np.asarray()$ was called instead of $np.array()$ as it employs zero-copy by default, such that the pointer remains in memory unmodified until it is passed to the Torch module.

The machine learning model is also not the major bottleneck here. As the model is called locally at each thread with no additional global communication, the Torch backend process for each linear layer is simply a *sgemm*/*cublasSgemm* call, depending on the CPU/GPU process. BLAS or cuBLAS [84, 85] are highly optimised linear algebra libraries, which in theory do not result in higher computational overhead compared to the advection operator, which is written in loops in Fortran.

Additional computational benefit can probably be achieved by using FTorch-ICON coupling (e.g. Kumar *et al.* [86]), C++ ComIn+Torch coupling or hard-coding the Neural Network into an internal routine in the ICON advection module. However, these are unnecessarily complex for a pilot study and do not bring scientific value. This shall be further discussed in Chapter 6.

The processing can also be offloaded to GPU, along with the host model. Lapillonne *et al.* [87] provided version 2024.10 of ICON which can be implemented on GPU. With this in mind, the Torch model can also benefit from the additional computational advantage of CUDA, without additional wait time from RAM-VRAM communication. This can potentially be the single largest contributor of computational overhead if the host model is hosted in CPU memory while the machine learning model is loaded on the GPU. Delays in memory access would lead to additional wait time, resulting in unnecessary delays. Unfortunately, the current setup in the DKRZ cluster has limited GPU resources and cross-sector use of GPU+compute nodes at the same slurm job is not allowed, and the nodes required for the DUST-RAD test suite is more than all available GPU node on levante. Therefore the actual computational overhead cannot be quantitatively assessed. This could of course be attempted in a coarser simulation that fits on one GPU node; however, this goes beyond the scope of this experiment.

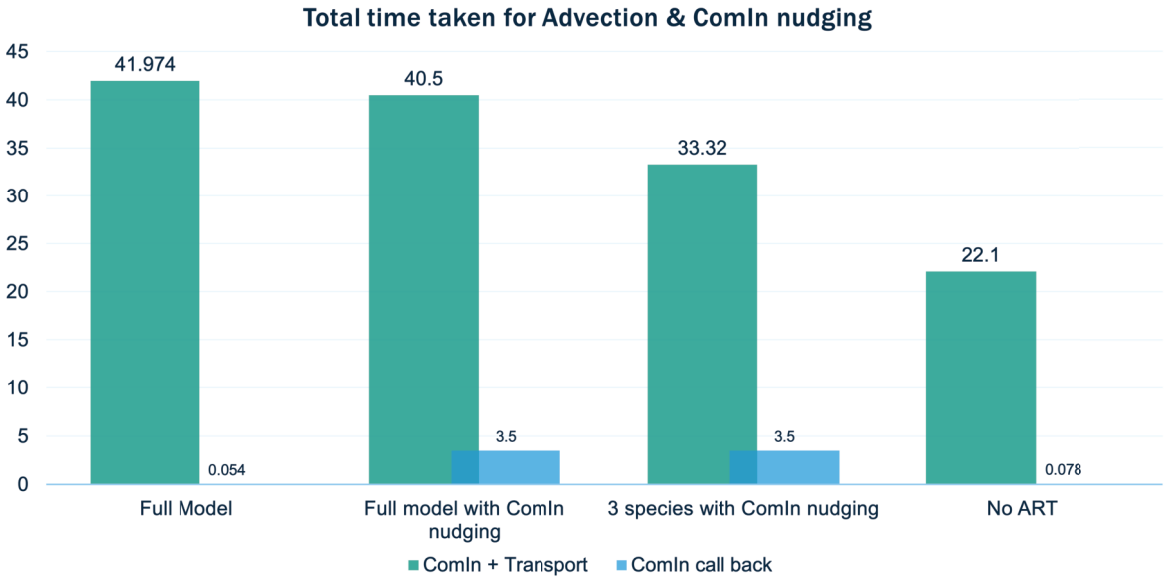

Figure 4.10.: Average time taken (in seconds) for advection and ComIn nudging per MPI thread in a 2-day simulation. Four configurations are compared: the full model (6 advected species), the full model with ComIn-coupled ML nudging, the reduced setup advecting only 3 species with ComIn nudging, and a baseline with no ART species advection. The ComIn + Transport bars represent the combined cost of chemical integration and tracer advection, while the ComIn callback bars isolate the overhead introduced by the ML nudging interface.

5. Model inter comparison

This section summarises and discusses the results from previous sections in an inter-comparative manner. The author identifies the strengths and weaknesses of different models by analysing the absolute magnitude and the manifold distance of the reconstruction. All metrics are recomputed in this section with the testing dataset from the **FULL MODEL**. This examines the generalisability of the models when source and sink processes are factored into the tracer tendencies.

5.1. Spatial error propagation

This section discusses the spatial propagation of error patterns for the Multi-head Autoencoder and Non-negative Matrix Factorisation. The author examines the spatial error distribution at lead times 03h and 144h. Although the data used in this section are drawn from the **FULL MODEL**, evaluating the error at the earliest time steps inevitably reflects the extent to which the model has memorised its training data rather than demonstrating genuine predictive skill; hence the 03h results serve solely as a baseline reference. It is worth noting that the normalisation is performed with respect to the field mean of the original dataset, which may yield slightly different scaling compared to other sections; accordingly, only the spatial error pattern should be interpreted.

Figure 5.1 presents the column mean normalised RMSE of the NMF reconstruction at lead times 03h and 144h. The normalised concentration field is provided in Figure 5.3 as a reference; note that the colour scale is reversed to better represent patches with low normalised mass. In the training data regime (03h, Figure 5.1, top), the column mean normalised RMSE is uniformly low across all species, consistent with expectations: the model reproduces patterns it encountered during training, and the residual reflects memorisation error rather than genuine predictive skill.

The spatial distribution of error at 03h is predominantly oceanic, collocated with regions of lower relative concentration. Figures 5.1 and 5.3 (top) reveals a striking spatial correspondence: regions of lower relative concentration also exhibit higher normalised RMSE. However, the error pattern extends spatially well beyond the low-concentration regions themselves. This behaviour is consistent with the zero-variance problem described above, whereby the model fails to reconstruct lower absolute concentration cells correctly due to its linear scaling, resulting in severe overestimation of the concentration field.

A closer inspection of Figures 5.1 and 5.3 (bottom, 144h) reveals visible patches of elevated nRMSE (dark blue), some collocated with the occlusion band of an Extra-Tropical Cyclone

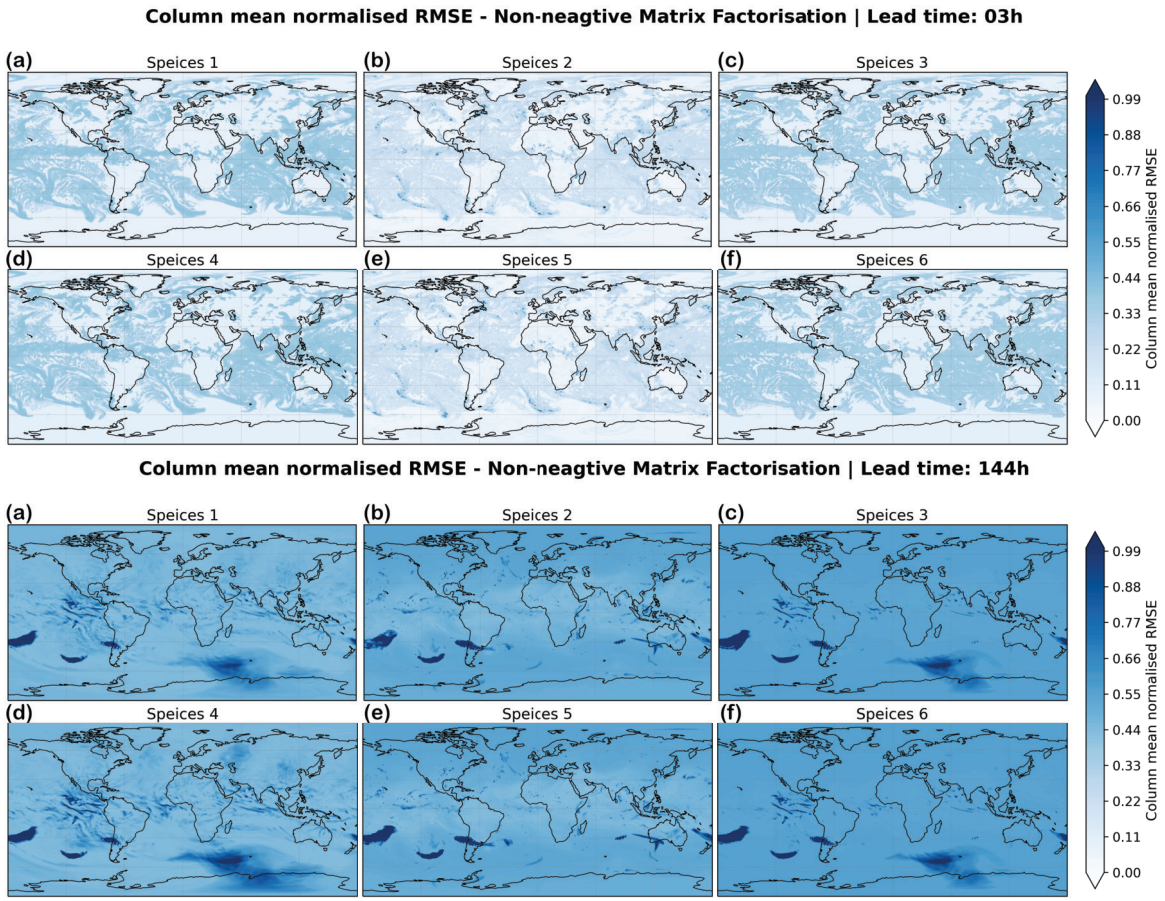

Figure 5.1.: Column mean normalised RMSE of the Non-negative Matrix Factorisation at lead times (top) 03 hr and (bottom) 144 hr. (a–c) mass concentration and (d–f) number concentration of Dust A, B and C respectively. The normalised RMSE is computed column-wise and normalised by the field mean of the original dataset. At 03 hr the error is uniformly low across all species, reflecting memorisation of the training distribution. At 144 hr, localised patches of elevated nRMSE emerge over oceanic and high-latitude regions where relative concentrations are suppressed.

(ETC) in the Southern Pacific and high-latitude disturbances over the Antarctic. However, this pattern is not replicated for other ETC occurrences in different ocean basins, rendering the observation inconclusive and insufficient to establish a robust synoptic link to the spatial evolution of the error.

The error pattern of the Multi-head Autoencoder bears a remarkable resemblance to that of the NMF, with several noteworthy distinctions. In the training data regime (Figure 5.2, top), the error pattern is broadly consistent with the NMF: uniformly low nRMSE across the domain, with noticeably higher error over oceanic regions where relative concentrations are suppressed by source-sink processes (convection, deposition, wash-out, etc.). However, the spatial structure is smoother for Dust A (Species 1, 4) and Dust B (Species 2, 5), likely reflecting the more continuous interpolation afforded by the non-linear Autoencoder across the lower-concentration sub-manifold of the data.

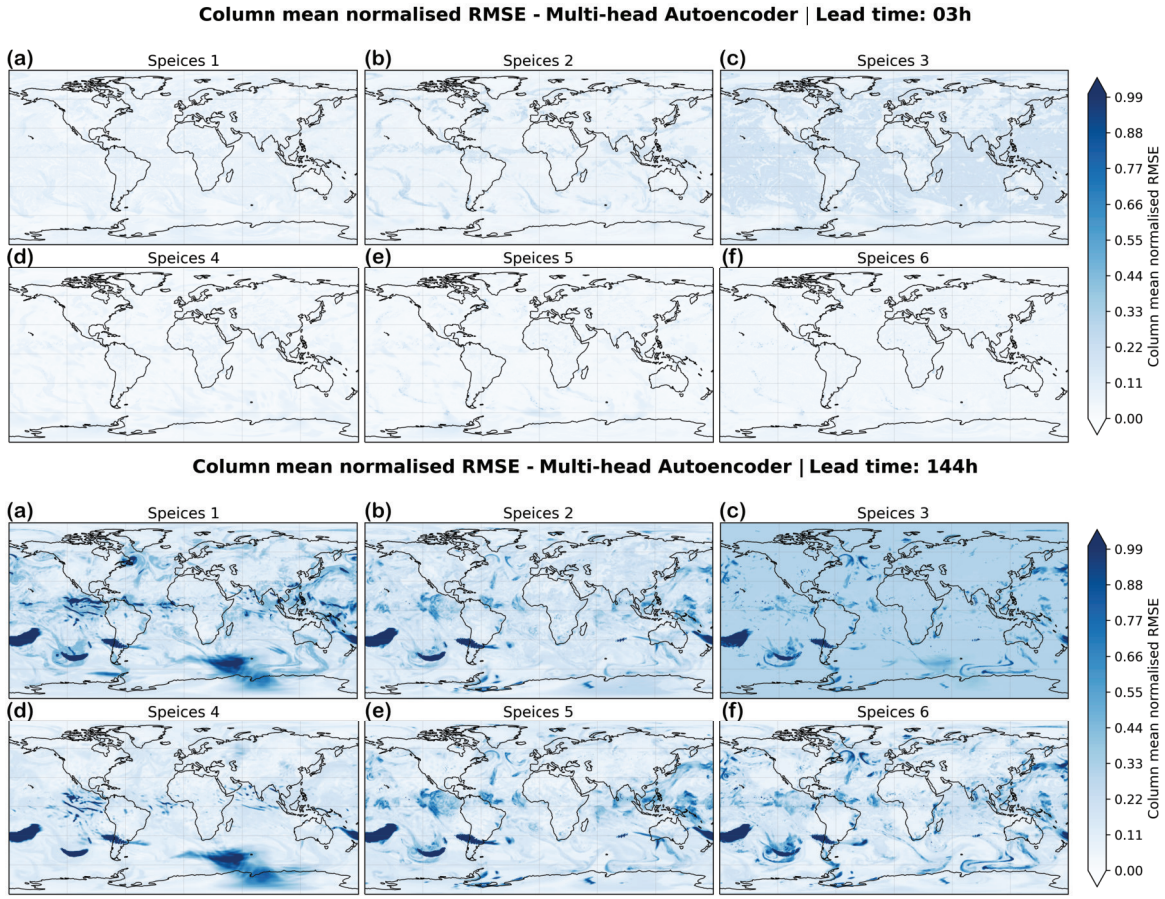

Figure 5.2.: Column mean normalised RMSE of the Multi-head Autoencoder at lead times (top) 03 hr and (bottom) 144 hr. (a–c) mass concentration and (d–f) number concentration of Dust A, B and C respectively. The normalised RMSE is computed column-wise and normalised by the field mean of the original dataset. At 03 hr the error is uniformly low across all species, comparable to the NMF baseline. At 144 hr, sharply localised patches of elevated nRMSE emerge, collocated with synoptically active features such as extra-tropical cyclones and high-latitude disturbances where source-sink processes deform the tracer sub-manifold.

At 144h (Figure 5.2, bottom), the Multi-head Autoencoder diverges from the NMF. The darkest patches, denoting very high nRMSE, are sharply collocated with two synoptically identifiable features: an ETC complex in the Southern Pacific and the high-latitude disturbances described above. Unlike the NMF error pattern, where only the ETC complex and Antarctic disturbances are identifiable, additional error maxima in the Multi-head Autoencoder are consistently observed along the occlusion branch and within convection active regions.

Comparison with the normalised concentration field (Figure 5.3, bottom) reveals a collocation of high-error regimes with regions of weak column mean normalised concentration—precisely the zones characterised by strong lifting, cloud interaction, and scavenging. These processes act on the three size distributions (Dust A, B, C) at different rates due to aerosol microphysics: the mass and number concentration sub-manifolds within a given size bin

are deformed similarly, whereas the sub-manifolds across the three size distributions are deformed at different rates. In these regions, the sub-manifold geometry therefore differs substantially from the general distribution, where source and sink processes are less vigorous. This tight spatial correspondence suggests that the error is not an artefact of the model construction, but is driven by the lack of representativeness of source-sink-affected sub-manifolds in the training data, as the author disabled all source and sink processes in pursuit of isolating the pure advection tendency. This demonstrates that such strategy may result in degraded reconstruction fidelity in convection active regions. Therefore, retraining with a mixed full model – nss data set or a full model only dataset can be used to train the data for better generalisability.

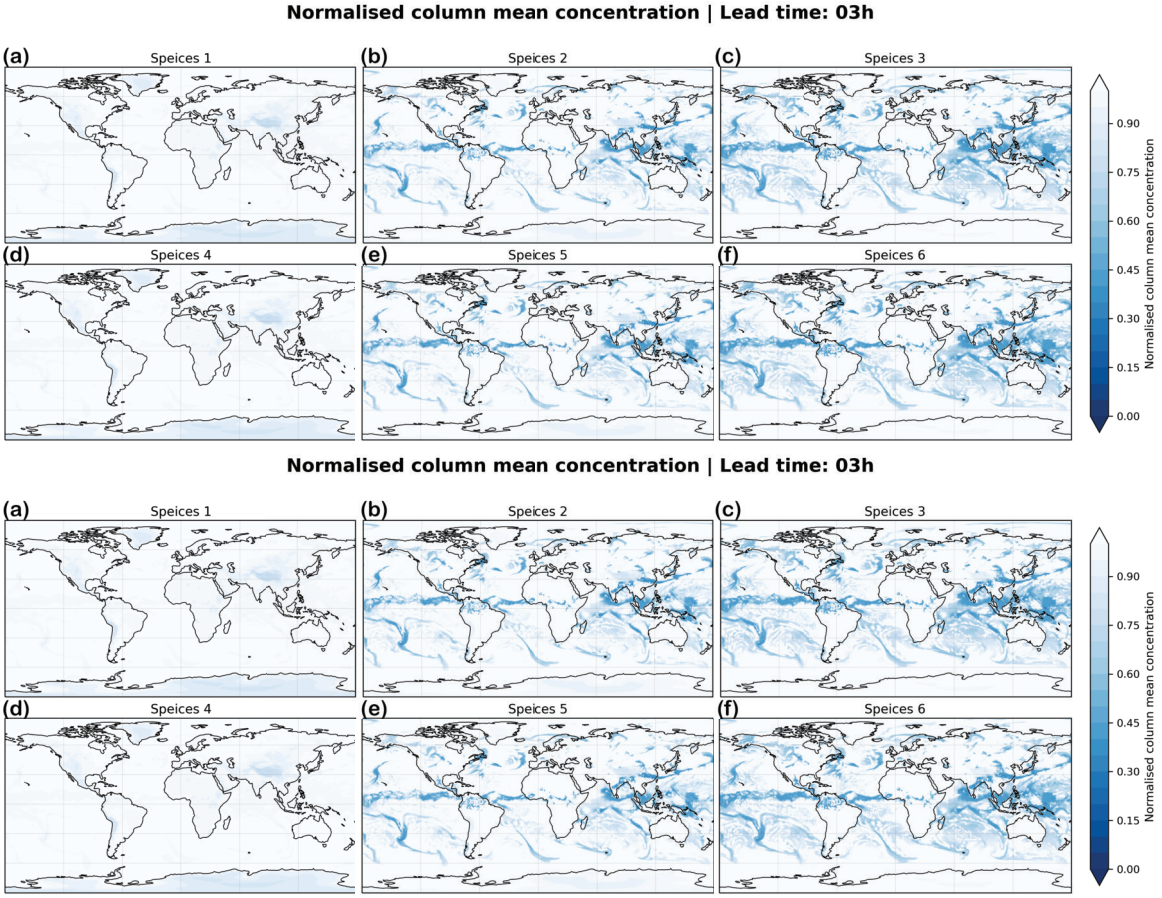

Figure 5.3.: Normalised column mean concentration of the full-model reference simulation at lead times (top) 03 hr and (bottom) 144 hr. (a–c) Mass concentration and (d–f) number concentration of Dust A, B and C respectively. Each grid cell is normalised by the global field maximum of the corresponding species so that values range from 0 to 1; the colour scale is reversed relative to the error maps in Figures 5.1 and 5.2 to highlight regions of suppressed concentration. This figure serves as a spatial reference for interpreting the collocation of elevated nRMSE with low-concentration regimes.

5.2. Geometric preservation

The NMF model performed relatively worse compared to the Multi-head Autoencoder in the relative Wasserstein metric. To compare across the different species and models, Wasserstein metric are scaled with the Interquartile Range of the source manifold [88], returning a $[0, +\infty)$ metric with lower being a better construction of the CDF. mass & number concentration dust A (Species 1+4) performed equally bad in both NMF and multi-head Autoencoder, suggesting a deeper, geometry factor in play, then a simple training/scaling issue. This is because, although the author have theorised that lower order of magnitude data will contains higher error, and it is indeed the case in Section 4.2 & 2.3.3, the scatter plot (Figure 4.6 a & 4.1 a) shows the highest deviation from the slope-1 line. This partly answers the phenomenon observed in 5.1, while the low Wasserstein metric in number concentration Dust A remains a puzzle. This may be caused by the underlying geometric similarity between the mass and number concentration of dust A, meaning that both model are actually being able to exploit the hidden connection between the tracer species, denoting a "success" in our machine learning development despite the failure in the absolute metrics.

5.3. Magnitude preservation

In previous sections, individual model performance, the mathematical foundation of the failure modes and manifold reconstruction were analysed on a single forward pass. This section extends the analysis by introducing an offline drift experiment. The analysis is performed by applying the embedding-reconstruction cycle for 100 iterations, using the testing dataset from the **FULL MODEL** as a seed to examine the drift in reconstruction quality over long iterations. This exposes the numerical stability of the solution beyond a single forward pass, where some models may indeed perform very well in one forward pass yet fail dramatically in repeated iterations. This acts as a proxy for the repeated nudging that occurs in online, coupled experiments. It must be stressed that this implementation is not the true online coupling, as

Normalised RMSE and the variance ratio $\sigma(\widehat{x})/\sigma(x)$ will be examined, which gives a standardised measurement of bias and distributional change (such as widening, suppression or preservation of the spread) respectively. The 2 metrics together distinguish between distributional shift (high nRMSE with near unity σ ratio) and a distortion of the distribution (high nRMSE & high anomaly in σ ratio).

In PCA, the nRMSE is identical for iteration 0 and 100 for all species. This is not surprising, as from Equation 4.1, $V_k V_k^T$ is an idempotent operator which introduces no additional error by construction, and the repeated projection can be simplified as $V_k(V_k^T V_k)\ldots(V_k^T V_k)V_k^T X \rightarrow (V_k^T V_k) = I_k$, such that I_k is an identity matrix. The only error is rounding due to numerical precision. PCA therefore serves as a theoretical lower bound on iterative error, albeit with no non-negativity guarantee and mean dependence on the training set.

For Non-negative Matrix Factorisation, the total nRMSE is mediocre at ~11.0% at iteration 0, drifting towards 15.5% after 100 iterations. The relative growth is about 41%, with the error growth scaling linearly with the number of iterations. The drift itself is unevenly distributed, with mass & number concentration of dust A (Species 1,4) & dust C (Species 3 & 6) remain stable. The nRMSE of these species lies between 10.1 – 19.9% at iteration 0, rising to 12.9 – 26.4% at iteration 100. The mass and number concentration dust B (Species 2, 5) are the most problematic. They start from 104.9 & 141.9% respectively, and worsening up to 135.9% and 181.3% for mass and number concentration respectively.

The 2 species also exhibit the largest variance degradation across all species. At iteration 0, the σ ratio of mass concentration B (Species 2) starts at 59.9%, reducing to 54.3%, while the number concentration starts from 97.3%, falling to 50.0% after 100 iterations. The other species remains relatively well represented, with the iteration 0 and 100 σ ratio at or above 89%. The total variance ratio degrades from 96.4% to 92.5%, meaning that the NMF preserved magnitude reasonably well over long iterations, except mass and number concentration dust B.

The Autoencoder with Softplus continues to exhibit catastrophic failure. Mass concentration shows a nRMSE of 4900x, 49x, 8.45x respectively for dust A-C (Species 1-3), the error only increased marginally after 100 iterations, indicating that the model are predicting within the narrow band of values discussed in Section 4.4.2. For σ ratio, All number concentration exhibited a near 0 variance ratio, suggesting that all source concentration are suppressed to a narrow band of values, consistent to the observation in Section 4.4.2 and Figure 4.8. This will therefore be excluded for further consideration as a candidate model.

The Multi-head Autoencoder achieves good initial reconstruction, while slightly worse than the no-source-sink dataset in Section 4.4.1. The total nRMSE starts at 12.6%, and the σ ratio are predominantly within the ideal range (93.2-98.7%), with number concentration dust A failed exceptionally at 69.2%.

After 100 iterations, however, the model drift becomes substantial, where species specific failure mode becomes apparat. The total nRMSE and σ ratio degrades to 27.9% & 67.4% respectively. The drift is most severe in dust C, nRMSE of mass concentration (Species 3) increased substantially from 12.9% to 63.9%, with the σ reduces from 93.2% to 38.1% correspondingly; for number concentration, the nRMSE & σ deteriorates from 10.5%, 98.5% to 82.1% & 67.4%. This collinear reduction in prediction skill and dynamic range suggest that the model losses its stability over long iterations.

It is particularly noticeable that the initial reconstruction geometry was badly preserved for number concentration dust A (Species 4). The initial forecast error us 17.4% and the variance ratio is catastrophic at 69.2%, indicating that the reconstruction resulted in a smoothened, amplitude dampened, consistent to the observation in Figure 4.6 d, where the explain variance is the worse over all species.

Overall, PCA results in near-perfect reconstruction but is inadmissible as a transferable model to different scenarios. NMF shows bounded error growth with good variance preservation for the ma'jority of species. The multi-head Autoencoder matches NMF at iteration 0 but degrades significantly over 100 iterations in dust species C. The contrast between the

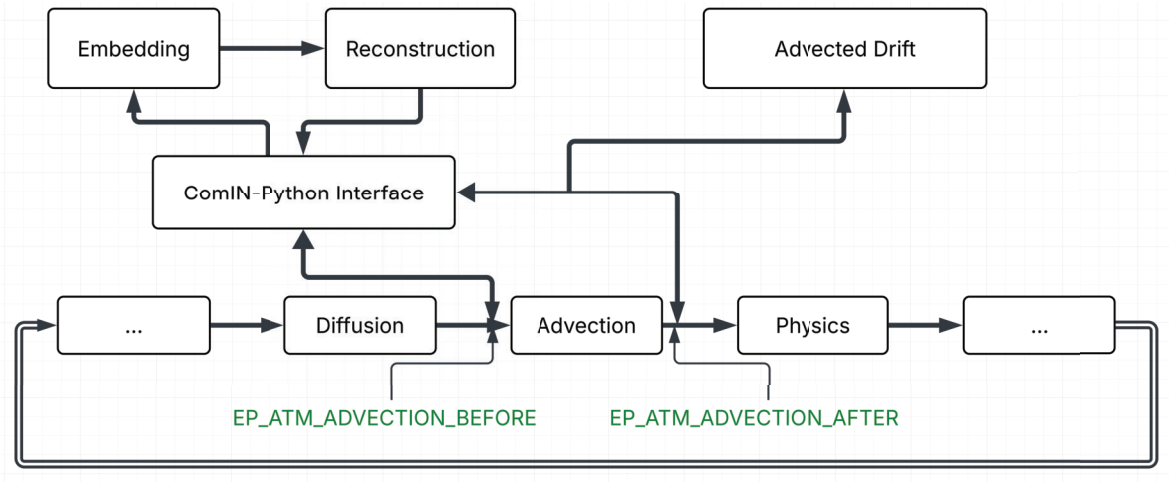

Figure 5.4.: Schematic diagram of the online coupling module for the proxy drift experiment. The machine learning model receives the tracer field at entry point *EP_ATM_ADVECTION_BEFORE*, applies the encode-decode cycle, and writes the reconstructed field back to the host model. After advection by the ICON transport scheme, the advected data are extracted at *EP_ATM_ADVECTION_AFTER* for evaluation. This proxy approach nudges the tracer field at every time step without modifying the advection operator itself.

models—NMF preserving the variance ratio well (92.5%) while the multi-head Autoencoder is crippled by one particular species—raises the question of whether the error extends to mass conservation, our main physical constraint.

5.4. Mass conservation

The previous section showed that the multi-head Autoencoder suffers from species-dependent variance collapse under iteration while NMF remains near-idempotent in cell-wise accuracy. This section examines whether these errors translate into a net mass budget imbalance. The author compares and interprets the results from the offline drift experiment, then introduces a simple online drift experiment with Torch-ComIn-ICON coupling as described in Section 4.5, where the data is nudged by the model before advection (Figure 5.4). This acts as a proxy study for the interaction of the processed tracer field with other components of the model (e.g. source/sink, physics, flux limiter).

Offline drift experiment

In PCA, the repeated embedding-reconstruction cycle yields no model drift and remains 100 % mass conserved from iteration 0 to 100. This is not surprising due to the consistently good performance from previous metrics.

NMF offline drift

Non-negative Matrix Factorisation shows minimal model drift in offline (Table 5.3), uncoupled experiment. All species preserves good mass conservation with error range (-2.16, 9.58%), with the largest bias occur at the number concentration of dust B & C, showing a positive and negative bias respectively. After the model is drifted for 100 iterations, the mass loss did not scale linearly with the number of iterations. The total mass ratio have decreased from 98.32% to 94.74%, much lower reduction than source→iteration 0. The largest drift remains in number concentration of dust B & C (7.98 %, 5.33 % respectively), with uneven mass losses observed across all species. This means that the model remained stable under repeated iteration, showing near idempotency behaviour.

For non-linear models, the results are catastrophic for Autoencoder with 2 layer Softplus and very uneven for Multi-head Autoencoder.

For Autoencoder with Softplus, mass concentration species have up to 500000% bias at iteration 0, increasing for about 8286% after 100 iterations. Number concentration was suppressed to almost 0, showing a calamitous failure in predictions in conventional Autoencoder. This is not surprising, as the model behave similarly in previous section, meaning that the model may be saturated by the sparse structure of the manifold, indicating the need for a non-zero classifier.

Multi-head Autoencoder offline drift

Multi-head Autoencoder preserves almost perfect mass conservation properties initially, only failing in number concentration of Dust C (Species 3, +4.74%), whilst the other species have an error range of (-1.31,0.8). However, the mass conservation properties are not well preserved, with mass and number concentration Dust C failing at -42.19% & 61.25% respectively. This collinear behaviour of mass and number concentration is no coincidence as a similar distribution was observed in the initial reconstruction (Figure 4.6 c, f). Other species exhibits relatively worse quality compared to NMF after 100 iterations.

It is clear that Non-negative Matrix Factorisation is superior in terms of total mass conservation over long simulations. The total and individual mass remains stable even after 100 iterations. This can be explained intuitively by the mathematical structure of the two models.

By construction, the NMF model is in essence a finite sum of basis vectors, such that insofar as the model learns from a stable, physical system, the constructed basis vector W remains non-negative. The single source of error is formulated as $r = ||Y - WH||$, a linearly dependent system. This means that error growth is bounded and stationary, and does not result in diffusive solutions. As long as the initial reconstruction error is sufficiently small, the solution will remain semi-idempotent.

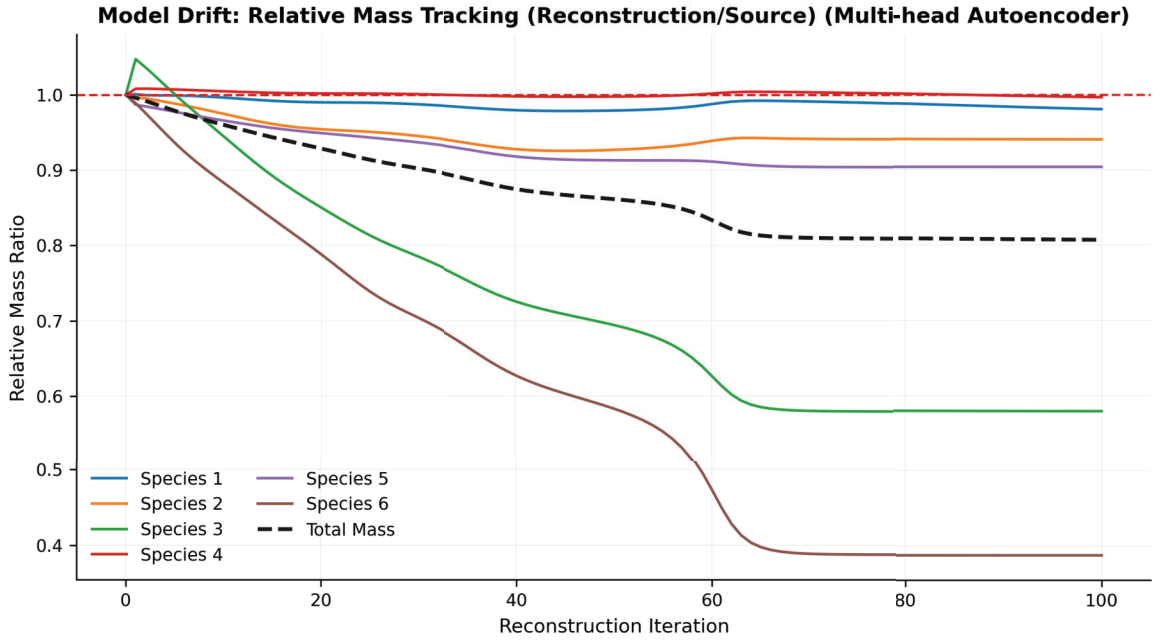

Figure 5.5.: Relative change in total mass for NMF and multi-head Autoencoder over 100 offline embedding-reconstruction iterations. The horizontal axis denotes the iteration number, and the vertical axis the mass ratio $\sum \hat{V} / \sum V$ relative to the source field. A value of 100 % indicates perfect mass conservation. The testing dataset from the full-model simulation is used as the seed.

This however, is not the case for non-linear solutions. Activation functions are non-invertible functions, the choice of activation function may partially contributes to the problem encountered in long term stability.

The output of the decoder is constructed as Equation 2.23, where the relative error is formulated as

$$(\hat{V} - V)/V = (m \cdot \sigma)/V, \tag{5.1}$$

for a small initial error $\epsilon_{m,\sigma}$, the magnitude grows as

$$\hat{V}/V \approx (1 + \epsilon_m)(1 + \epsilon_\sigma) = 1 + \epsilon_m + \epsilon_\sigma + \epsilon_m \epsilon_\sigma \tag{5.2}$$

This means the error grow unboundedly particularly in the 3^{rd} term.

The Softplus function constrains a non-zero magnitude head, which proves to be better as the mapping $(-\infty \to \infty) \mapsto (0 \to \infty)$ is strictly positive, monotonic, smooth function. unlike ReLU, Softplus does not result in mass losses due to the cut off at $x < 0$. However, it introduces a systematic positive bias from the mapping, particularly in values close to 0 as the function has an open interval at 0. This can be observed in mass concentration dust A (Species 1, Figure 4.6 a,c), the lower quartile of the data shows a systematic overshoot, discussed above.

The non-zero head may contributes the most on the systematic mass loss. The output interval of Sigmoid is $(0, 1)$, intended as a classification gate for 0-non-zero classification.

The Sigmoid function is symmetric at (0, 0.5), meaning that the gradient in the output of extreme values vanishes. Small reconstruction error will be mapped towards the centroid of the function, amplifying the error. At each iterations, the Sigmoid gate will slide towards the centre of the function, resulting in a systematic bias as the non-zero gate cannot produce exact 1.0 by construction. This resulted in a slight negative bias over iterations, and is most likely the major contributor of the undershoot identified.

The observed pattern of collinear, large errors for both number and mass conservation is most likely due to several factors.

The dynamic range of dust A and B likely sits in the unsaturated domain of the sigmoid function, where the function is approximately linear, resulting in error propagation similar to that of the NMF model. The systematic errors discussed above will likely cancel out, resulting in high-quality output. Values from dust C likely sit at the saturated sector of the sigmoid function, resulting in extreme values being nudged downwards.

The collinear behaviour of mass and number concentration of dust C is likely due to the loss of signal in the Tanh function. The degraded representations are both poorly conditioned, and the multiplicative update only worsens the output quality.

Online drift experiment

In the online drift experiment, the host model is allowed to run freely for 150 time steps before the machine learning model is called. It must be stressed that this experiment only acts as a proxy for the interaction between the machine learning model and the physical processes in the host model. Actual error growth might differ from that of the true embedding-advection-reconstruction pipeline.

NMF online Drift

The NMF model (Figure 5.6) performed worse compared to the offline drift experiment. The most problematic species at the start of the nudging remain dust C (Species 3,6), with an initial bias of 96%. The remaining species unanimously begin with slight positive bias (ranging from 1–3.5%), gradually decreasing at a much higher rate than number concentration dust C. It is worth noting that the slopes of dust A (Species 1,4) are the highest; the remaining dust B sits between dust A and C. The crossover point occurs at approximately time step 90, where the mass loss of dust A and B overtakes that of dust C, declining at a constant rate.

The initial error is determined by the collinearity of the convex cone produced by the basis vector rotation $\widehat{\mathbf{V}} = \mathbf{WBV}$ and the ground truth. The alignment is then determined by how well the rotated manifold matches the manifold of the training data. In the online drift experiment, the particular entry point (50 time steps post-initialisation) may have a manifold that is particularly dissimilar from the training data, resulting in a low-quality initial construction of dust C.

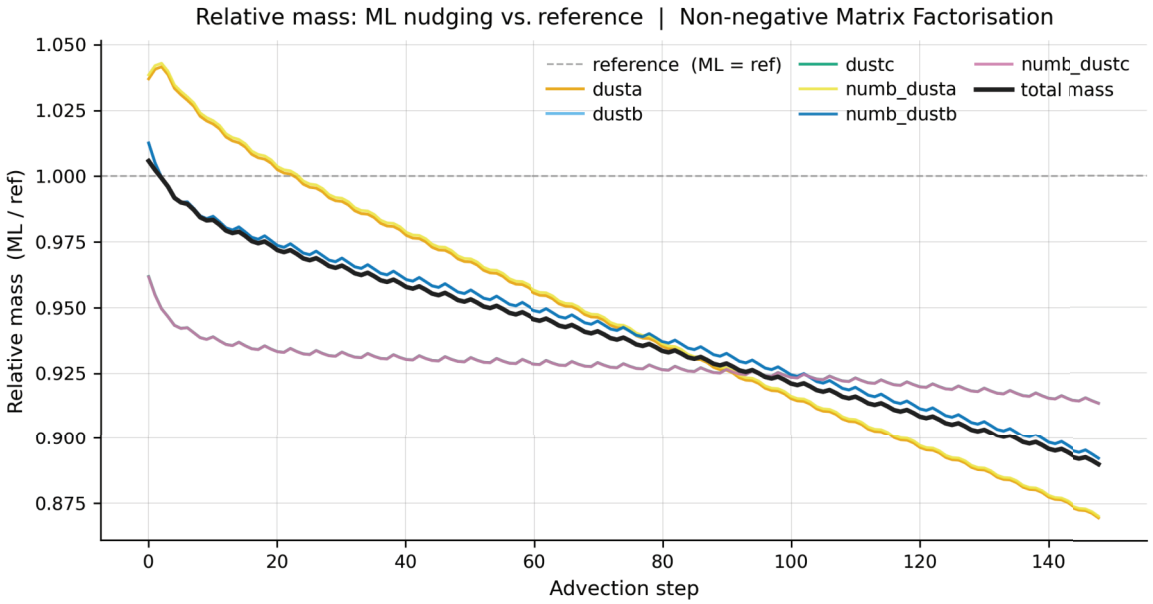

Figure 5.6.: Relative mass evolution of all six species under the NMF online drift experiment. The horizontal axis denotes the time step after nudging begins (150 time steps post-initialisation), and the vertical axis the mass ratio $\sum \hat{V}_k / \sum V_k$ for each species k, where 100 % indicates perfect mass conservation relative to the reference simulation.

The drift rate (dm/dt) is determined by the deformation and translation of the manifold by the host model. Any mathematical operator acting on a manifold can be viewed as a translation and deformation of the data in abstract space. Advection alone denotes a volume-preserving transformation of the space, while processes such as deposition and washout contract the manifold asymmetrically.

For instance, the dry deposition rate is proportional to d^2 in the Stokes regime, while scavenging is strongest in finer modes. The removal processes exhibit different properties; conceptually, sedimentation and dry deposition from the coarser mode (dust C) exhibit better homogeneity than washout, which may result in more complex deformation of the data. This is perhaps the culprit of the difference in drift rate: simpler processes collapse the manifold structure into a subdomain that can be represented by fewer degrees of freedom, which translates to better representation and a lower drift rate in the NMF model.

Multi-head Autoencoder online drift

The most striking feature is the transient mass conservation failure in the first 15 time steps. The initial reconstruction error is remarkably small, consistent with Figure 4.6 and Table 5.3. The initial reconstruction begins at almost 0%, indicating excellent reconstruction quality; however, a sharp decline in mass conservation starts immediately after the first iteration, reaching as high as 35% in mass and number concentration dust B. Dust A exhibits the minimal mass loss, stabilising at 10%, while dust C shows a steeper slope, reaching a

maximum of 25% mass loss. What is unexpected is the recovery of mass conservation with increasing iterations: dust C shows the highest recovery overall, reaching 12.5% mass loss towards the end of the nudging period, with dust B and A recovering 5% and 2.5% mass respectively.

This non-linear behaviour is abnormal and warrant further investigation. This might be explained by an interaction of several mechanisms.

In a coupled online drift experiment, the non-linear response from processes in the host model—namely deposition, washout, and coagulation—plays an essential role in the stability of the solution. Shrinkage and deformation of the manifold structure, as discussed in the NMF online drift experiment, influences the basis vector projection. Let $\sigma(z)$ and $\text{softplus}(z)$ be the output map of the decoder; the reconstruction is

$$\widehat{V} = \sigma(z) \cdot \text{softplus}(z) \qquad | \qquad \sigma \in (0, 1), \text{softplus} \in (0, \infty) \tag{5.3}$$

As discussed in the sections above, σ in the non-zero head is strictly < 1; the magnitude head is therefore conditioned to overshoot the ground truth in order to compensate for the error. This results in a delicate balance between under- and overshoot that could respond non-linearly to an initial perturbation. Here Equation 5.2 is linearised about a state V; the output of the decoder is then:

$$\widehat{V}' = \epsilon_m \epsilon_\sigma V' \tag{5.4}$$

If $\epsilon_m \epsilon_\sigma = \epsilon_{ac}$, the linearised error growth rate (mass loss factor) is then, by substituting to Equation 5.3

$$\epsilon_{ac} = \sigma(z) \cdot \text{softplus}(z)/V' \tag{5.5}$$

For a perfect reconstruction, $\epsilon_{ac} = 1$. However, from construction, $\sigma < 1$, which results in the error gain to be < 1, resulting in systematic mass losses.

Nevertheless, the error growth rate in the online coupled experiment reveals a feedback cycle that results in violent mass loss. With the mass loss factor ϵ_{ac}, the reconstructed data is $\widehat{V} = V + \epsilon_{ac} V$; the sharp features in the original state space are smoothed, as observed above, and the model physics and advection are applied to data with smoothed geometry. As the advection is approximately mass conserving (a fundamental property of the mass transport scheme), the transport applied to the altered field will start to drift away from the ground truth. The physical processes (deposition, coagulation, washout, etc.) respond in various powers of the parameters (e.g. d in deposition and d^2 for coagulation), resulting in asymmetric non-linear responses from the different processes and vastly different drift speeds. The sigmoid gate and Softplus at these larger values may exhibit different dynamic ranges, quickly saturating and producing the mass decay observed in Figure 5.7.

The stabilisation and partial recovery is even more puzzling: mass loss is the expected failure mode by construction, but the monotonically increasing relative mass is not. This behaviour might have two manifestations:

1. True mass gain. Ghost masses are introduced to the system by zero-variance or Softplus leakage, which is not the expected behaviour. As the mean mass has shrunk in the first 15

iterations, the input to the Softplus or sigmoid function returns to the unsaturated domain away from $-/+\infty$. The shrinkage means that the basis vector is rotated to represent a lower-concentration manifold with geometry that resembles a high-concentration manifold. This mismatch could trigger the model to auto-regress to its equilibrium fixed point, resulting in mass gain. 2. Physical processes' compensation. As the mass loss most likely shrinks the extreme values, the source and sink processes may respond non-linearly to the change in extreme values, resulting in a difference in $\frac{dm}{dt}$ between the unaltered model and the nudged model. This shift in deposition rate could potentially drift the absolute concentration values of the ground truth model downwards, closer to that of the nudged model, as the extreme values in the ground truth are well preserved. This will then result in the apparent "mass gain" in terms of relative mass, while in reality it is due to the difference in the time derivative of the absolute concentrations.

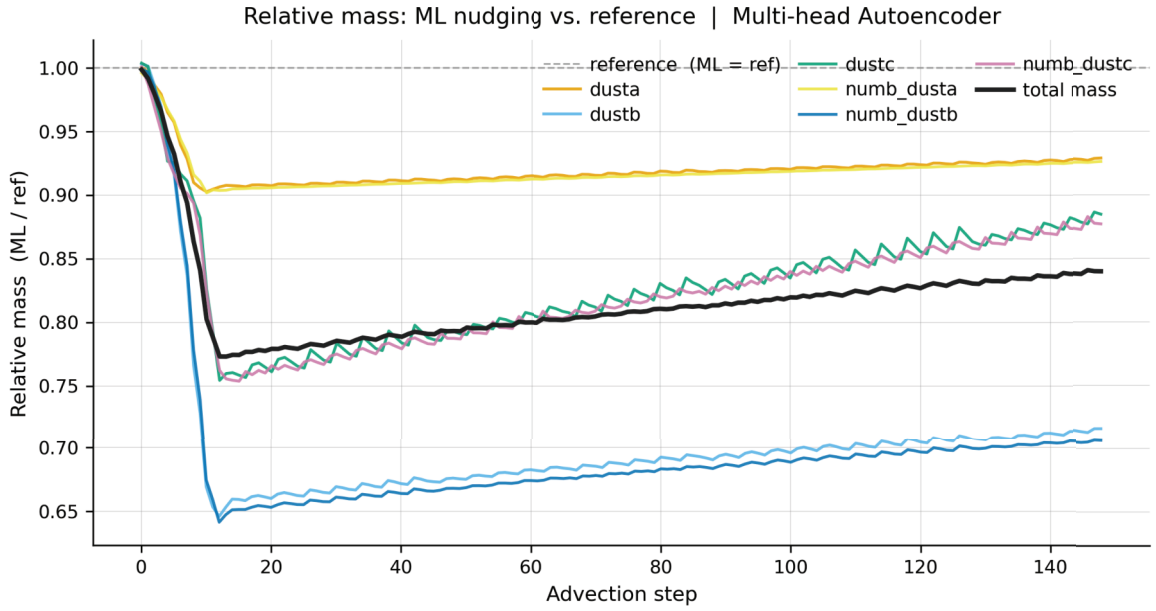

Figure 5.7.: Relative mass evolution of all six species under the multi-head Autoencoder online coupled drift experiment. The horizontal axis denotes the time step after nudging begins, and the vertical axis the mass ratio $\sum \hat{V}_k / \sum V_k$ for each species k relative to the reference simulation (100 %). A sharp initial mass loss is observed in the first 15 time steps, followed by partial recovery in most species.

Such behaviour may raise a suspicion of overfitting. This is, however, not the classical failure mode of overfitting. Overfitting would indeed predict almost perfect reconstruction at the start, but in subsequent iterations the small error would behave similarly to zero-mean noise with no systematic bias, exhibiting a random-walk-like error curve. The multi-head model, however, exhibits error in a monotonically decreasing and then recovering manner, vastly different from the failure mode in overfitting models. Moreover, the drift should be agnostic to all species, resulting in all species degrading at a similar rate, as overfitting would occur on the manifold rather than on individual species.

5.5. Improvement to the Multi-head Autoencoder

Base on the limitations and failure modes identified in the above sections, the author proposes possible avenues of improvement to the Autoencoder architecture, and discusses the reason for input normalisation techniques such as log-scaling and linear scaling (IQR/variance) is unfeasible.

Additional penalty term

Additional penalty can be introduced to the loss function

$$\mathcal{L} = \mathcal{L}_{\text{multi head}} + \alpha_3 \cdot (\sum \hat{V} - \sum V)^2 \tag{5.6}$$

The term penalises departures from mass conservation, where α_3 is the scaling factor relative to the other loss terms, consistent with Equation 2.24. It must be stressed that this will not constrain absolute mass conservation, as the loss function only acts as an attractor for the optimiser. For strict mass conservation, a further mass conservation gate must be introduced; however, in this thesis, the performance benefit of such a gate remains questionable, as the geometry of the manifold may itself be non-conservative.

Output renormalisation

Output renormalisation can be introduced as mass conservation gates

$$\begin{aligned} \mathrm{S} &= \frac{\sum_{i=1}^{m} v_i}{\sum_{j=1}^{r} h_j}, \qquad s.t. \quad \mathrm{S} \in \mathbb{R}^{1\times n} \\ \mathbf{H}_{mass} &= \mathbf{HS}. \end{aligned} \tag{5.7}$$

The error is redistributed across the field; this, however, may contribute negatively to the overall accuracy. For instance, a spatially heterogeneous error pattern (in this case under-representation in a particular region) will cause the mass conservation gate to shift the distribution positively. The error is redistributed, but it introduces unnecessary error at other grid points, deforming the manifold and introducing geometric error that could cause unexpected non-linear responses.

Moreover, for a field-sum mass conservation strategy, global communication is unavoidable—precisely what the entire project aims to avoid. This introduces unnecessary computational overhead and wait time for thread synchronisation, which diminishes the computational benefit of embedding. If local mass conservation is applied within the stencil of each thread, the manifold will no longer be smooth and continuous, as the renormalisation factor will differ between neighbouring cells, causing discontinuities in the solution. As discussed in

Section 2.1.1, the strength of the flux limiter will then inevitably decrease as the solution becomes discontinuous. The advection thus relies on the low-order flux term with lower accuracy and smoother solution, further reducing the accuracy of the advection.

Halo exchange [89] can potentially be introduced, as it is available for native MPI implementations. By exchanging data with the neighbouring stencil, the discontinuities can be smoothed by a Gaussian filter or other smoothing function. This, however, introduces additional wait time and may slow down the embedding-advection-reconstruction cycle. In addition, this requires extensive work to integrate the renormalisation routine with the current MPI implementation in the advection module, which in its entirety is out of scope for a Master's thesis.

Alternative to the Sigmoid gate in the non-zero head

The key issue with the sigmoid gate is the asymptotic saturation at extreme values. This can potentially be solved by introducing a steepness parameter k to the exponential term of the sigmoid function:

$$\sigma_k(x) = \frac{1}{1 + e^{-kx}} \tag{5.8}$$

The steepness parameter k is proportional to the steepness of the well-defined region of the sigmoid function; this means that the transitional region from $\sigma(x) \to 0$ to $\sigma(x) \to 1$ shrinks increasingly as k increases, producing step-function-like behaviour with a smoother gradient. This will allow the non-zero head to perform like a true classification gate with a smooth, differentiable backward trajectory.

Additional heads/parameter

Additional skip connections can be introduced to predict the local mass conservation tendencies. By including new parameters, such as ddt_adv_q* (the advective tracer tendency), an auxiliary correction head can be trained as a residual term at the multiplication gate at output. The target of this auxiliary network will be the ratio of the total mass of the source and reconstructed species, and it will be added as a residual term in Equation 2.23.

Failure of input normalisation

A conventional remedy for datasets spanning multiple orders of magnitude is input normalisation. Two fundamentally different strategies exist: *linear scaling*, which divides the raw input by a dispersion statistic (e.g. IQR or variance) computed in the original, linear space; and *log scaling*, which applies a logarithmic transform $x \mapsto \log(x)$ to compress the dynamic range before training. Each fails for distinct reasons. Most importantly, all architectures are exclusively designed with explicit/implicit non-negativity constraint in mind, applying *log* scaling remaps the input to $(-\infty, \infty)$, necessitating the need for a complete redesign of ALL architectures proposed.

Linear scaling (IQR/variance normalisation). Linear scaling divides each channel by a characteristic spread such as the IQR or standard deviation. Although this centres the data in the unsaturated domain of Tanh, the normalisation constant is itself a linear-space quantity, dominated by the upper tail when the dynamic range spans $O(10^{-5})$ to $O(10^{3})$. A physically meaningful shift from 10^{-6} to $10^{-5.5}$ ($\sim$3.2×) becomes indistinguishable from numerical noise in the scaled space, as a linear dispersion statistic cannot resolve multiplicative structure. Furthermore, the backward pass must propagate through the denormalisation, restoring the original scale disparity, so the gradient dominance of high-magnitude species persists regardless of the scaling applied at the input. Operationally, IQR and variance evolve with the tracer field as source and sink processes alter the distribution, requiring per-time-step global communication that would negate the computational benefit of the embedding.

Log scaling (logarithmic transform). A log transform $x \mapsto \log(x)$ compresses the dynamic range by mapping multiplicative relationships to additive ones, but introduces three failure modes. First, $\log(0)$ is undefined. A floor value $x \mapsto \log(x + \epsilon)$ leaves the non-zero data essentially unaffected (for $x \gg \epsilon$ the shift is negligible), but all true zeros are mapped to a single point mass at $\log(\epsilon)$. A large ϵ collapses zeros and small non-zero values together, destroying the zero/non-zero distinction; a small ϵ pushes the point mass deep into the saturated domain of Tanh (e.g. $\log(10^{-10}) \approx -23$, far below the non-zero data at $\sim [-11.5, 6.9]$), reintroducing the vanishing gradient problem. Second, a small additive error δ in log space becomes multiplicative with denormalisation ($\hat{x} = x \cdot e^{\delta}$), producing large absolute errors at high magnitudes and catastrophic relative errors near zero. Third, mass conservation cannot be maintained, as the field sum $\sum e^{\hat{z}_i}$ bears no algebraic relationship to the original $\sum x_i$, and any spatially correlated bias δ compounds into unbounded mass drift under iteration, analogous to the Sigmoid gate drift identified in Section 5.4.

In summary, linear scaling preserves the zero structure but cannot resolve multi-scale dynamic range, while log scaling compresses the dynamic range but destroys the zero structure and violates mass conservation. This complementary failure further motivates the multi-head architecture, which addresses both constraints simultaneously through separate heads rather than through input normalisation.

		PCA	Non-negative matrix factorisation	Autoencoder with Softplus	Multi-head Autoencoder
Species 1	ACC	−0.0030	0.8339	−0.4538	0.6782
	Wasserstein / IQR	0.1296	2.64	1.1×10^5	0.6599
Species 2	ACC	−0.0001	0.9173	0.0723	0.9646
	Wasserstein / IQR	0.0351	0.265	134.63	0.0533
Species 3	ACC	−0.0030	0.9405	0.6601	0.9799
	Wasserstein / IQR	0.0141	0.183	5.232	0.0627
Species 4	ACC	−0.0003	0.8303	0.4750	0.7176
	Wasserstein / IQR	0.1334	2.65	18.2	0.2353
Species 5	ACC	−0.0001	0.9239	−0.4357	0.9624
	Wasserstein / IQR	0.0365	0.264	0.8913	0.0407
Species 6	ACC	−0.0030	0.9561	0.6541	0.9925
	Wasserstein / IQR	0.0149	0.184	0.8913	0.0130
Total	ACC	0.7627	0.9732	0.0896	0.9986
	Wasserstein / IQR	0.0034	0.0450	0.4128	0.0035

Table 5.1.: Summary of normalised RMSE and manifold statistics of various models. Anomaly Correlation Coefficient & normalised Wasserstein statistics are used. Colour coding for **ACC** (higher = better): green ≥ 0.90; blue 0.75–0.90; amber 0.50–0.75; red 0–0.50; **bold red** < 0 (negative). Colour coding for **Wasserstein / IQR** (lower = better): green ≤ 0.10; blue 0.10–0.50; amber 0.50–5; red 5–100; **bold red** > 100.

		PCA		Non-negative matrix factorisation		Autoencoder		Multi-head Autoencoder	
		Iteration 0	Iteration 100	Iteration 0	Iteration 100	Iteration 0	Iteration 100	Iteration 0	Iteration 100
Species 1	nRMSE	0.19 %	0.19 %	10.1 %	14.6 %	**489,774 %**	**497,903 %**	18.1 %	19.1 %
	$\sigma(\hat{x})/\sigma(x)$	—	—	97.0 %	93.1 %	**50,873 %**	1.1×10^{-5} %	98.7 %	87.4 %
Species 2	nRMSE	1.14 %	1.14 %	104.9 %	135.9 %	**4,916 %**	**4,872 %**	12.5 %	18.4 %
	$\sigma(\hat{x})/\sigma(x)$	—	—	59.9 %	54.3 %	**320.2 %**	**0 %**	95.3 %	92.9 %
Species 3	nRMSE	0.62 %	0.62 %	10.3 %	12.9 %	**845.0 %**	**847.1 %**	12.9 %	63.9 %
	$\sigma(\hat{x})/\sigma(x)$	—	—	101.6 %	97.8 %	29.5 %	**0 %**	93.2 %	38.1 %
Species 4	nRMSE	0.19 %	0.19 %	12.8 %	15.5 %	102.2 %	102.2 %	17.4 %	19.1 %
	$\sigma(\hat{x})/\sigma(x)$	—	—	101.6 %	97.8 %	**0.065 %**	**0 %**	69.2 %	88.3 %
Species 5	nRMSE	1.14 %	1.14 %	141.6 %	**181.3 %**	108.2 %	108.2 %	12.9 %	19.8 %
	$\sigma(\hat{x})/\sigma(x)$	—	—	97.3 %	50.0 %	**0.029 %**	**0 %**	98.5 %	89.3 %
Species 6	nRMSE	0.62 %	0.62 %	19.9 %	26.4 %	115.7 %	115.7 %	10.5 %	82.0 %
	$\sigma(\hat{x})/\sigma(x)$	—	—	93.4 %	89.4 %	**0.077 %**	**0 %**	98.6 %	38.1 %
Total	nRMSE	1.02 %	1.02 %	11.0 %	15.5 %	104.4 %	104.3 %	12.6 %	27.9 %
	$\sigma(\hat{x})/\sigma(x)$	—	—	96.4 %	92.5 %	**0.079 %**	2.3×10^{-8} %	93.7 %	67.4 %

Table 5.2.: Summary of model drift (nRMSE) when applied iteratively. Values for the Autoencoder columns have been converted to percentages (original raw ratios ×100). Colour coding for **nRMSE**: green < 20 % (good); amber 20–50 % (moderate); red 50–200 % (poor); **bold red** > 200 % (extreme). Colour coding for $\sigma(\hat{x})/\sigma(x)$: blue 90–110 % (ideal variance preservation); amber 70–90 % (moderate variance loss); red < 70 % or > 110 % (poor); **bold red** near-zero (variance collapse).

	PCA		Non-negative matrix factorisation		Autoencoder		Multi-head autoencoder	
	Source	Iteration 100	Iteration 0	Iteration 100	Iteration 0	Iteration 100	Iteration 0	Iteration 100
Species 1	100.00 %	100.00 %	98.78 %	95.18 %	**489,714.84 %**	**498,002.34 %**	100.08 %	98.07 %
Species 2	100.00 %	100.00 %	97.84 %	96.16 %	**5,013.10 %**	**4,971.79 %**	99.61 %	94.05 %
Species 3	100.00 %	100.00 %	101.09 %	97.01 %	**943.02 %**	**944.93 %**	104.74 %	57.91 %
Species 4	100.00 %	100.00 %	101.09 %	97.01 %	**0.0008 %**	**0.0008 %**	100.80 %	99.66 %
Species 5	100.00 %	100.00 %	109.58 %	107.98 %	**0.0010 %**	**0.0010 %**	98.69 %	90.40 %
Species 6	100.00 %	100.00 %	98.05 %	94.67 %	**0.0017 %**	**0.0017 %**	98.89 %	38.75 %
Total	100.00 %	100.00 %	98.32 %	94.74 %	**0.0021 %**	**0.0021 %**	99.61 %	80.70 %

Table 5.3.: Offline mass conservation over iteration (values expressed as percentages of source mass; 100 % = perfect conservation). Colour coding: green 95–105 % (conserved); blue > 105 % (over-production); amber 80–95 % (moderate loss); red < 80 % (severe loss); **bold red** > 50× or < 0.5 % (extreme / physically unrealistic).

6. Conclusion & Outlook

This thesis investigated whether a reduced-order representation can serve as a physically consistent and computationally beneficial surrogate model for tracer advection in ICON-ART. The central question is whether a model with its embedding and reconstruction operators $\mathcal{P}, \mathcal{R}$ can project the 3 dust species with mass and number concentration each (3+3) into a lower-dimensional representation. Such projection and reconstruction must preserve non-negativity, mass conservation and manifold geometry insofar as downstream processes such as slow/fast physics and dynamics remain stable. The author constructed an offline training setup in which 4 different embedding models of increasing complexity were adapted and assessed for their physical fidelity, iterative stability and computational benefit.

6.1. Summary of findings

The baseline PCA resulted in the lowest reconstruction error across all species (total normalised RMSE $\sim$ 1% at 4 latent dimensions) and exhibited near-perfect idempotency under iteration. These properties arise exactly from the SVD solution, where the basis vector is rotated to the direction of maximum variance (or remaining variance for principal component > 1). However, PCA is mean-centred and therefore distribution-mean dependent. A shift in the distribution manifold, caused by source and sink processes over long iterations or when applied in different simulations, will result in systematic drift unless the PCA is recomputed on the new dataset. This, however, defeats the purpose of this pilot study, as the aim is to develop a mass-conserving, physically consistent latent representation for operational use. Furthermore, PCA does not guarantee non-negativity, with no way to constrain its numerical behaviour or its physical admissibility. It is therefore understood as the theoretical degrees of freedom of the manifold in abstract space, not a deployable compression strategy.

Non-negative Matrix Factorisation achieved acceptable reconstruction fidelity for 4 out of 6 species ($R^2 = 0.71$–0.88, nRMSE $\sim$ 10–18 %), while mass and number concentration of dust A (Species 1 and 4) remained poorly reconstructed ($R^2 = 0.15$), exhibiting a systematic underestimation bounded by an implicit limiter at approximately 0.6× the source magnitude. The scatter analysis revealed a species-dependent bias pattern, where the direction of the systematic error correlates with the absolute magnitude of the species mean: lower magnitude species (dust A) are underestimated, while the highest magnitude species (dust C) are overestimated, suggesting that the linear basis vectors learnt by NMF project towards

the centroid of the joint distribution. The most prominent deficiency is the zero-variance problem, whereby the model predicts non-zero concentrations in grid cells where the source field is identically zero, introducing spurious "ghost" mass throughout the domain. This arises from the absence of a binary zero/non-zero constraint in the factorisation, a limitation that directly motivated the BCE-driven non-zero head in the multi-head Autoencoder. Despite these shortcomings, NMF guarantees non-negativity by construction through its multiplicative update rule, and exhibited near-idempotent mass conservation in the offline drift experiment: the total mass ratio degraded from 98.32 % to 94.74 % after 100 iterations, with the error growth bounded and stationary as a consequence of the linear algebraic structure $\|\mathbf{Y} - \mathbf{WH}\|$. In the full-model generalisation test, NMF achieved strong manifold preservation (total ACC = 0.9732), though the normalised Wasserstein metric (0.0450) was an order of magnitude larger than PCA and the multi-head Autoencoder (0.0035), confirming that the linear factorisation captures the dominant distributional modes but fails to represent the peripheral states and extreme values that occupy the tails of the tracer manifold.

The gradient-based PCA (Autoencoder with no activation function) failed dramatically for mass concentration dust A (Species 1), as the input signal spans one order of magnitude higher in dynamic range, resulting in the scatter plot (Figure 4.4 a) resembling a vertical line. All input values in the range $(0–10^{-4})$ are compressed into $(0–1 \times 10^{-3})$, signifying that the signal is 20× off (nRMSE: 20.2) from the source mean. This catastrophic failure is understood through scale analysis: the gradient from the larger-magnitude input (number concentration, Species 4–6) overwhelms the optimiser. Results from the model statistics also indicate that the model failed to learn any meaningful structure for mass concentration dust A (Species 1), meaning that the low-order-of-magnitude structure is essentially absent due to the absence of a non-linear activation.

The proposed multi-head Autoencoder, the key novel contribution of this thesis, addresses the limitations identified in the other approaches. By splitting the decoder into a magnitude head (Softplus activation, interval $(0, \infty)$) and a non-zero head (Sigmoid activation, interval $(0, 1)$) after the first dense layer with Tanh activation function, the architecture decouples the representation of absolute magnitude from the zero/non-zero classification. The element-wise product $\widehat{V} = m(z) \cdot \sigma(z)$ extends the effective dynamic range of the output, as arbitrarily large and small values are represented in different heads.

The gradient analysis (Equations 4.5–4.4) showed that the three gradient pathways operate in complementary regimes: the BCE classifier dominates at zero elements, as the gradient of the BCE loss scales in first order whilst other terms scale in second order. A simple scale analysis reveals that they approach the extrema at different rates, confirming the BCE loss's dominance. This means that at this extremum, the classification accuracy dominates the gradient descent, whilst at the other extremum ($p = 1$, non-zero), the BCE loss gradient and the MSE loss gradient through the non-zero head become 0 by construction, leaving the MSE loss gradient as the only non-zero term for gradient computation, preserving concentration fidelity. This prevents the vanishing gradient problem that crippled the conventional Autoencoder.

Analysis of the loss function reveals a similar mathematical construction. The MSE loss is weighted for the non-zero term (in this case, 10), while the factor α_2 scales the BCE

loss to 0.8 of the MSE loss. A simple analysis reveals that in non-zero elements, the total contribution of the BCE loss term to the loss function is only 8% of the total loss, while in zero elements, the contribution of the BCE loss term jumps to 80%, as the MSE loss is not weighted in this case.

The multi-head model achieved strong reconstruction for 4 out of 6 species ($R^2 = 0.82$–0.91, nRMSE ≈ 10–$14\,\%$), with excellent manifold preservation (total ACC = 0.9986). Mass and number concentration of dust A (Species 1 and 4) remained challenging, exhibiting systematic underestimation in the high-value domain and overestimation in the medium-value domain. The author attributes this to the low absolute magnitude of dust A relative to other species, which results in a weaker gradient signal despite the weighted MSE loss. In the full-model generalisation test, the multi-head Autoencoder achieved the best total normalised Wasserstein metric (0.0035) alongside PCA, outperforming NMF (0.0450), confirming that the non-linear embedding captures distributional structure that the linear factorisation cannot.

The conventional 2-layer Autoencoder with Softplus activation, acting as an ablation study, exhibited catastrophic failure. The decoder output collapsed to a narrow band $[0.31, 1.31]$, exactly the composite output interval of the two activation functions $\log(1 + e^{\tanh(\cdot)})$, confirming that the architecture is structurally incapable of representing the full dynamic range of the tracer field once it reaches convergence (Section 4.4.2). The gradient analysis demonstrated again that, for float32 precision, the gradient of the tanh activation becomes indistinguishable from zero beyond $|h| \approx 8.67$, severely limiting the capacity of the latent representation. This is the classical vanishing gradient problem. This finding further consolidates the necessity of the non-zero head, as the raw magnitude head (or, in other words, a conventional Autoencoder with such architecture) will mathematically fail for the desired application.

Spatial error correlation was investigates, revealing a strong correlation of diabatic processes and grid cells with enhanced error. Analysis to the relative mass shows 2 significant pattern: a cyclonic-like mass loss pattern, resembling the Warm Conveyor belt or the occlusion branch of the Extra Tropical Cyclones; and the enhanced mass loss near the Inter-tropical Convergence Zone, and region where convection are the most violence. These region have strong mass loss rate, due to savaging, deposition, washout and other processes. Both model exhibited considerably worse reconstruction at these region, possibly due to the construction of the training data. Training data is comprised of the no-source-sink data, where no mass gain/loss occur in this setup. Although the author successfully trained the model to rely on advection only data, this experiment shows that the generalisation, or rather, the extrapolation from no-source-sink to full model experiment with mass change. This is because, the non-volume preserving deformation brought by physical processes, unlike the volume preserving rotation in the abstract space from the no-source-sink dataset, are new to the model, therefore exhibit stronger error in these sub-manifold.

Furthermore, the iterative stability of the multi-head model proved to be its most significant limitation. After 100 offline iterations, the total mass ratio degraded to 80.70 %, with dust C (Species 3 and 6) exhibiting severe losses of 57.91 % and 38.75 % respectively. The author traced this behaviour to two architectural sources. First, the sigmoid gate in the non-zero

head is strictly less than unity by construction, introducing a systematic negative bias that accumulates over iterations. Second, the multiplicative coupling of the two heads results in error growth that depends on the product $\epsilon_m \cdot \epsilon_\sigma$ (Equation 5.2), which, unlike the additive error of NMF, can grow unboundedly under repeated application. The online drift experiment revealed unexpected non-monotonic behaviour: an initial sharp mass loss in the first 15 time steps, followed by partial recovery. The author hypothesises that this reflects a feedback between the shrinkage of the manifold by the sigmoid gate and the non-linear response of the host model's physical processes, which reduce the deposition rate as extreme values are suppressed by the Autoencoder, partially compensating the mass loss. It must be emphasised that the proxy online drift experiment does not represent the true online embedding-advection-reconstruction pipeline, as such an implementation requires substantial modifications to the advection code base.

6.2. Synthesis

This thesis demonstrated that a reduced-order representation of tracers for the advection module in ICON-ART is indeed feasible, but limitations exist for all models, making them unsuitable for deployment to the ICON code base at the current stage. No model performed satisfactorily and complied with all physical criteria simultaneously. NMF emerges as the most robust candidate in its current form, with non-negativity by construction, near-idempotent mass conservation and acceptable reconstruction fidelity, particularly for species with high absolute magnitudes (dust C). Its linear nature, however, limits its capacity to capture non-linear dependencies in the tracer manifold, which is also reflected in the spatial error analysis: the NMF error pattern at extended lead times shows elevated nRMSE collocated with low-concentration oceanic regions, consistent with the zero-variance problem, though without the sharply localised synoptic signature exhibited by the Multi-head Autoencoder. This suggests that the linear factorisation captures the dominant distributional modes but fails to represent the peripheral states and extreme values that occupy the tails of the tracer manifold.

The multi-head Autoencoder improves upon the previously identified physical constraints, achieving satisfactory geometrical reconstruction and distributional fidelity in a single-pass experiment, outperforming NMF in this respect. The spatial error analysis (Chapter 5) further revealed that the Multi-head Autoencoder produces error maxima sharply collocated with convectively active regions—precisely the zones characterised by strong lifting, cloud interaction and scavenging—where source and sink processes deform the tracer sub-manifolds at species-dependent rates. This tight spatial correspondence confirms that the dominant error source is not an artefact of the model construction, but rather a consequence of the training strategy: by disabling all source and sink processes in pursuit of isolating the pure advection tendency, the training data lack representativeness of the source-sink-affected sub-manifold geometry. Nevertheless, the iterative instability, driven by the asymptotic sigmoid gate and the unbounded multiplicative error growth, precludes the model from practical deployment in its current state. Architectural improvements, such

as those discussed in Section 5.5, are necessary to raise the stability of the solution to a satisfactory level. These include a steepened sigmoid function that more closely approximates a true binary classifier, an additional mass conservation penalty term in the loss function, and auxiliary decoder heads, all offering directions to be explored and discussed again below.

The author further established in the timing study (Section 4.5) that, in the current implementation, the machine-learning–Python–ComIn coupled 4-species advection scheme could potentially achieve a 21% computational benefit compared to advecting 6 species in the base experiment, suggesting that the approach can be financially and computationally attractive especially for complex atmospheric chemistry simulations (such as Vol-aero-rad, Table 1.1) where tens to hundreds of species are advected.

In conclusion, the author has demonstrated—at least partially—that the scientific question identified at the beginning of this thesis is addressable: a physically consistent, reduced-order representation can be constructed for ICON-ART tracers, though not yet deployed without further refinement. The pilot study, using a 6-species Dust-Rad configuration in the ICON standard test suite, establishes a methodological foundation in which several embedding models were proposed and assessed against the design constraints. The dominant failure modes—including the training-strategy-induced degradation in convectively active regions and the multiplicative error growth under iteration—were identified through systematic inter-comparison and supported by concrete mathematical derivations, from which subsequent architectural improvements were proposed. Further development is necessary both architecturally (Autoencoder stability and sub-manifold representativeness) and operationally (ICON-Torch coupling) before a discussion on the operational viability of a coupled deployment can proceed.

6.3. Outlook

This thesis has identified many issues with the currently proposed model, and some possible improvements to the model architecture and operational implementation are already discussed in Section 5.5. This section continues to explore these possible directions.

6.3.1. Further ablation studies on the source of error

The model inter-comparison (Chapter 5) identified several error modes, including the zero-variance problem; poor reconstruction in the lower-absolute-magnitude species (dust A); shrunk variance and high error in some species over long iterations (drift experiments); a strong initial drift in the online coupled experiment; and a catastrophic failure in the ablation study (Autoencoder with the magnitude head only). The author hypothesised several possible reasons for these phenomena, including vanishing gradient problems that arise from the architectural design limiting the range of the converged model, and unbounded error growth. These hypotheses, however, are not rigorously isolated. A

systematic ablation study should be performed to disentangle these complex relations into explainable causalities:

1. **Input rescaling.** Section 5.5 established that both linear scaling (IQR/variance normalisation) and log scaling fail for complementary reasons: the former cannot resolve multiplicative structure across orders of magnitude and does not eliminate gradient dominance, while the latter is incompatible with the zero-rich tracer field and violates mass conservation under denormalisation. Nevertheless, a controlled ablation with linear rescaling can clarify whether the reconstruction failure in low-magnitude species is indeed a gradient domination problem, where the model never converged with respect to these species, or a more fundamental representational limitation. This would also reveal whether the proposed weighted MSE loss is sufficient to compensate for the sparse nature of the manifold, or whether a more aggressive weighting strategy is required.
2. **Zero-variance contribution.** The zero-variance problem, identified in the scatter plot of NMF (Figure 4.1), revealed that the model is very uncertain at or near zero, possibly due to the model's inability to identify zero/non-zero elements. This led to the proposal of the non-zero binary head and the corresponding BCE loss function. Ablating the weight of the BCE loss (adjusting α_3) in the loss function (Equation 2.24) over a wider range could identify the sensitivity of the reconstruction fidelity to the binary constraint, and whether a stronger, true binary constraint such as a steepened sigmoid is necessary.
3. **Activation function saturation.** The gradient analysis in Section 4.4.2 demonstrated that the tanh activation saturates beyond $|h| \approx 8.67$ in float32, a classical vanishing gradient problem. Replacing tanh with alternative bounded activations such as GELU [90] or Mish [91], which exhibit smoother saturation behaviour, could extend the effective channel capacity of the latent representation without violating the non-zero constraint applied at the 2^{nd} layer of the Autoencoder.
4. **Iterative error accumulation.** The author proved mathematically (Section 5.4) that NMF's error growth is bounded and stationary, whereas the error grows unboundedly under the multiplicative coupling of ϵ_m and ϵ_σ in the multi-head Autoencoder. A controlled ablation in which the sigmoid gate is replaced by a hard threshold (step function with straight-through estimator [92]) or a steepened sigmoid (as proposed in Section **??**) would isolate the contribution of the sigmoid's asymptotic saturation to the observed mass loss.

These ablation experiments require no modification to the ICON code base and can be performed entirely within the existing offline framework, making them a natural first step for future work. These ablation studies should be performed simultaneously and presented in a square matrix, such that an intuitive understanding of the error sources can be conveyed.

6.3.2. Architectural improvements

Several architectural improvements can be introduced once ablation studies correctly attribute the sources of error to the corresponding failure modes.

The current architecture adopts a 2-layer encoder/decoder structure with a bottleneck of 4 latent representations, resulting in 33% compression. As the model is introduced to complex atmospheric chemistry simulations with 10+ species, the bottleneck size and neuron count of each layer must adapt to the increase in input channels. Although Grinsztajn *et al.* [64] argues that a 2-layer network is sufficient for tabular-style data, as the input channel scales with increasing number of species, a deeper architecture with possibly residual connections [93] would allow the model to learn more complex non-linear feature maps that are necessary to achieve an even higher compression ratio than the pilot study. A variant of the attention mechanism can also be implemented by evaluating the transport tendency ddt_adv_q*, addressing the magnitude imbalance discussed above.

The multi-head loss function (Equation 2.24) currently combines weighted MSE and BCE with fixed coefficients. Adaptive loss weighting strategies, such as uncertainty-based weighting [94] or gradient normalisation [95], could dynamically balance the contributions of the two heads during training, reducing the sensitivity to hyperparameter choices and potentially improving reconstruction quality for the poorly performing species.

Additionally, the auxiliary correction head proposed in the improvement section of Chapter 5 deserves further attention. By training an additional head to predict the local mass conservation tendency as a residual correction, the architecture could learn to compensate for the systematic bias introduced by the Sigmoid gate, without requiring global communication. This approach is conceptually similar to the residual learning framework of He *et al.* [93], where the network learns the deviation from an identity mapping rather than the full transformation.

6.3.3. GPU acceleration and computational coupling strategies

The timing study (Section 4.5) demonstrated a 21 % reduction in advection cost under the current Python-ComIn coupling. However, the Python interface introduces overhead in Fortran-Python coupling and I/O, and the model inference is executed on CPU. Several coupling strategies can potentially improve the computational benefit:

FTorch coupling. FTorch [96] provides a direct Fortran interface to the PyTorch C++ backend (LibTorch), eliminating the Python interpreter overhead entirely. Kumar *et al.* [86] implemented a FTorch-ICON coupling for Mie radiation emulation, achieving near-native Fortran performance for the inference step. Porting the inference to a native Fortran routine using FTorch would remove the I/O and efficiency problem and enable near-native execution speed. This, however, requires substantial changes to the ICON code base and would therefore be the final step in the operational deployment of the reduced-order model.

GPU offloading. Lapillonne *et al.* [87] introduced GPU support for ICON version 2024.10, enabling GPU-only host model computation. This enables the workload of the embedding to also be transferred to the GPU, such that the embedding-advection-reconstruction cycle can be implemented without host-device memory I/O. This eliminates the additional wait time and communication latency brought by RAM-VRAM access in heterogeneous CPU-GPU deployment, which would become a major slowdown for the model. For the Autoencoder, the forward pass—including Tanh, Softplus, and sigmoid activations and linear layers—maps naturally to CUDA tensor operations. For NMF, the vectorised matrix operations **BV**, **WH** are standard `cublasSgemm` calls. A coarser-resolution simulation that can be run on a single GPU node could serve as the test bed for quantifying the actual speedup under full GPU implementation.

HALO exchange for local mass conservation The output renormalisation strategy discussed in Section 5.5 requires global communication to compute the field-sum mass ratio **S** (Equation 5.7), which violates the principle of avoiding global communication and the associated wait time for process synchronisation. A purely local mesh renormalisation, however, is applied independently within the stencil of each MPI thread, resulting in discontinuities at the mesh boundaries. A discontinuous field, as discussed in Section 2.1.1, forces the flux limiter to fall back to the low-order, diffusive flux term, further resulting in accuracy loss in the subsequent advection step.

Halo exchange [89], currently used by ICON, offers neighbourhood grid communication. This can serve as an intermediate solution between global and purely local renormalisation. Data are communicated at the boundaries of each stencil; when combined with a Gaussian filter or weighted average over the neighbourhood cells, this offers a smoother gradient in the normalisation factor. This eases the discontinuities, increasing the proportion of the high-order flux, resulting in higher-accuracy advection and lower diffusion while maintaining minimal inter-process communication.

The implementation would require substantial modification of the ICON code base, such that the renormalisation routine is embedded within the existing MPI communication infrastructure of the ICON advection. This means that a C++–ComIn coupling is no longer possible; a native FTorch implementation must be introduced for the embedding-reconstruction cycle. ICON already performs halo exchanges natively for the prognostic variables, and an additional exchange for the renormalisation could in principle be introduced with minimal additional latency. However, the non-linearity of the flux limiter, renormalisation and the Gaussian smoothing must be carefully validated, as artefacts and reconstruction error may be introduced. A systematic study of the halo-smoothing-embedding's impact on both mass conservation and advection accuracy is therefore necessary to identify the sources of error.

Extension to full chemistry configurations This thesis introduced a pilot study using DUST-RAD (3 dust species with mass and number concentration each). The compression ratio is therefore 33%, a modest compression. The true potential, as discussed in Chapter 1, lies in comprehensive atmospheric chemistry simulations where tens to hundreds of species are advected.

In such configurations, the tracer manifold may be embedded at a much higher compression rate, as species sharing common emission sources, chemical pathways or removal mechanisms occupy a manifold with substantially lower intrinsic dimension. The computational benefit could then be significantly higher than in the pilot study. However, the heterogeneity of the species may increase, as chemically distinct species groups (e.g. VOCs, secondary aerosol and mineral dust) occupy disjoint sub-manifolds arising from different removal and chemical reaction properties. This can be alleviated by a similar strategy to that adopted in Sturm *et al.* [15], where each group of species has an independent embedding model.

Fully coupled online experiment The author implemented a proxy online drift experiment in Section 5.4, where the machine learning model nudges the tracer field before advection. This experiment reveals mass conservation failures and species-specific shrinkage in dynamic range and reconstruction fidelity. Even if these issues are addressed, it does not necessarily lead to a successful implementation of an embedding-advection-reconstruction scheme. Additional experiments with coupled embedding–latent-species advection–reconstruction must be performed to study the problems introduced by advecting latent species, e.g. mass loss by diffusion and discontinuity.

A realistic implementation can begin by extending the experiment to a full chemistry setup, for instance Volaero-rad introduced in Table 1.1. It features 51 tracers of several types, allowing various configurations to be tested.

Subsequently, GPU offloading should be explored to quantify the computational overhead of a CPU-host GPU-ML model implementation, as a pure GPU implementation faces limitations from the available GPU resources on computer clusters. Furthermore, the ICON-GPU implementation may not have full functionality and could introduce additional bugs and computational issues. From this, a FTorch coupling of the ML model can be introduced as an internal routine in ICON advection, such that the model can flexibly adapt to GPU- or CPU-based emulation.

Halo exchanges for non-local stencil mass conservation can be implemented alongside the fully coupled online experiments. Such experiments or implementations would require substantial modifications of the ICON advection module, as discussed, since the advection currently accepts by default the same number of species as other processes. The advection must be modified such that the number of species advected can adapt to its input shape, and assumptions in the flux limiter or advection module may be altered by the embedding (e.g. boundary conditions, boundedness of the solution). This is an open and technically challenging question that can only be answered empirically through a fully coupled experiment. It is a natural continuation of the work performed in this thesis, where the author has provided the necessary methodological framework, evaluation tactics and directions for future work.

Bibliography

1. Pöschl, U. Atmospheric aerosols: composition, transformation, climate and health effects. *Angewandte Chemie International Edition* **44,** 7520–7540 (2005).
2. Lauritzen, P. H., Nair, R. D. & Ullrich, P. A. A conservative semi-Lagrangian multi-tracer transport scheme (CSLAM) on the cubed-sphere grid. *Journal of Computational Physics* **229,** 1401–1424 (2010).
3. Dubey, S., Mittal, R. & Lauritzen, P. H. A flux-form conservative semi-Lagrangian multitracer transport scheme (FF-CSLAM) for icosahedral-hexagonal grids. *Journal of Advances in Modeling Earth Systems* **6,** 332–356 (2014).
4. Prill, F., Reinert, D., Rieger, D. & Zängl, G. *ICON tutorial* tech. rep. (Deutsche Wetterdienst, 2022).
5. Weimer, M. *et al.* An emission module for ICON-ART 2.0: implementation and simulations of acetone. *Geoscientific Model Development* **10,** 2471 (2017).
6. Lin, S.-J. & Rood, R. B. Multidimensional flux-form semi-Lagrangian transport schemes. *Monthly weather review* **124,** 2046–2070 (1996).
7. Herzfeld, M. & Engwirda, D. A Flux-Form Semi-Lagrangian advection scheme for tracer transport on arbitrary meshes. *Ocean Modelling* **181,** 102140 (2023).
8. Asmouh, I., El-Amrani, M., Seaid, M. & Yebari, N. A conservative semi-Lagrangian finite volume method for convection–diffusion problems on unstructured grids. *Journal of Scientific Computing* **85,** 11 (2020).
9. Li, Z. *et al.* Aerosol optical properties and their radiative effects in northern China. *Journal of Geophysical Research: Atmospheres* **112** (2007).
10. Stier, P., Seinfeld, J. H., Kinne, S. & Boucher, O. Aerosol absorption and radiative forcing. *Atmospheric Chemistry and Physics* **7,** 5237–5261 (2007).
11. Chung, C. E. Aerosol direct radiative forcing: a review. *Atmospheric aerosols—regional characteristics—chemistry and physics,* 379–394 (2012).
12. Gordon, H., Glassmeier, F. & T. McCoy, D. An overview of aerosol-cloud interactions. *Clouds and Their Climatic Impacts: Radiation, Circulation, and Precipitation,* 13–45 (2023).
13. Cairo, F., Di Liberto, L., Dionisi, D. & Snels, M. Understanding aerosol–cloud interactions through lidar techniques: a review. *Remote Sensing* **16,** 2788 (2024).
14. Dagan, G., Yeheskel, N. & Williams, A. I. Radiative forcing from aerosol–cloud interactions enhanced by large-scale circulation adjustments. *Nature Geoscience* **16,** 1092–1098 (2023).

15. Sturm, P. O. *et al.* Advecting superspecies: Efficiently modeling transport of organic aerosol with a mass-conserving dimensionality reduction method. *Journal of Advances in Modeling Earth Systems* **15,** e2022MS003235 (2023).
16. Kramer, M. A. Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal* **37,** 233–243 (1991).
17. Zhai, J., Zhang, S., Chen, J. & He, Q. *Autoencoder and its various variants* in *2018 IEEE international conference on systems, man, and cybernetics (SMC)* (2018), 415–419.
18. Févotte, C. & Idier, J. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural computation* **23,** 2421–2456 (2011).
19. Strassmann, D. *Investigation of Higher-Order Tracer Advection in ICON* (2024).
20. Skamarock, W. C. Positive-definite and monotonic limiters for unrestricted-time-step transport schemes. *Monthly weather review* **134,** 2241–2250 (2006).
21. Zalesak, S. T. Fully multidimensional flux-corrected transport algorithms for fluids. *Journal of computational physics* **31,** 335–362 (1979).
22. LeVeque, R. J. *Finite volume methods for hyperbolic problems* (Cambridge university press, 2002).
23. Sengupta, T. K., Ganerwal, G. & Dipankar, A. High accuracy compact schemes and Gibbs' phenomenon. *Journal of Scientific Computing* **21,** 253–268 (2004).
24. Prill, F. & Eser, C. *Reports on ICON* tech. rep. (Deutsche Wetterdienst, 2022).
25. Griebel, M., Rieger, C. & Schier, A. in *Transport Processes at Fluidic Interfaces* 145–175 (Springer, 2017).
26. Zängl, G., Reinert, D., Rípodas, P. & Baldauf, M. The ICON (ICOsahedral Non-hydrostatic) modelling framework of DWD and MPI-M: Description of the non-hydrostatic dynamical core. *Quarterly Journal of the Royal Meteorological Society* **141,** 563–579 (2015).
27. Malardel, S. *et al.* A new grid for the IFS. *ECMWF newsletter* **146,** 321 (2016).
28. De Moura, C. A. & Kubrusly, C. S. The courant–friedrichs–lewy (cfl) condition. *AMC* **10,** 45–90 (2013).
29. Rieger, D. *et al.* ICON-ART 1.0–a new online-coupled model system from the global to regional scale. *Geoscientific Model Development Discussions* **8,** 567–614 (2015).
30. Schröter, J. *et al.* ICON-ART 2.1: A flexible tracer framework and its application for composition studies in numerical weather forecasting and climate simulations. *Geoscientific Model Development* **11,** 4043–4068 (2018).
31. Hogan, R. J. & Bozzo, A. *ECRAD: A new radiation scheme for the IFS* (European Centre for Medium-Range Weather Forecasts Reading, UK, 2016).
32. Wang, C. & Picaut, J. Understanding ENSO physics—A review. *Earth's Climate: The Ocean–Atmosphere Interaction, Geophys. Monogr* **147,** 21–48 (2004).
33. Hurrell, J. W., Kushnir, Y., Ottersen, G. & Visbeck, M. An overview of the North Atlantic oscillation. *Geophysical monograph-American geophysical union* **134,** 1–36 (2003).

34. Wang, J. & Ikeda, M. Arctic oscillation and Arctic sea-ice oscillation. *Geophysical Research Letters* **27,** 1287–1290 (2000).

35. Del Sozzo, E., Fleming, M., Flamm, K. & Thompson, N. How Much Progress Has There Been in NVIDIA Datacenter GPUs? *arXiv preprint arXiv:2601.20115* (2026).

36. Rasp, S. & Thuerey, N. Data-driven medium-range weather prediction with a resnet pretrained on climate simulations: A new model for weatherbench. *Journal of Advances in Modeling Earth Systems* **13,** e2020MS002405 (2021).

37. Scher, S. & Messori, G. Weather and climate forecasting with neural networks: using general circulation models (GCMs) with different complexity as a study ground. *Geoscientific Model Development* **12,** 2797–2809 (2019).

38. Grönquist, P. *et al.* Deep learning for post-processing ensemble weather forecasts. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **379** (2021).

39. Shi, X. *et al.* Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems* **28** (2015).

40. Weyn, J. A., Durran, D. R. & Caruana, R. Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *Journal of Advances in Modeling Earth Systems* **12,** e2020MS002109 (2020).

41. Lam, R. *et al.* Learning skillful medium-range global weather forecasting. *Science* **382,** 1416–1421 (2023).

42. Bi, K. *et al.* Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecast. *arXiv preprint arXiv:2211.02556* (2022).

43. Lang, S. *et al.* AIFS–ECMWF's data-driven forecasting system. *arXiv preprint arXiv:2406.01465* (2024).

44. Prill, F. & Jacob, M. Apr. 2025. https://events.ecmwf.int/event/464/contributions/5252/attachments/3223/5514/HPC-WS_Prill.pdf.

45. Grundner, A. *et al.* Deep learning based cloud cover parameterization for ICON. *Journal of Advances in Modeling Earth Systems* **14,** e2021MS002959 (2022).

46. Hannachi, A., Jolliffe, I. T., Stephenson, D. B., *et al.* Empirical orthogonal functions and related techniques in atmospheric science: A review. *International journal of climatology* **27,** 1119–1152 (2007).

47. Prasad, K. M. & Bapat, R. The generalized moore-penrose inverse. *Linear Algebra and its Applications* **165,** 59–69 (1992).

48. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *nature* **401,** 788–791 (1999).

49. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

50. Ward, R., Wu, X. & Bottou, L. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. *Journal of Machine Learning Research* **21,** 1–30 (2020).

51. Seung, D., Lee, L., *et al.* Algorithms for non-negative matrix factorization. *Advances in neural information processing systems* **13,** 35 (2001).
52. Choromanska, A., Henaff, M., Mathieu, M., Arous, G. B. & LeCun, Y. *The loss surfaces of multilayer networks* in *Artificial intelligence and statistics* (2015), 192–204.
53. Olaya, J. & Otman, C. *Beta-divergence for Nonnegative Matrix Factorization* in *2021 International Conference on Digital Age & Technological Advances for Sustainable Development (ICDATA)* (2021), 15–22.
54. Hinton, G. E. & Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *science* **313,** 504–507 (2006).
55. Prince, S. J. *Understanding deep learning* (MIT press, 2023).
56. Goodfellow, I. *Deep learning* 2016.
57. Cybenko, G. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems* **2,** 303–314 (1989).
58. Hornik, K., Stinchcombe, M. & White, H. Multilayer feedforward networks are universal approximators. *Neural networks* **2,** 359–366 (1989).
59. Le Cun, Y. & Fogelman-Soulié, F. Modèles connexionnistes de l'apprentissage. *Intellectica* **2,** 114–143 (1987).
60. Bourlard, H. & Kamp, Y. Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics* **59,** 291–294 (1988).
61. Hinton, G. E. & Zemel, R. Autoencoders, minimum description length and Helmholtz free energy. *Advances in neural information processing systems* **6** (1993).
62. Popescu, M.-C., Balas, V. E., Perescu-Popescu, L. & Mastorakis, N. Multilayer perceptron and neural networks. *WSEAS transactions on circuits and systems* **8,** 579–588 (2009).
63. Hara, K., Saito, D. & Shouno, H. *Analysis of function of rectified linear unit used in deep learning* in *2015 international joint conference on neural networks (IJCNN)* (2015), 1–8.
64. Grinsztajn, L., Oyallon, E. & Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems* **35,** 507–520 (2022).
65. Ho, N.-H., Yang, H.-J., Kim, S.-H. & Lee, G. Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network. *IEEE Access* **8,** 61672–61686 (2020).
66. Zou, Z. & Karniadakis, G. E. L-HYDRA: multi-head physics-informed neural networks. *arXiv preprint arXiv:2301.02152* (2023).
67. Conceicao, R. *et al.* Saharan dust transport to Europe and its impact on photovoltaic performance: A case study of soiling in Portugal. *Solar Energy* **160,** 94–102 (2018).
68. Caruana, R. *Multitask learning: A knowledge-based source of inductive bias1* in *Proceedings of the Tenth International Conference on Machine Learning* (1993), 41–48.
69. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32** (2019).

70. Blanchard, P., Higham, D. J. & Higham, N. J. Accurately computing the log-sum-exp and softmax functions. *IMA Journal of Numerical Analysis* **41,** 2311–2330 (2021).

71. Nielsen, F. & Sun, K. Guaranteed bounds on information-theoretic measures of univariate mixtures using piecewise log-sum-exp inequalities. *Entropy* **18,** 442 (2016).

72. Dekking, F. M., Kraaikamp, C., Lopuhaä, H. P. & Meester, L. E. *A Modern Introduction to Probability and Statistics: Understanding why and how* (Springer, 2005).

73. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

74. Bock, S., Goppold, J. & Weiß, M. An improvement of the convergence proof of the ADAM-Optimizer. *arXiv preprint arXiv:1804.10587* (2018).

75. Amari, S.-i. Backpropagation and stochastic gradient descent method. *Neurocomputing* **5,** 185–196 (1993).

76. Ceneda, N. *Quantile Regression of High-Frequency Data Tail Dynamics via a Recurrent Neural Network* PhD thesis (Dissertation, 05 2020, 2020).

77. https://scikit-learn.org/stable/modules/cross_validation.html.

78. Wong, T.-T. & Yeh, P.-Y. Reliable accuracy estimates from k-fold cross validation. *IEEE Transactions on Knowledge and Data Engineering* **32,** 1586–1594 (2019).

79. Hartung, K., Jöckel, P., Kerkweg, A., Kern, B. & Loch, W. Connecting Codes to ICON via the Community Interface (ComIn)-The Modular Earth Submodel System as a first complex ComIn plugin (2024).

80. Chai, T., Draxler, R. R., *et al.* Root mean square error (RMSE) or mean absolute error (MAE). *Geoscientific model development discussions* **7,** 1525–1534 (2014).

81. Wilks, D. S. *Statistical methods in the atmospheric sciences* (Academic press, 2011).

82. Panaretos, V. M. & Zemel, Y. Statistical aspects of Wasserstein distances. *Annual review of statistics and its application* **6,** 405–431 (2019).

83. Li, X. *et al. Discovery of Floating-Point Differences Between NVIDIA and AMD GPUs* in *2024 IEEE 24th International Symposium on Cluster, Cloud and Internet Computing (CCGrid)* (2024), 663–666.

84. Duff, I. S., Heroux, M. A. & Pozo, R. An overview of the sparse basic linear algebra subprograms: The new standard from the BLAS technical forum. *ACM Transactions on Mathematical Software (TOMS)* **28,** 239–267 (2002).

85. Habibzadeh, M., Shishvan, O. R. & Soyata, T. in *GPU Parallel Program Development Using CUDA* 383–395 (Chapman and Hall/CRC, 2018).

86. Kumar, P., Vogel, H., Bruckert, J., Muth, L. J. & Hoshyaripour, G. A. MieAI: a neural network for calculating optical properties of internally mixed aerosol in atmospheric models. *npj Climate and Atmospheric Science* **7,** 110 (2024).

87. Lapillonne, X. *et al.* Operational numerical weather prediction with ICON on GPUs (version 2024.10). *Geoscientific Model Development* **19,** 755–772 (2026).

88. Piccoli, B. & Rossi, F. On properties of the generalized Wasserstein distance. *Archive for Rational Mechanics and Analysis* **222,** 1339–1365 (2016).

89. Spies, L., Bienz, A., Moulton, D., Olson, L. & Reisner, A. Tausch: A halo exchange library for large heterogeneous computing systems using MPI, OpenCL, and CUDA. *Parallel Computing* **114,** 102973 (2022).

90. Hendrycks, D. & Gimpel, K. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415* (2016).

91. Misra, D. Mish: A self regularized non-monotonic activation function. *arXiv preprint arXiv:1908.08681* (2019).

92. Bengio, Y., Léonard, N. & Courville, A. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432* (2013).

93. He, K., Zhang, X., Ren, S. & Sun, J. *Deep residual learning for image recognition* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), 770–778.

94. Kendall, A., Gal, Y. & Cipolla, R. *Multi-task learning using uncertainty to weigh losses for scene geometry and semantics* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), 7482–7491.

95. Chen, Z., Badrinarayanan, V., Lee, C.-Y. & Rabinovich, A. *Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks* in *International conference on machine learning* (2018), 794–803.

96. Atkinson, J. *et al.* FTorch: a library for coupling PyTorch models to Fortran. *The Journal of Open Source Software* (2025).

A. Hyperparameters

Loss function	Name	Value	Description
WeightedMSELoss	nonzero_weights	10	Weight of the non-zero elements have on the loss score
MultiHeadLoss	a3	0.8	Weight of BCE Loss to *WeightedMSELoss* (MSELoss is 1.0)

Table A.1.: Final hyperparameters for the multi-head loss function (Equation 2.24). The *nonzero_weights* parameter scales the contribution of non-zero elements in the *WeightedMSELoss*, and *a3* controls the relative weight of the *BCEWithLogitsLoss* to the magnitude loss. Values are selected via cross-validated grid search.

Hyper Parameter	Value	Detail
Beta	0.5	Shpae of the map x->d(x), See 2.4.0.1
L1-ratio	2.0	Ratio of L1-> L2, L2 regularisation=0 and L1 regularisation=1
Alpha	0.0005	Strength of regularisation
Rank	4	Number of superspecies
Tolerance	0.0001	Tolerance value for convergence

Table A.2.: Hyperparameters of the Non-negative Matrix Factorisation.

Hyper Parameter	Value	Detail
	Model + AdamW	
Batch size	1048576	Batch size of each mini batch
Learning Rate	0.0057	Ratio of L1-> L2, L2 regularisation=0 and L1 regularisation=1
Neuron size	12	Size of hidden dimension
Rank	4	Bottleneck layer size (Latent dimension)
Weight decay	10^{-3}	Decay rate of weights, L2 regularisation like
Betas	(0.89, 0.92)	Coefficient for g_t, g_t^2 of gradients respectively
	LR Scheduler	
Factor	0.95	Decay rate of the learning rate
Patience	20	Number of epoches of stagnation in training loss before the factor is applied to the learning rate
Threshold	10^{-4}	Threshold for measuring the new optimum
Minimum LR	10^{-5}	Lower bound of learning rate

Table A.3.: Hyperparameters of Autoencoder with no activation function.

Hyper Parameter	Value	Detail
	Model + AdamW	
Batch size	1048576	Batch size of each mini batch
Learning Rate	0.0057	Ratio of L1-> L2, L2 regularisation=0 and L1 regularisation=1
Neuron size	12	Size of hidden dimension
Rank	4	Bottleneck layer size (Latent dimension)
Weight decay	10^{-4}	Decay rate of weights, L2 regularisation like
Betas	(0.8,0.8)	Coefficient for g_t, g_t^2 of gradients respectively
	LR Scheduler	
Factor	0.99	Decay rate of the learning rate
Patience	20	Number of epoches of stagnation in training loss before the factor is applied to the learning rate
Threshold	10^{-4}	Threshold for measuring the new optimum
Minimum LR	10^{-5}	Lower bound of learning rate

Table A.4.: Hyperparameters of Multi-head Autoencoder.

Hyper Parameter	Value	Detail
	Model + AdamW	
Learning Rate	0.01	Ratio of L1-> L2, L2 regularisation=0 and L1 regularisation=1
Neuron size	12	Size of hidden dimension
Rank	4	Bottleneck layer size (Latent dimension)
Weight decay	10^{-2}	Decay rate of weights, L2 regularisation like
Betas	(0.95,0.99)	Coefficient for g_t, g_t^2 of gradients respectively
	LR Scheduler	
Factor	0.955	Decay rate of the learning rate
Patience	20	Number of epoches of stagnation in training loss before the factor is applied to the learning rate
Threshold	10^{-4}	Threshold for measuring the new optimum
Minimum LR	10^{-5}	Lower bound of learning rate

Table A.5.: Hyperparameters of the Softplus Autoencoder in ablation study.

B. Source Code modification - No source sink

B.1. Sedimentation

Fortran Code

```
! path:
externals/art/interface/mo_art_sedimentation_interface.f90
!Line 409
sedim_update = (dt_sub * ( p_upflux_sed(jc,jk,  jb)   &
&                      - p_upflux_sed(jc,jkp1,jb) ) &
&         / rhodz_new(jc,jk,jb))
tracer(jc,jk,jb,jsp) =   tracer(jc,jk,jb,jsp) - &
&           (0.0_wp * sedim_update)

! Line 511
sedim_update = (dt_sub * ( p_upflux_sed(jc,jk,  jb)   &
  &                      - p_upflux_sed(jc,jk+1,jb) ) &
  &         / rhodz_new(jc,jk,jb))
tracer(jc,jk,jb,jsp) =   tracer(jc,jk,jb,jsp) - &
&           (0.0_wp * sedim_update)

! Line 552
sedim_update = (dt_sub * ( p_upflux_sed(jc,jk,  jb)   &
&                      - p_upflux_sed(jc,jkp1,jb) )  &
&         / rhodz_new(jc,jk,jb))
tracer(jc,jk,jb,jsp) =   tracer(jc,jk,jb,jsp) - &
&           (0.0_wp * sedim_update)
```

B.2. 2 Moment Deposition Velocity

Fortran Code

```
! path:
externals/art/aerosol_dynamics/mo_art_depo_2mom.f90
! Line 121
! ---------------------------------
! --- Moment 0
! ---------------------------------
<...>
vdep0(jc) = (0.0_wp / (raerody + rd0))

! Line 134
! ---------------------------------
! --- Moment 3
! ---------------------------------
<...>
vdep3(jc) = 0.0_wp * ( 1.0_wp / (raerody + rd3))
```

B.3. Dry Deposition - Radioactive aerosol

Fortran Code

```
! path:
externals/art/aerosol_dynamics/mo_art_drydepo_radioact.f90
! Line 72
DO jc = istart, iend
    p_art_data%turb_fields%vdep(jc,jb,jsp) = &
    &       0.0_wp * vdep_const
ENDDO
```

B.4. 1 Moment Sedimentation Scheme

Fortran Code

```
! path:
externals/art/aerosol_dynamics/mo_art_sedi_1mom.f90
! Line 174
DO jc = istart, iend
    p_art_data%turb_fields%vdep(jc,jb,jsp) = &
    &       (0.0_wp * vsed(jc,nlev))
END DO
```

B.5. 2 Moment Sedimentation Scheme

Fortran Code

```
! path:
externals/art/aerosol_dynamics/mo_art_sedi_2mom.f90
! Line 99
! CALL art_clip_gt(vsed3(jc,jk),1.0_wp)
! CALL art_clip_gt(vsed0(jc,jk),1.0_wp)
! Set vsed to 0
vsed0(jc,jk) = 0.0_wp
vsed3(jc,jk) = 0.0_wp
```

B.6. Aerosol Washout scheme

Fortran Code

```
! path:
externals/art/aerosol_dynamics/mo_art_washout_aerosol.f90
! Line 250
 wash_rate_m0(i,k) = 0._wp * ((-1.0_wp)*dnmbdt)
 wash_rate_m3(i,k) = 0._wp * ((-1.0_wp)*dnmbdt *
    &             (diam_aero(i,k))**(3.0_wp) &
    &           * EXP(4.5_wp *(LOG(sg_aero))**(2.0_wp)))
```

B.7. Radioactive Aerosol washout scheme

Fortran Code

```
! path:
externals/art/aerosol_dynamics/mo_art_washout_radioact.f90
! Line 91
! Washout by rain
IF ( qr(jc,jk)> 0.0_wp .OR. rr_con3d(jc,jk)> 0.0_wp) THEN
    IF ((rr_con(jc) + rr_gsp(jc)) > 0.0_wp) THEN
      washout_rate  = 0.0_wp * factor_a * wdf * &
      &             ((rr_gsp(jc)+rr_con(jc)) ** wde)
      tracer(jc,jk) = tracer(jc,jk) - (tracer(jc,jk) * &
      &             washout_rate * dtime)
    ENDIF
```

```
ENDIF

! Washout by snow
IF ( qs(jc,jk)> 0.0_wp .OR. ss_con3d(jc,jk)> 0.0_wp) THEN
    IF ((ss_con(jc) + ss_gsp(jc)) > 0.0_wp) THEN
      washout_rate  = 0.0_wp * factor_b * wdf * &
      &             ((ss_gsp(jc)+ss_con(jc)) ** wde)
      tracer(jc,jk) = tracer(jc,jk) - (tracer(jc,jk) * &
      &             washout_rate * dtime)
    ENDIF
ENDIF
```

B.8. Volcanic Ash washout scheme

Fortran Code

```
! path:
externals/art/aerosol_dynamics/mo_art_washout_volc.f90
! Line 73
tracer(jc,jk)= tracer(jc,jk)                                &
&             - 0.0_wp * (3.*(prr_con(jc) + prr_gsp(jc)) &
&             * tracer(jc,jk)                               &
&             * dtime)
```

C. Additional plots to Softplus ablation study

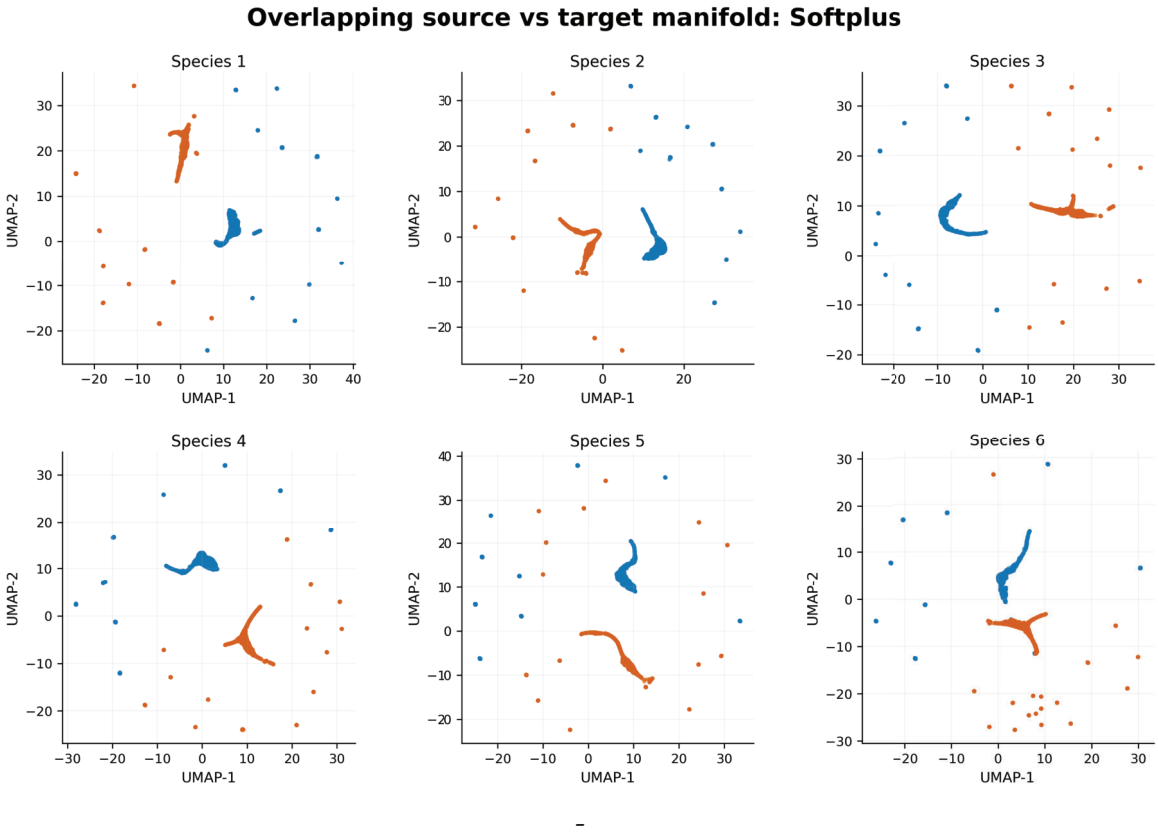

Figure C.1.: UMAP projection of the latent representations z (source, blue) and reconstructed representations ẑ (target, orange) produced by a Softplus autoencoder, shown separately for six species. Each point corresponds to a single sample embedded into a two-dimensional UMAP space. Tight, co-localised clusters indicate high-fidelity reconstruction and good manifold alignment, whereas spatial separation between source and target clouds reflects reconstruction error in the latent geometry.

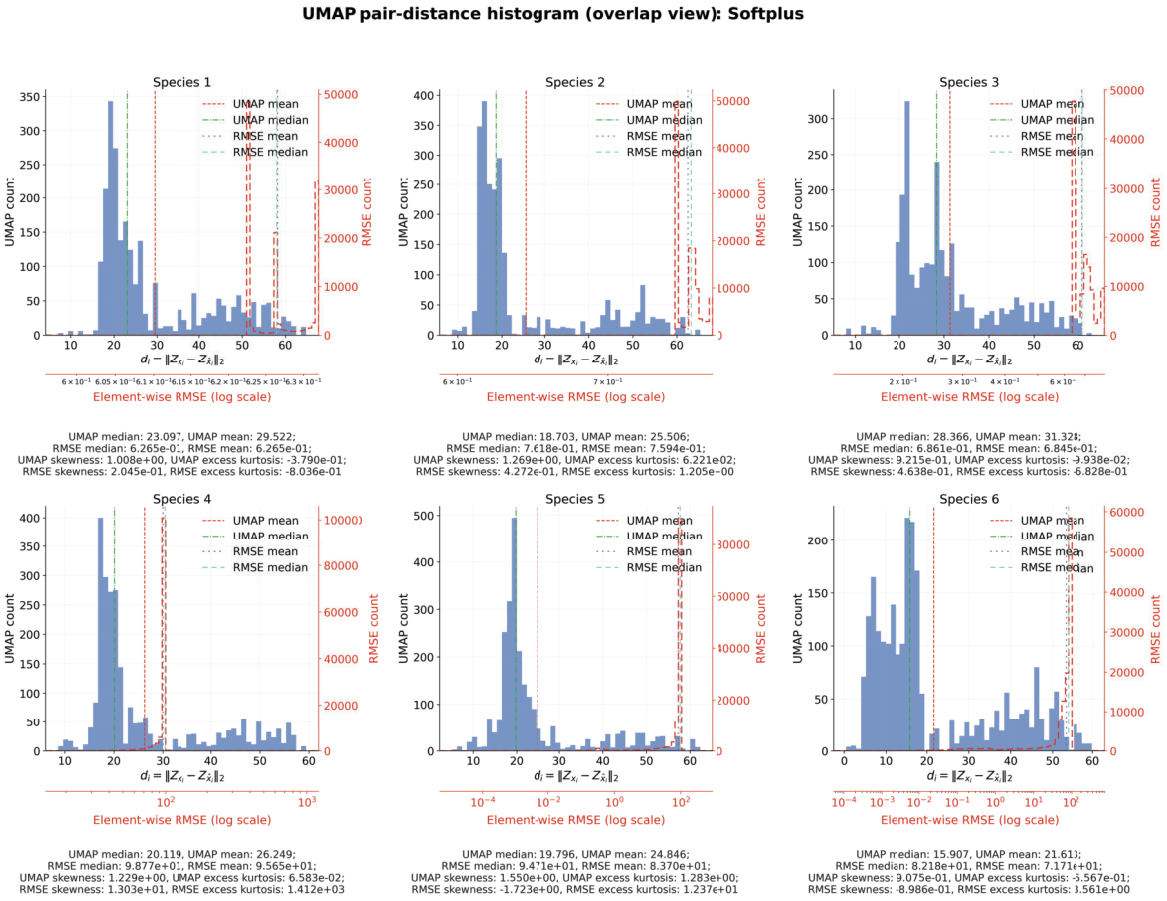

Figure C.2.: Pair-wise distance and element-wise reconstruction error distributions for the Softplus Autoencoder across six species. The left y-axis (blue histogram) shows the UMAP pair-distance $d_i = \|\mathbf{z}_{x_i} - \hat{\mathbf{z}}_{x_i}\|_2$ between source and target embeddings; the right y-axis (right-aligned ticks) shows the corresponding element-wise RMSE distribution on a logarithmic scale. Vertical dashed lines indicate the mean (red, dotted) and median (green,dashed) of the UMAP distances, and the mean (pink, dotted) and median (cyan, dashed) of the RMSE values, respectively. Summary statistics (median, mean, skewness, and excess kurtosis) for both distributions are reported below each panel. Right-skewed UMAP distance distributions across all species indicate that the majority of samples are reconstructed with low error, while a minority of outlier samples exhibit substantially larger latent displacement.

D. Loss curve of trained model

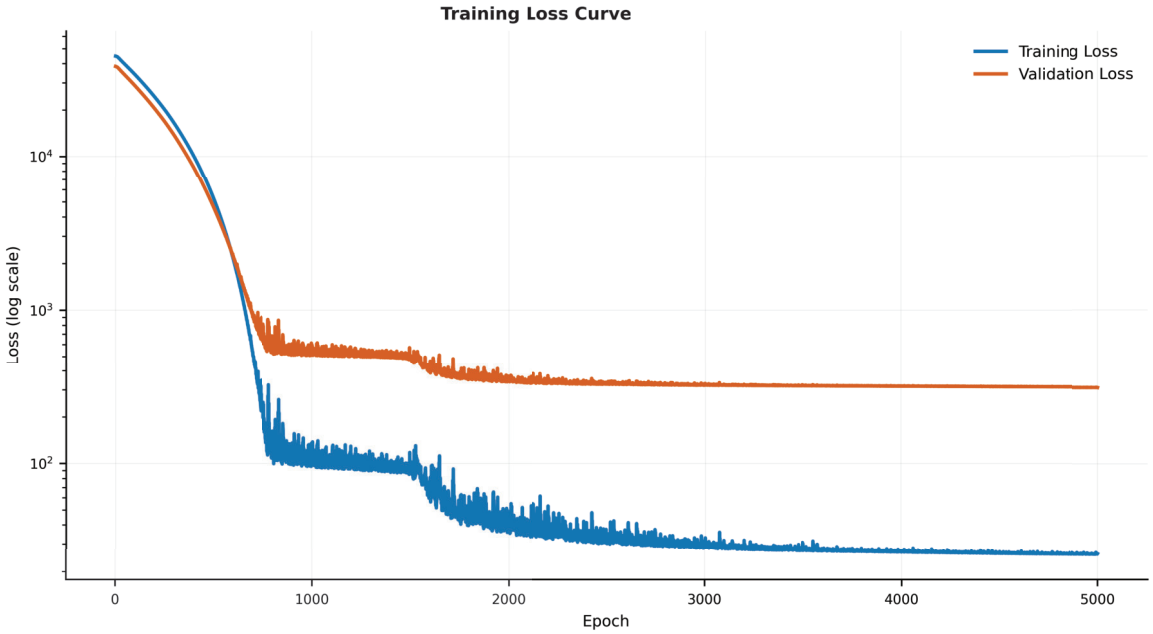

Figure D.1.: Training & Validation loss of Multi-head Autoencoder. Blue curve denotes the training loss and orange the validation loss

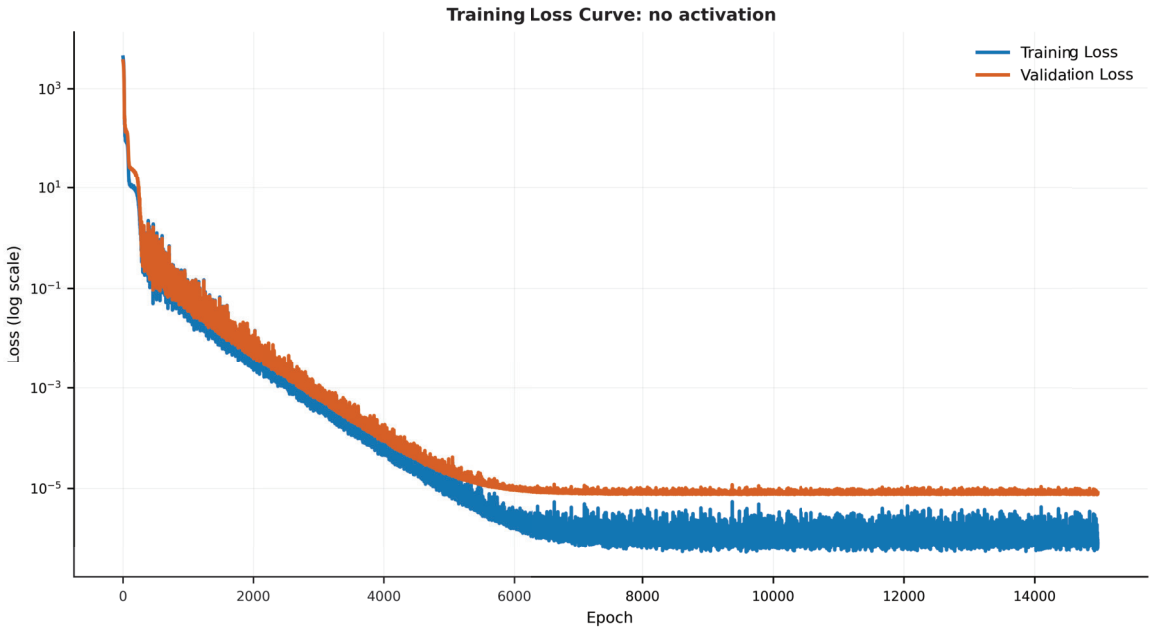

Figure D.2.: Training & validation loss of 1 layer Autoencoder with no activation function. Blue curve denotes the training loss and orange the validation loss

E. Timer for training Data

		Processes					
		Total	Total w/o ComIn	Transport (Advection)	Horizontal (including Flux Limiter)	Vertical (including Flux Limiter)	ART
Model	Full Model	320.8 CH	320.7 CH	13.6 CH (4.2%)	11.2 CH	2.1 CH	11.7 CH (3.6%)
	Source/Sink off	238.0 CH	237.9 CH	13.6 CH (5.7%)	11.2 CH	2.2 CH	12.9 CH (5.4%)
	Full Model (ART off)	198.0 CH	197.9 CH	7.1 CH (3.6%)	6.0 CH	0.94 CH	N/A

Table E.1.: Process timer for obtaining the training data. Reference run of ICON-ART using the exp.dust_rad in the ICON-NWP test suite. The simulation is done with full model enabled, source/sink turned off in ART source code and some namelist/xml options off and the full model with lart = .false. option in run_nml. Note that as source/sink off model is done by multiplying 0s in front of the concentration update, the ART model spent a bit more time executing the additional scalar multiplication. Unit in CPU hour (CH), bracket is the proportion of the CH used in respect to the total CPU timeing except ComIn.

	Full model	Full model with ComIn nudging	3 Advected species
Total	1015.4	1005.9	1073.9
ComIn call back	0.054	3.50	0.054
Transport	41.92	37.00	29.82
Horizontal Advection	35	30.35	24.54
Vertical Advection	5.90	6.00	4.18

Table E.2.: Average time taken (in seconds) for individual processes per MPI thread in a 2-day simulation. Three configurations are compared: the full model (6 advected species), the full model with ComIn-coupled ML nudging, and a reduced setup advecting only 3 species. Highlighted values (blue) denote the quantities used to estimate the computational benefit of the reduced-order embedding. Back-trajectory computation and horizontal flux limiter are sub-components of horizontal advection.