Karlsruhe Institute of Technology

Machine Learning Approaches for Composition Attribution of Ice Nucleating Particles

Master's Thesis of

Loghman Fathollahi

At the KIT Institute of Meteorology and Climate Research

First examiner:	Prof. Dr. Corinna Hoose
Second examiner:	Prof. Dr. Thomas Leisner
First advisor:	Dr. Gabriella Linnea Wallentin
Second advisor:	Dr. Alexander Böhmländer

01. 09 2025 – 01. 09 2026

Machine Learning Approaches for Composition Attribution of Ice Nucleating Particles (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, 01. 09 2026

. . . . . .

(Loghman Fathollahi)

Abstract

Ice-nucleating particles (INPs) are aerosol particles that facilitate the heterogeneous freezing of supercooled cloud droplets at higher temperatures than those required for homogeneous freezing. They influence the development of mixed-phase clouds and can thereby affect precipitation and cloud radiative properties. A freezing spectrum reports the cumulative INP concentration as a function of temperature but does not identify the aerosol populations responsible for the measured activity. Because ice residuals are often not analysed further, the measured freezing activity cannot usually be linked directly to the aerosol particles responsible for it.

This thesis investigates whether the Snomax surface fraction can be recovered from a single perturbed synthetic freezing spectrum using supervised machine learning. The problem is simplified to a mixture of mineral dust and Snomax®, which is used as a laboratory proxy for a biogenic component. The prediction target is the Snomax surface fraction, defined as the fraction of the total particle surface area contributed by Snomax. Because no suitable measured dataset with known composition for each sample was available, $100,000$ synthetic spectra were generated using published temperature-dependent ice-nucleation parameterisations for dust and Snomax. Variations in ice nucleation activity and prescribed pointwise spectrum variability were included. Only spectra in which both components exceeded a 5σ detectability threshold were retained. Features describing the spectral shape were derived to separate composition from total aerosol loading, and several regression models were compared using a common training and validation procedure.

A random forest model was selected. On an independent test dataset, it achieved a coefficient of determination of $R^2 = 0.90$ and a root mean square error of 0.018 in surface fraction, with negligible bias. About $97\ \%$ of the predictions had an absolute error below 0.05. The model required only two features: the ratio of the slopes in the higher and lower-temperature regions and the overall steepness of the spectrum. The SHAP analysis showed that these features represented spectral regions in which the modelled freezing behaviour of dust and Snomax differed.

The results show that composition attribution can be addressed as a regression problem for this simplified synthetic mixture when both components are detectable. Validation with measured laboratory mixtures is required before application to ambient aerosol.

Keywords: ice-nucleating particles, freezing spectra, composition attribution, mineral dust, Snomax®, INAS density, machine learning, random forest.

Zusammenfassung

Eisnukleierende Partikel (INPs) sind Aerosolpartikel, die das heterogene Gefrieren unterkühlter Wolkentröpfchen bei höheren Temperaturen ermöglichen, als für das homogene Gefrieren erforderlich sind. Sie beeinflussen die Entwicklung von Mischphasenwolken und können dadurch den Niederschlag und die Strahlungseigenschaften von Wolken verändern. Ein Gefrierspektrum beschreibt die kumulative INP-Konzentration als Funktion der Temperatur, identifiziert jedoch nicht die Aerosolpopulationen, die für die gemessene Aktivität verantwortlich sind. Da Eisresiduen häufig nicht weiter analysiert werden, kann die gemessene Gefrieraktivität meist nicht direkt mit den dafür verantwortlichen Aerosolpartikeln verknüpft werden.

In dieser Arbeit wird untersucht, ob der Snomax-Oberflächenanteil mithilfe von überwachtem maschinellem Lernen aus einem einzelnen verrauschten Gefrierspektrum bestimmt werden kann. Das Problem wird auf eine Mischung aus Mineralstaub und Snomax® vereinfacht, wobei Snomax als Laborersatz für eine biogene Komponente dient. Die Zielgröße ist der Snomax-Oberflächenanteil, definiert als der von Snomax stammende Anteil an der gesamten Partikeloberfläche. Da kein geeigneter Messdatensatz mit bekannter Zusammensetzung für jede Probe verfügbar war, wurden 100,000 synthetische Spektren mithilfe veröffentlichter temperaturabhängiger Parametrisierungen der Eisnukleation für Staub und Snomax erzeugt. Variationen der Eisnukleationsaktivität und eine vorgegebene punktweise Variabilität des Spektrums wurden berücksichtigt. Es wurden nur Spektren verwendet, in denen beide Komponenten eine Nachweisgrenze von 5σ überschritten. Merkmale zur Beschreibung der Spektrumsform wurden abgeleitet, um die Zusammensetzung von der gesamten Aerosolbeladung zu trennen. Mehrere Regressionsmodelle wurden mit einem gemeinsamen Trainings- und Validierungsverfahren verglichen.

Ein Random-Forest-Modell wurde ausgewählt. Auf einem unabhängigen Testdatensatz erreichte es ein Bestimmtheitsmaß von $R^2 = 0{,}90$ und eine Wurzel des mittleren quadratischen Fehlers von 0,018 für den Oberflächenanteil bei vernachlässigbarer systematischer Abweichung. Etwa 97 % der Vorhersagen hatten einen absoluten Fehler von weniger als 0,05. Das Modell benötigte nur zwei Merkmale: das Verhältnis der Steigungen in den höheren und niedrigeren Temperaturbereichen sowie die gesamte Steilheit des Spektrums. Die SHAP-Analyse zeigte, dass diese Merkmale Bereiche des Spektrums abbildeten, in denen sich das modellierte Gefrierverhalten von Staub und Snomax unterschied.

Die Ergebnisse zeigen, dass die Bestimmung der Zusammensetzung für diese vereinfachte synthetische Mischung als Regressionsproblem behandelt werden kann, wenn beide Komponenten nachweisbar sind. Vor einer Anwendung auf atmosphärische Aerosole ist eine Validierung mit gemessenen Labormischungen erforderlich.

Schlüsselwörter: eisnukleierende Partikel, Gefrierspektren, Bestimmung der Zusammensetzung, Mineralstaub, Snomax®, INAS-Dichte, maschinelles Lernen, Random Forest.

Preface

Declaration on the use of AI tools

During the preparation of this thesis, I used *ChatGPT (OpenAI), Claude (Anthropic), NotebookLM (Google), Grammarly, Elicit, and Semantic Scholar* as supporting tools.

The tools were used to identify potentially relevant publications and develop search terms; obtain preliminary summaries before studying the original publications; structure and improve the grammar, spelling, style, and readability of text that I had written; check and correct LaTeX code that I had written; and obtain technical assistance with Python debugging, code structure and refactoring, understanding compatibility changes between different versions of software packages, and Matplotlib formatting.

The tools were not used to formulate the research questions or hypotheses, determine the scientific methodology, select the machine-learning models or evaluation criteria, make scientific decisions about the data, parameter ranges, features, model selection, numerical results or diagnostics, interpret the results, draw the conclusions, or generate references for direct inclusion. All cited sources were obtained from the original publications and were personally read and verified.

No output from these tools was treated as scientific evidence or as a scientific result. All language suggestions were checked, and all code included in the final workflow was inspected, tested, and corrected where necessary. The datasets, results, and figures were produced and validated through the documented workflow. I remain fully responsible for the scientific content, methods, analysis, interpretation, and conclusions of this thesis.

Contents

Abstract i

Zusammenfassung iii

Preface v

List of Symbols 1

List of Abbreviations 7

Terminology Used in This Thesis 9

1 Introduction 13
- 1.1 Motivation . . . 13
- 1.2 Inferring aerosol composition from freezing spectra . . . 14
- 1.3 Aim and scope . . . 15
- 1.4 Research questions . . . 16
- 1.5 Structure of the thesis . . . 16

2 Theoretical Background 17
- 2.1 Fundamentals of Ice-Nucleating Particles . . . 17
- 2.2 Importance of INPs for Clouds and Climate . . . 17
- 2.3 Measurement of INPs and Sources of Variability . . . 18
- 2.4 Temperature Dependence and INP Types Relevant to This Work . . . 19
- 2.5 The Singular Description and the INAS Framework . . . 19
- 2.6 Lognormal Aerosol Size Distributions and Surface Area . . . 21
- 2.7 INAS Parameterisations Used in This Work . . . 22
- 2.8 The INP Composition-Attribution Problem . . . 23
- 2.9 Machine Learning Background . . . 24

3 Methodology 33
- 3.1 Overview of the modelling pipeline . . . 33
- 3.2 Ice nucleation active surface site density parameterisations . . . 35
- 3.3 Mixed aerosol populations and composition recovery . . . 38
- 3.4 Surface-area concentration ensemble . . . 39
- 3.5 Synthetic mixed INP spectra . . . 45
- 3.6 Detectability and regime classification . . . 51
- 3.7 Feature engineering . . . 55

3.8 Feature selection . . . 58
3.9 Dataset partitioning . . . 61
3.10 Regression models and training protocol . . . 61
3.11 Model selection . . . 64

4 Results **69**
4.1 The generated ensemble . . . 69
4.2 Detectability and the retained dataset . . . 71
4.3 Feature selection . . . 75
4.4 Feature interpretation and the number of features . . . 81
4.5 Model comparison and selection . . . 82
4.6 Final model . . . 83
4.7 Summary of the results . . . 86

5 Conclusions and Outlook **89**
5.1 Answers to the research questions . . . 89
5.2 Limitations . . . 90
5.3 Outlook . . . 92
5.4 Concluding remarks . . . 93

Acknowledgements **95**

Bibliography **97**

List of Figures

3.1 Overview of the modelling pipeline 34
3.2 Unperturbed fitted INAS curves of the two species 37
3.3 Construction of the surface-area concentration ensemble 41
3.4 Two-dimensional Latin hypercube sampling 42
3.5 Representative lognormal surface-area distributions 44
3.6 Dust INAS measurements and fitted parameterisation 46
3.7 Decomposition of the INAS parameterisation perturbation 48
3.8 The two perturbation layers 50
3.9 Detection statistics of both species 53
3.10 Model-selection procedure 65

4.1 Realised activation curves 70
4.2 Dust contribution and the dust-detectability threshold 72
4.3 Dust contribution around the detection limit 72
4.4 Regime phase diagram 74
4.5 Representative spectra per regime 75
4.6 Learning curve 76
4.7 Feature correlation before and after pruning 77
4.8 SHAP importance of the surviving features 79
4.9 Robustness of the importance ranking 80
4.10 Cross-validation performance against the number of features 82
4.11 Model selection funnel 84
4.12 Cross-validation and test RMSE near the selection band 84
4.13 Predictions and residuals on the test set 87
4.14 Error against composition 88
4.15 Feature importance of the selected model 88

List of Tables

3.1 Parameters used for dust and Snomax 44
3.2 Detectability-regime criteria 52
3.3 Engineered spectral features 59
3.4 Labels of the shape-intrinsic features 60
3.5 Regression models and hyperparameter grids 63
3.6 Study-specific simplicity ranking used for model selection. 67

4.1 Detectability regimes of the ensemble 73
4.2 Correlation pruning decisions 76
4.3 SHAP ranking of the surviving features 78
4.4 Model comparison and selection outcome 83
4.5 Tuned hyperparameters of the final model 85
4.6 Final model performance by split 85

List of Symbols

Latin symbols

Symbol	Description	Unit
$\bar{a}$	Mean particle surface area of a particle population	m^2
$\bar{a}_x$, $\bar{a}_{x,i}$	Mean particle surface area of species x or of species x in sample i	m^2
$\bar{a}_x^{\mathrm{ref}}$	Mean particle surface area of species x at the reference diameter	m^2
$\bar{a}_{\max}$	Mean particle surface area at the upper diameter bound	m^2
$\mathbf{A}$	Matrix containing the dust and Snomax reference spectra in the NNLS decomposition	–
$a_{x,i}$	Temperature-independent log-amplitude perturbation for species x in sample i	–
$b_{x,i}$	Temperature-dependence perturbation for species x in sample i	K^{-1}
$\boldsymbol{b}^{(\ell)}$	Bias vector of neural-network layer ℓ	–
c_{aer}	Aerosol particle number concentration, $c_{\mathrm{aer}} = S/\bar{a}$	m^{-3}
$c_{\mathrm{INP}}(T)$	Cumulative INP number concentration at temperature T	m^{-3}
$c_x(T)$, $c(T)$	Contribution of species x and total mixed INP concentration	m^{-3}
$c^{(i)}(T)$	Mixed spectrum of sample i after the INAS parameterisation perturbation and before the pointwise perturbation	m^{-3}
$\tilde{c}^{(i)}(T)$	Synthetic observed spectrum of sample i after the pointwise spectrum perturbation	m^{-3}
$c_x^{\mathrm{ref}}(T)$	Single-species reference spectrum calculated from the unperturbed fitted parameterisation	m^{-3}
$\mathcal{D}$	Supervised-learning dataset of input–target pairs	–
d	Snomax particle diameter used in the mass-to-surface conversion	m
d_{lo}, d_{hi}	Lower and upper bounds of the sampled number-median-diameter range	m

Symbol	Description	Unit
d_p	Particle diameter in the lognormal number-size distribution	m
d_{pm}	Number-median diameter, used in this study as a numerical control variable for particle loading	m
$d_{pm,i}$	Sampled number-median diameter of sample i	m
d_{pm2}	Median diameter of the surface-area-size distribution	m
d_{pm}^{ref}	Reference number-median diameter	m
$f(\boldsymbol{x};\boldsymbol{\theta})$	Regression model with input $\boldsymbol{x}$ and parameters $\boldsymbol{\theta}$	–
f_m	Snomax mass-to-surface conversion factor, $f_m = \rho d/6$	$\mathrm{mg\,m^{-2}}$
f_{slope}	Amplification factor used for the prescribed high-temperature dust decay	–
$F_M(\boldsymbol{x})$	Gradient-boosting ensemble after M steps	–
$\mathcal{F}$	Set of all input features in the SHAP formulation	–
$g_x(T)$	Fraction of the mixed spectrum contributed by species x	–
$g_{\mathrm{dust}}^{\mathrm{lin}}(T;R)$	Dust fraction under the dilute-activation approximation	–
$H[p]$	Shannon entropy of the normalised freezing-increment distribution	–
$\boldsymbol{h}^{(\ell)}$	Output vector of neural-network layer ℓ	–
I_j	Impurity-based importance of feature j	–
I_j^{SHAP}	Mean absolute SHAP importance of feature j	–
$\Delta i(t)$	Reduction in impurity produced by the split at tree node t	–
$K(T)$	Cumulative nucleus spectrum per unit liquid volume in the Vali framework	$\mathrm{m^{-3}}$
$k(T)$	Density of ice-active sites becoming active per unit temperature	$\mathrm{m^{-2}\,K^{-1}}$
k	Detectability threshold of the matched-filter statistic, set to 5	–
k_{decay}	Decay coefficient of the prescribed high-temperature dust extension	$\mathrm{K^{-1}}$
K_m	Number of cross-validation folds used for model m	–
$\mathcal{L}$	Empirical loss function	–
M	Number of trees or boosting steps in an ensemble	–
m	Snomax mass concentration associated with surface area concentration S	$\mathrm{kg\,m^{-3}}$
m^*	Model with the lowest mean cross-validation RMSE	–
N	Number of samples in a regression dataset	–
N_0	Total particle number concentration of a lognormal population	$\mathrm{m^{-3}}$
$N_{\mathrm{ice}}(T)$	Cumulative number of frozen drops at temperature T	–
N_{tot}	Total number of drops examined	–

Symbol	Description	Unit
N_s	Number of generated synthetic samples	–
$\mathcal{N}_k(\boldsymbol{x})$	Set of the k nearest training samples to $\boldsymbol{x}$	–
$\bar{n}_{s,\mathrm{dust}}(T)$	Unperturbed fitted dust INAS density	m^{-2}
$\bar{n}_{m,\mathrm{snomax}}(T)$	Unperturbed fitted mass-based Snomax ice nucleation activity	mg^{-1}
$n^{(i)}_{s,\mathrm{dust}}(T)$	Parameterisation-perturbed dust INAS density for sample i	m^{-2}
$n^{(i)}_{m,\mathrm{snomax}}(T)$	Parameterisation-perturbed mass-based Snomax ice nucleation activity for sample i	mg^{-1}
n_p	Particle number concentration represented by the lognormal number-size distribution	m^{-3}
$n_s(T)$	Ice-nucleation active surface-site density	m^{-2}
$n_{s,\mathrm{dust}}(T)$, $n_{s,\mathrm{snomax}}(T)$	Dust INAS density and surface-area-equivalent Snomax INAS density	m^{-2}
$n_{m,\mathrm{snomax}}(T_C)$	Mass-based Snomax ice nucleation activity	mg^{-1}
n_{ref}	Dust INAS density at T_{high}	m^{-2}
p	Number of input features; also the Minkowski exponent in the k-NN distance metric	–
$p(T)$	Normalised freezing-increment weight on the temperature grid	–
$p(t)$	Proportion of training samples reaching tree node t	–
$q_x(T)$	INP response of species x per unit particle surface-area concentration	m^{-2}
R	Snomax surface fraction and regression target; denoted `ssf` in code and output files	–
R_i	Snomax surface fraction of sample i	–
R_{min}, R_{max}	Lower and upper bounds of the sampled Snomax surface fraction	–
R^*	Dust-detectability threshold in Snomax surface fraction	–
R^2	Coefficient of determination	–
R^2_{CV}, $R^2_{\mathrm{train,CV}}$	Mean held-out-fold and mean training-fold coefficients of determination	–
r	Pearson correlation coefficient used for feature pruning	–
r_{NNLS}	Relative residual of the NNLS reference-shape reconstruction	–
S	Total particle surface area concentration	$\mathrm{m}^2\,\mathrm{m}^{-3}$
$S_{x,i}$	Surface area concentration of species x in sample i	$\mathrm{m}^2\,\mathrm{m}^{-3}$
$S_{\mathrm{tot},i}$	Total particle surface area concentration of sample i	$\mathrm{m}^2\,\mathrm{m}^{-3}$
S_x^{ref}	Reference surface area concentration of species x	$\mathrm{m}^2\,\mathrm{m}^{-3}$
$\mathcal{S}$	Subset of input features in the SHAP formulation	–

Symbol	Description	Unit
s_m	Sample standard deviation of the fold RMSE values for model m	–
SE_m	Standard error of the mean cross-validation RMSE for model m	–
$\overline{\mathrm{RMSE}}_m$	Mean cross-validation RMSE of model m	–
SNR_x	Matched-filter detection statistic for species x	–
T	Absolute temperature	K
T_C	Temperature in degrees Celsius	°C
T_0	Melting-point reference temperature in the cumulative INAS relation	K
T_{high}	Upper temperature limit of the fitted dust parameterisation, 259 K	K
T_{ref}	Reference temperature of the INAS parameterisation perturbation, 253.15 K (−20 °C)	K
ΔT	Temperature interval; in the dust extension, $T - T_{\mathrm{high}}$	K
T_{10}, T_{50}, T_{90}	10th, 50th and 90th quantile temperatures of $p(T)$	°C
$T_m(\boldsymbol{x})$	Prediction of tree m for input $\boldsymbol{x}$	–
$u_{i,1}$, $u_{i,2}$	Latin-hypercube unit coordinates of sample i	–
V	Drop volume in the Vali framework	m^3
$v(t)$	Feature used to split tree node t	–
$v_{\boldsymbol{x}}(\mathcal{S})$	Expected model output when feature subset $\mathcal{S}$ is known for sample $\boldsymbol{x}$	–
$\boldsymbol{W}^{(\ell)}$	Weight matrix of neural-network layer ℓ	–
$w_i(\boldsymbol{x})$	Weight assigned to neighbour i in k-NN regression	–
$\boldsymbol{x}_i$, x_{ij}	Input-feature vector of sample i and its value for feature j	–
y_i, $\hat{y}_i$	True target and model prediction for sample i	–
$\bar{y}$	Mean true target value	–

Greek symbols

Symbol	Description	Unit
α	Regularisation-strength hyperparameter used by selected regression models	–
$\boldsymbol{\beta}$, β_0, β_j	Regression-coefficient vector, intercept and coefficient of feature j	–
$\boldsymbol{\gamma}^*$	Non-negative coefficient vector fitted in the NNLS decomposition	–
γ_m	Contribution weight of tree m in gradient boosting	–

Symbol	Description	Unit
$\delta_x(T)$	Log-space change caused by removing species x from the mixed spectrum	–
$\varepsilon_i(T)$	Independent log-space draw of the pointwise spectrum perturbation	–
η	Elastic-net mixing parameter between the ℓ_2 and ℓ_1 penalties	–
θ	Angle between two vectors in the cosine similarity	–
$\boldsymbol{\theta}$	Parameter vector of a regression model	–
λ	Overall regularisation strength in regularised regression	–
ν	Learning rate in gradient boosting	–
ρ	Bulk density of Snomax used in the mass-to-surface conversion	$\mathrm{kg\,m^{-3}}$
$\sigma_{a,x}$	Standard deviation of the log-amplitude perturbation for species x	–
$\sigma_{b,x}$	Standard deviation of the temperature-dependence perturbation for species x	$\mathrm{K^{-1}}$
σ_g	Geometric standard deviation of the lognormal particle size distribution	–
σ_ℓ	Activation function of neural-network layer ℓ	–
$\sigma_{\log}$	Log-space standard deviation of the pointwise spectrum perturbation	–
ϕ_0	SHAP reference value, equal to the expected model prediction	–
$\phi_j(\boldsymbol{x})$	SHAP contribution of feature j to the prediction for input $\boldsymbol{x}$	–

Operators and statistical notation

Symbol	Description
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2
$\mathbb{E}[\,\cdot\,]$	Expectation value
$\arg\min$	Argument that minimises an objective function
$\lVert\cdot\rVert_2$	Euclidean (ℓ_2) norm
$\lvert\cdot\rvert$	Absolute value
ℓ_1, ℓ_2	Absolute-value and squared-coefficient penalties or norms
$\cos\theta(\mathbf{a}, \mathbf{b})$	Cosine similarity between vectors $\mathbf{a}$ and $\mathbf{b}$
$\bar{\cdot}$	Overbar: mean quantity or unperturbed fitted parameterisation, according to context
$\hat{\cdot}$	Hat: estimated parameter or predicted quantity

Symbol	Description
$\tilde{\cdot}$	Tilde: synthetic spectrum after the pointwise spectrum perturbation
$(\cdot)^{(i)}$	Superscript i: quantity belonging to synthetic sample i
$(\cdot)^{\mathrm{ref}}$	Superscript ref: reference quantity evaluated at the reference diameter

List of Abbreviations

Abbreviation	Meaning
AI	Artificial intelligence
AIDA	Aerosol Interaction and Dynamics in the Atmosphere cloud chamber
CV	Cross-validation
DNN	Deep neural network
IMK-AAF	Institute of Meteorology and Climate Research–Atmospheric Aerosol Research
IMK-TRO	Institute of Meteorology and Climate Research–Troposphere Research
INAS	Ice-nucleation active surface-site density
INP	Ice-nucleating particle
KIT	Karlsruhe Institute of Technology
k-NN	k-nearest neighbours
LHS	Latin-hypercube sampling
LightGBM	Light Gradient Boosting Machine
MAE	Mean absolute error
MaxErr	Maximum absolute error
MLP	Multilayer perceptron
NNLS	Non-negative least squares
ReLU	Rectified linear unit
RMSE	Root mean square error
SD	Standard deviation
SE	Standard error
SHAP	SHapley Additive exPlanations
SNR	Signal-to-noise ratio
ssf	Snomax surface fraction; name of the target R in code and output files
TreeSHAP	Exact SHAP algorithm for tree-based models
XGBoost	Extreme Gradient Boosting

Terminology Used in This Thesis

Term	Meaning
Spectrum generation and perturbations	
Synthetic spectrum generation framework	The complete computational procedure that combines the particle population, composition, INAS parameterisations and perturbations to generate mixed synthetic freezing spectra. This term is used instead of *forward model*.
Unperturbed fitted INAS parameterisation	The published fitted temperature-dependent relationship before any sample-specific perturbation is applied. It does not represent an arithmetic mean over the perturbed ensemble.
INAS parameterisation perturbation	The coherent per-sample modification of the amplitude and temperature dependence of an entire INAS curve. It is applied separately to dust and Snomax.
Parameterisation-perturbed spectrum	The mixed spectrum after applying the species-specific INAS parameterisation perturbations but before applying the pointwise spectrum perturbation. Detectability is evaluated using this spectrum.
Pointwise spectrum perturbation	The independent multiplicative factor applied at each temperature after the dust and Snomax contributions have been combined.
Prescribed pointwise spectrum variability	The assumed log-space variability level, represented by $\sigma_{\log}$, from which the pointwise spectrum perturbation is generated. It is a prescribed synthetic modelling choice and not measured instrumental uncertainty.
Synthetic observed spectrum	The final synthetic spectrum after both the INAS parameterisation perturbation and the pointwise spectrum perturbation have been applied. This spectrum is used for feature extraction.
Perturbed synthetic freezing spectrum	A general description of a synthetic freezing spectrum containing both perturbation layers.
Spectrum without the pointwise perturbation	A spectrum to which the pointwise multiplicative factor has not been applied. The INAS parameterisation perturbation may still be present and is therefore specified separately when necessary.

Term	Meaning
Single-species reference spectrum	A fixed pure-dust or pure-Snomax spectrum calculated from the unperturbed fitted parameterisation at the reference surface-area concentration, without either perturbation.
Prescribed high-temperature dust extension	The exponential continuation applied above 259 K to define the dust INAS density over the complete common temperature grid. It is a modelling choice of this study and is not part of the published dust parameterisation.
Composition, loading and spectral features	
Composition attribution	Estimation of the relative dust and Snomax contributions from the shape of a mixed freezing spectrum.
Snomax surface fraction	The fraction R of the total particle surface area contributed by Snomax. It is the regression target and is not a particle-number or mass fraction.
Total particle surface-area concentration	The total surface-area concentration $S = S_{\mathrm{dust}} + S_{\mathrm{snomax}}$. It represents the total particle loading in this study.
Component particle surface-area concentration	The surface-area concentration assigned to one component, S_{dust} or S_{snomax}.
Surface-area concentration ensemble	The collection of particle populations generated by sampling the Snomax surface fraction R and the number-median diameter d_{pm}.
Numerical control variable d_{pm}	The sampled number-median diameter used to control the total surface-area concentration S and mean particle surface area $\bar{a}$. It is not claimed to represent the physical particle size of dust or Snomax.
Surface-area-equivalent Snomax INAS density	The published mass-based Snomax ice-nucleation activity converted to a surface-area basis so that it can be combined with the dust contribution.
Dilute-activation approximation	The approximation $n_s(T)\bar{a} \ll 1$, under which $1 - \exp[-n_s(T)\bar{a}] \approx n_s(T)\bar{a}$ and $c_{\mathrm{INP}}(T) \approx S\, n_s(T)$.
Partial activation	The condition in which the activation probability remains below saturation and the spectrum retains sensitivity to the INAS parameterisation.
Activation saturation	The condition in which the activation probability approaches unity and the INP concentration approaches the particle number concentration $S/\bar{a}$.
Scale-invariant spectral feature	A feature that is unchanged when the complete spectrum is multiplied by a positive constant. This invariance applies to purely multiplicative scaling and not necessarily to shape changes caused by partial saturation.

Term	Meaning
Shape-intrinsic feature	A feature calculated directly from one spectrum without requiring the exact pure-dust or pure-Snomax reference spectra.
Reference-shape feature	A feature calculated by comparing a spectrum with the exact pure-species reference spectra. These features are excluded from model training.
Conceptual leakage	Unrealistically direct access to composition information through the exact reference curves used to generate the synthetic spectra.
Detectability terminology	
Species detection statistic	The statistic SNR_x, which combines the standardised logarithmic contribution of species x across the complete temperature grid.
Detectability criterion	A species is classified as detectable when $\mathrm{SNR}_x \geq k$. The value $k = 5$ is a conservative study-specific modelling choice.
Dust-detectability threshold	The value $R^* = 0.0701$, defined as the largest Snomax surface fraction for which dust satisfies the adopted detectability criterion. It is calculated using the unperturbed fitted parameterisations and the dilute-activation approximation.
Detectability regime	A category assigned according to whether the dust and Snomax contributions individually satisfy the adopted detectability criterion.
Mixed regime	The regime in which both dust and Snomax satisfy the detectability criterion. Only this regime is retained for model training and evaluation.
Dust-dominated regime	The regime in which dust satisfies the detectability criterion but Snomax does not. Snomax is nevertheless present in the generated mixture.
Snomax-dominated regime	The regime in which Snomax satisfies the detectability criterion but dust does not. Dust is nevertheless present in the generated mixture.
Undetectable regime	The regime in which neither component satisfies the adopted detectability criterion. This classification does not represent an absolute instrumental detection limit.
Conditional model performance	Model performance evaluated only for mixed-regime spectra under the adopted $k = 5$ criterion. It does not represent performance over the complete synthetic ensemble or unrestricted atmospheric measurements.
Uncertainty and temperature terminology	

Term	Meaning
Measurement uncertainty	Uncertainty associated with measured data. This term is not used to describe the prescribed synthetic pointwise spectrum perturbation.
Prediction error	The difference between the predicted and true Snomax surface fractions.
Higher-temperature region	The part of the spectrum at relatively higher temperatures. For the slope features, the division between the two temperature regions is −15 °C.
Lower-temperature region	The part of the spectrum at relatively lower temperatures.
Higher-temperature Snomax shoulder	The spectral region in which the contribution from Snomax is particularly visible.
Lower-temperature dust branch	The spectral region in which the increasing dust contribution is particularly visible.

1 Introduction

1.1 Motivation

Clouds influence both the atmospheric radiation budget and the hydrological cycle. Their effects depend partly on cloud phase because liquid droplets and ice particles differ in their optical properties, growth processes, and sedimentation rates. Mixed-phase clouds contain supercooled liquid droplets and ice particles at the same time. The partitioning of water between these two phases influences precipitation formation, cloud lifetime, and cloud radiative properties (Murray et al. 2012; Burrows et al. 2022; Knopf and Alpert 2023). A correct representation of ice formation in mixed-phase clouds is therefore important for weather and climate models.

The first ice particles in a cloud are formed through primary ice production. In pure water droplets of micrometre size, homogeneous freezing usually becomes important only at temperatures below approximately −38 °C (Hoose and Möhler 2012). At higher temperatures, the freezing of supercooled droplets generally requires a foreign particle that supports the formation of an ice embryo. Following the terminology of Vali et al. (Vali et al. 2015), such a particle is called an ice-nucleating particle (INP). Several modes of heterogeneous ice nucleation exist. This thesis focuses on immersion freezing, in which an INP located inside a supercooled liquid droplet initiates freezing (Hoose and Möhler 2012; Vali et al. 2015; Kanji et al. 2017).

The formation of primary ice does not mean that the whole cloud immediately changes to the ice phase. The later development depends on temperature, humidity, cloud dynamics, liquid water content, and interactions between hydrometeors. Under suitable conditions, ice crystals may grow by vapour deposition while liquid droplets evaporate through the Wegener–Bergeron–Findeisen process (Korolev and Isaac 2003; Murray et al. 2012). Riming, aggregation, sedimentation, and direct droplet freezing can also change the amount and distribution of liquid water and ice. In addition to primary ice production, new ice particles can form through secondary ice production, for example by rime splintering, fragmentation during droplet freezing, and fragmentation caused by ice–ice collisions. Secondary ice production can therefore increase the ice crystal concentration beyond that produced directly by INPs (Georgakaki et al. 2022; Korolev et al. 2022). Primary and secondary ice processes together influence cloud phase, precipitation, and radiative properties.

Atmospheric INPs originate from several aerosol classes. These include mineral dust, marine aerosol, biological material such as bacteria, fungal spores, pollen, and biological macromolecules, and some combustion particles such as soot. Their relative importance

depends on temperature, location, and atmospheric conditions (Hoose and Möhler 2012; Kanji et al. 2017; Burrows et al. 2022). This thesis focuses on desert dust and Snomax® as two reference materials with clearly different freezing activities.

The ability of an INP to initiate freezing depends on its composition, size, surface properties, and temperature (DeMott et al. 2010; Kanji et al. 2017; Burrows et al. 2022). Mineral dust is an important atmospheric INP type because it is abundant and can initiate ice formation over a broad temperature range. In immersion freezing, feldspar minerals, particularly K-feldspar, can be highly active for ice nucleation (Atkinson et al. 2013). Mineral particles may also nucleate ice through other pathways, including deposition nucleation (Yakobi-Hancock et al. 2013), but these pathways are outside the scope of the present work.

Some biological INPs can initiate freezing at higher temperatures than many mineral particles. Among these, strains of *Pseudomonas syringae* contain proteinaceous ice-nucleating material in their outer membrane (Orser et al. 1985). Snomax, a commercial product derived from non-viable *Pseudomonas syringae*, is commonly used as a biological reference material in laboratory studies of ice nucleation (Wex et al. 2015; Whale et al. 2015). The selected dust and Snomax parameterisations have different temperature dependences of ice-nucleating activity. Consequently, their relative contributions can influence the shape of a freezing spectrum.

Freezing measurements provide a spectrum of ice-nucleating activity rather than direct identification of the particles responsible for that activity. This limitation is also relevant to atmospheric models. A global model simulation underestimated measured INP concentrations at higher freezing temperatures, suggesting that an INP source may be missing or that the properties assigned to a represented source are inaccurate (Herbert et al. 2025). Better constraints on the relative contributions of different INP types could therefore support the development and evaluation of INP parameterisations. This thesis addresses a controlled version of this problem by investigating whether the Snomax surface fraction in a mixture of dust and Snomax can be inferred from a perturbed synthetic immersion-freezing spectrum.

1.2 Inferring aerosol composition from freezing spectra

An immersion freezing measurement provides information about the number of ice-nucleating particles in a sample, but it does not directly identify their composition. In a droplet freezing assay, a known volume of a particle suspension is divided into many droplets. The droplets are cooled, and the temperature at which each droplet freezes is recorded. The frozen fraction at each temperature is used to calculate the cumulative INP concentration, $c_{\mathrm{INP}}(T)$, which represents the concentration of particles that are active for ice nucleation at or above that temperature (Vali 1971; Murray et al. 2012; Herbert et al. 2014). If the total particle surface area is known, the concentration can also be expressed as the ice nucleation active surface site (INAS) density, $n_s(T)$, which represents the cumulative density of sites that are

ice active at or above temperature T per unit particle surface area(Hoose and Möhler 2012; Murray et al. 2012).

For a sample containing only one known material, the freezing spectrum can be compared with the characteristic activity of that material. The interpretation is more difficult when the sample contains several INP types. For the external mixture considered here, the contributions of dust and Snomax are calculated independently using their respective surface areas and are then added. Because their activities change differently with temperature, changing their relative contributions also changes the shape of the mixed spectrum. The exact mixing relation is derived in Chapter 2.

When the component spectra and their relative surface areas are known, the mixed spectrum can be calculated using the synthetic spectrum generation framework. This thesis considers the inverse problem, in which only the mixed spectrum is available and the relative composition must be estimated. The target variable is the Snomax surface fraction, denoted by R, which is the fraction of the total particle surface area contributed by Snomax.

In practice, this inference is affected by measurement uncertainty and natural variability in ice-nucleating activity. Because INP concentrations can change steeply with temperature, even a small uncertainty in the measured freezing temperature can produce a large uncertainty in the derived INP concentration (Hiranuma et al. 2015; Schrod and Bingemer 2025). In addition, the signal of a component with a small surface fraction may be hidden by the stronger contribution of the other component. This makes retrieval difficult when one component does not leave a detectable signal in the mixed spectrum.

A supervised retrieval method requires many spectra paired with known compositions. Such paired datasets are difficult to obtain because the measurement capabilities required for INP closure and composition attribution are not all available in a single method (Burrows et al. 2022). Synthetic spectra can provide known composition labels while allowing the INAS parameterisation perturbation and the pointwise spectrum perturbation to be prescribed and controlled. They can therefore be used to develop and evaluate the retrieval method before future tests with measured spectra of known mixtures.

1.3 Aim and scope

The central aim of this thesis is to test whether the Snomax surface fraction can be inferred from the shape of an immersion freezing spectrum. To address this problem, published activity parameterisations for desert dust and Snomax are expressed on a common INAS basis (Ullrich et al. 2017; Wex et al. 2015). They are used to generate synthetic freezing spectra for mixtures with known Snomax surface fractions. Machine learning regression models are then trained to estimate this fraction from the shape of each spectrum.

The study is a controlled proof of concept. It is limited to two externally mixed materials whose contributions are assumed to be independent. The analysis uses only synthetic spectra and does not attempt to identify all possible atmospheric INP species or retrieve

the composition of measured freezing spectra. Its purpose is to test whether the proposed approach is feasible in a controlled setting before future tests with measured spectra of known mixtures.

1.4 Research questions

The work is organised around the following questions.

1. Can a machine learning model recover the Snomax surface fraction, a continuous composition parameter, from a single perturbed synthetic INP spectrum of a mixture of dust and Snomax components, and with what accuracy?
2. Which machine learning model gives the best result for this task, and by which criteria should it be chosen among the candidates?
3. What detectability criterion determines when either component's contribution becomes too weak to detect relative to the prescribed pointwise spectrum variability, and how does the resulting threshold restrict the composition range over which the retrieval is meaningful?

1.5 Structure of the thesis

Chapter 2 presents the theoretical background: the fundamentals of heterogeneous ice nucleation, the importance of INPs for clouds and climate, the singular description and the surface site formulation used in the synthetic spectrum generation framework, the two parameterisations used here together with the normalisation issue raised in Section 1.3, the additive mixing of an external mixture, and the measurement variability that provides context for the idealised perturbations used in this study.

Chapter 3 describes the methodology as an end-to-end pipeline: the computation of the two parameterisations on a common temperature grid, the generation of the surface area ensemble, the synthesis of the mixed spectra including the stochastic parameterisation variability and the prescribed pointwise spectrum perturbation, the detectability classification and the restriction of the dataset, the feature engineering and feature selection, the partitioning of the dataset, the training protocol of the regression models, and the model selection rule.

Chapter 4 presents and discusses the results, following the same order as the pipeline, from the properties of the synthetic ensemble and the detectability regimes through the feature selection and interpretation to the comparison of the candidate models and the performance of the final model.

The closing chapter summarises the findings with respect to the research questions above, states the limitations which follow from the synthetic setting, and outlines the steps required to transfer the workflow to measured spectra.

2 Theoretical Background

2.1 Fundamentals of Ice-Nucleating Particles

Ice formation in the atmosphere proceeds along two routes. Below 0 °C, ice is thermodynamically more stable than liquid water, but a water droplet can remain in a metastable, supercooled state. Freezing begins only when an ice cluster reaches a critical size. Forming the ice-water interface increases the Gibbs free energy, whereas forming the bulk ice phase decreases it; the competition between these contributions creates the nucleation energy barrier (Murray et al. 2012; Knopf and Alpert 2023). In the absence of any foreign particle, a supercooled water droplet freezes homogeneously, but only once it is cooled below about −38 °C (Hoose and Möhler 2012; Knopf and Alpert 2023). At higher temperatures a suitable foreign surface reduces the energy barrier; this route is heterogeneous ice nucleation, and the particles that provide the surface are the ice-nucleating particles (INPs) (Murray et al. 2012; Kanji et al. 2017; Knopf and Alpert 2023).

INPs are only a very small subset of the atmospheric aerosol. At −30 °C the ice-active fraction is of the order of 10^{-5} of all particles, and it becomes very much smaller as the temperature rises towards the melting point (Canzi et al. 2025; Burrows et al. 2022; DeMott et al. 2010; Knopf and Alpert 2023). Which particles belong to this subset is governed mainly by their chemical composition, their size, and the state of their surfaces (Knopf and Alpert 2023; Burrows et al. 2022).

Heterogeneous nucleation is conventionally divided into four modes: immersion freezing, deposition nucleation, contact freezing, and condensation freezing (Huang et al. 2021; Vali et al. 2015).

Immersion freezing, in which an INP already suspended inside a supercooled droplet triggers the transition, is regarded as the dominant route to primary ice in mixed-phase clouds (Huang et al. 2021; Knopf and Alpert 2023; Kanji et al. 2017).

2.2 Importance of INPs for Clouds and Climate

Once ice is present, under suitable conditions the crystals grow at the expense of the surrounding supercooled droplets, because the equilibrium vapour pressure over ice is lower than over liquid water. This is the Wegener–Bergeron–Findeisen process, and it can contribute to precipitation formation and shorten the lifetime of the cloud (Korolev and Isaac 2003; Murray et al. 2012). Riming, aggregation, sedimentation, and direct droplet

freezing can also change the amount and distribution of liquid water and ice. In addition, existing ice particles can produce new ice particles through secondary ice production, for example through rime splintering, fragmentation during droplet freezing, and fragmentation caused by ice–ice collisions. Secondary ice production can therefore increase the ice crystal concentration beyond the concentration produced directly by INPs (Georgakaki et al. 2022; Korolev et al. 2022).

By affecting the partitioning between the liquid and ice phases, ice nucleation also alters the optical properties of the cloud and hence the radiative forcing at the top of the atmosphere (Murray et al. 2012; Knopf and Alpert 2023). Heterogeneous ice nucleation consequently remains among the least well constrained processes in atmospheric physics and a leading source of uncertainty in cloud and climate models (Murray et al. 2012; Burrows et al. 2022).

2.3 Measurement of INPs and Sources of Variability

INP measurements fall into two broad categories: chamber techniques, which expose particles to controlled temperature and humidity, and droplet-freezing assays, which cool a large number of droplets and record the frozen fraction as a function of temperature (Huang et al. 2021; Murray et al. 2012).

Droplet-freezing assays are widely used for immersion freezing because their comparatively large droplets contain enough nucleant material to make the rarer higher-temperature INPs observable. They measure the fraction of droplets that freeze, which must be related to the amount of nucleating material in each droplet and normalised by the aerosol surface area to derive $n_s(T)$ (Huang et al. 2021; Whale et al. 2015).

Measured spectra are affected by several distinct sources of uncertainty. The number of detected freezing events is uncertain because of instrumental uncertainty and the intrinsically stochastic nature of freezing. The latter can make the apparent n_s depend on the design of the experiment (Knopf et al. 2020; Burrows et al. 2022).

An intercomparison of seventeen immersion-freezing techniques applied to the same reference dust found deviations of about 8 K in activation temperature, or equivalently up to three orders of magnitude in n_s, larger than the stated uncertainty of any single instrument. This indicates that individual instruments may be reasonably precise while the accuracy across methods remains poorly established (Hiranuma et al. 2015).

Non-ideal behaviour in continuous-flow diffusion chambers leads to a systematic and variable underestimation of reported INP concentrations, with correction factors between about 1.5 and 9.5 (Garimella et al. 2017).

2.4 Temperature Dependence and INP Types Relevant to This Work

Measured INP concentrations spread across several orders of magnitude even when compared at one and the same temperature, and the whole distribution shifts upward steeply as the homogeneous freezing threshold is approached (David et al. 2019); at −30 °C, reported concentrations range from below 0.01 L^{-1} to above 100 L^{-1} (Burrows et al. 2022; DeMott et al. 2010).

The temperature at which a given particle becomes ice active depends strongly on its composition. Mineral dust generally becomes appreciably ice active below about −15 °C, whereas some highly active biological INPs, particularly ice-nucleating bacteria and bacterial preparations such as Snomax, can initiate freezing at higher temperatures (Kanji et al. 2017; Wex et al. 2015; Huang et al. 2021). The two materials studied here, desert dust and Snomax, are chosen as representatives of these different temperature-dependent regimes.

Mineral dust is regarded as the most important INP type for mixed-phase clouds, both because of its generally effective ice-nucleating ability and because emission rates of up to 5000 Tg yr^{-1} give it a near-global distribution (Kanji et al. 2017). Dust particles generally act as INPs below about −15 °C (Hoose and Möhler 2012), but can also facilitate heterogeneous ice nucleation at higher temperatures in the immersion mode, where the activity has been attributed largely to the feldspar group and specifically to K-feldspar: despite accounting for only about 3 % of atmospheric dust by mass, K-feldspar dominates the INP population above 248 K (Atkinson et al. 2013).

Biological INPs can be active at much higher temperatures than mineral dust. Strains of *Pseudomonas syringae* can initiate freezing at temperatures as high as about −2 °C (Huang et al. 2021; Wex et al. 2015). In the experiments of Möhler et al. (Möhler et al. 2008), the investigated bacterial cells were mainly ice active between approximately −7 and −11 °C, whereas Snomax showed strong activity around −8 °C. The activity arises from specialised ice-nucleating proteins that present a surface pattern capable of ordering nearby water molecules into an ice-like arrangement (Huang et al. 2021). In *Pseudomonas syringae*, these proteins are encoded by a single ice nucleation gene and are localised in the outer cell membrane (Orser et al. 1985; Huang et al. 2021).

Snomax, the biological material used in this work, is a commercial, freeze-dried preparation of non-viable *Pseudomonas syringae* and is commonly used as a biological reference material in laboratory studies of ice nucleation (Wex et al. 2015; Whale et al. 2015).

2.5 The Singular Description and the INAS Framework

Heterogeneous nucleation is usually described in terms of active sites, that is, localised features of the particle surface, such as cracks or chemical impurities, at which nucleation proceeds more readily than elsewhere (Hoose and Möhler 2012; Vali et al. 2015; Kanji et al.

2017). A direct consequence is that a particle offering more surface area is more likely to present somewhere on that surface a feature meeting the requirements for nucleation (Hoose and Möhler 2012).

In the singular, or deterministic, description one assumes that the time dependence of nucleation is of secondary importance next to the distribution of nucleus types, so that each active site becomes ice active at its own characteristic temperature (Murray et al. 2012; Connolly et al. 2009).

Under this assumption the ice-nucleating ability of a population is summarised by the INAS density $n_s(T)$, the cumulative number of sites per unit surface area that are active between the melting point and temperature T (Murray et al. 2012; Vali et al. 2015; Niemand et al. 2012).

The relation between this INAS density and the number of particles that actually nucleate goes back to Vali (Vali 1971), whose formulation is restated in the form used here by Murray et al. (Murray et al. 2012). Vali expressed the frozen fraction of an array of drops through a cumulative nucleus spectrum $K(T)$ defined per unit volume of liquid,

$$\frac{N_{\mathrm{ice}}(T)}{N_{\mathrm{tot}}} = 1 - \exp\left[-K(T)\,V\right], \tag{2.1}$$

where N_{tot} is the number of drops examined, $N_{\mathrm{ice}}(T)$ the cumulative number frozen at temperature T, and V the drop volume. Inverting Equation (2.1) gives the spectrum explicitly,

$$K(T) = -\frac{1}{V}\,\ln\left[1 - \frac{N_{\mathrm{ice}}(T)}{N_{\mathrm{tot}}}\right], \tag{2.2}$$

and the density of sites becoming active per unit temperature, $k(T)$, is related to the cumulative INAS density by $n_s(T) = -\int_{T_0}^{T} k(T')\,\mathrm{d}T'$, with T_0 the melting point (Murray et al. 2012).

In the INAS description the number of available sites scales with the surface area of the aerosol particle rather than with the volume of the surrounding liquid, so the drop volume in Equation (2.1) is replaced by the surface area of the individual nucleating particle. Writing $\bar{a}$ for the mean surface area per particle of a monodisperse or narrowly distributed population, the ice-active fraction of that population at temperature T is

$$\frac{c_{\mathrm{INP}}(T)}{c_{\mathrm{aer}}} = 1 - \exp\left[-n_s(T)\,\bar{a}\right], \tag{2.3}$$

where c_{aer} is the particle number concentration and $c_{\mathrm{INP}}(T)$ the concentration of particles that have nucleated ice by temperature T (Murray et al. 2012; Niemand et al. 2012). Inverting Equation (2.3) recovers the INAS density from a measured ice-active fraction,

$$n_s(T) = -\frac{1}{\bar{a}} \ln\left(1 - \frac{c_{\mathrm{INP}}(T)}{c_{\mathrm{aer}}}\right). \tag{2.4}$$

For a small ice-active fraction, $x = c_{\mathrm{INP}}(T)/c_{\mathrm{aer}} \ll 1$, the approximation $\ln(1-x) \approx -x$ reduces Equation (2.4) to $n_s(T) \approx c_{\mathrm{INP}}(T)/S$, where $S = c_{\mathrm{aer}}\,\bar{a}$ is the total particle surface area concentration. This linear form, valid while the ice-active fraction stays below about 0.1, is commonly used to derive surface-normalised INAS densities (Murray et al. 2012; Niemand et al. 2012; Ullrich et al. 2017), and it is the form in which the desert-dust parameterisation of Section 2.7 is reported.

For a monodisperse or narrowly distributed particle population, the total particle surface area concentration is related to the particle number concentration and the mean particle surface area by $S = c_{\mathrm{aer}}\,\bar{a}$. Therefore,

$$c_{\mathrm{aer}} = \frac{S}{\bar{a}}. \tag{2.5}$$

Combining Equation (2.5) with Equation (2.3) gives the standard INAS expression used here to calculate the INP concentration (Murray et al. 2012; Niemand et al. 2012):

$$c_{\mathrm{INP}}(T) = \frac{S}{\bar{a}}\left[1 - \exp\left(-n_s(T)\,\bar{a}\right)\right]. \tag{2.6}$$

Here, S is the total particle surface area per unit volume of air ($\mathrm{m^2\,m^{-3}}$), $\bar{a}$ is the mean surface area per particle ($\mathrm{m^2\,particle^{-1}}$), and their ratio $S/\bar{a}$ is the particle number concentration (particles $\mathrm{m^{-3}}$). The bracketed term in Equation (2.6) is bounded between zero and one, so the INP concentration cannot exceed the particle number concentration. For fixed $n_s(T)$ and $\bar{a}$, S controls the overall magnitude of the INP spectrum but does not change its shape. The temperature dependence of the spectrum is instead determined by the material through $n_s(T)$ and by the particle size through $\bar{a}$.

2.6 Lognormal Aerosol Size Distributions and Surface Area

Aerosol particle number-size distributions can often be approximated by a lognormal function (Heintzenberg 1994). Expressed with respect to the logarithm of the particle diameter, the number-size distribution is

$$\frac{\mathrm{d}n_p}{\mathrm{d}\ln d_p} = \frac{N_0}{\sqrt{2\pi}\ln\sigma_g} \exp\left[-\frac{\left(\ln d_p - \ln d_{pm}\right)^2}{2\ln^2\sigma_g}\right], \tag{2.7}$$

where N_0 is the total particle number concentration, d_p is the particle diameter, d_{pm} is the median diameter of the number-size distribution, and σ_g is the geometric standard deviation.

The median d_{pm} divides the number distribution such that half of the particles are smaller and half are larger (Heintzenberg 1994).

The median diameter of the surface-area-size distribution, d_{pm2}, differs from the median diameter of the number-size distribution. Applying the general moment-conversion relation given by (Heintzenberg 1994) to the second moment gives

$$\ln d_{pm2} = \ln d_{pm} + 2 \ln^2 \sigma_g. \tag{2.8}$$

Assuming spherical particles, the total particle surface area concentration is obtained from the second moment of the number-size distribution (Heintzenberg 1994):

$$S = \pi N_0 \exp\left(2 \ln d_{pm} + 2 \ln^2 \sigma_g\right), \tag{2.9}$$

where d_{pm} in Equation (2.9) is the median diameter of the number-size distribution, not the median of the surface-area-size distribution (Heintzenberg 1994).

The average surface area per particle is obtained by dividing the total surface area concentration by the particle number concentration:

$$\bar{a} = \frac{S}{N_0} = \pi d_{pm}^2 \exp\left(2 \ln^2 \sigma_g\right). \tag{2.10}$$

Therefore, $S = N_0 \bar{a}$. This relationship connects the lognormal particle size distribution to the INAS framework used in the following section, where the particle number concentration is denoted by c_{aer}.

2.7 INAS Parameterisations Used in This Work

Each species enters the synthetic spectrum generation framework through its own temperature-dependent ice nucleation activity. The two parameterisations used here come from different measurement traditions and are reported using different normalisations: one per unit surface area and one per unit mass. This section gives their published forms and the properties of those forms that decide how the two components appear in a mixed spectrum. The choices made in applying them, including the common temperature grid, the treatment outside the fitted range, and the numerical value of the mass-to-surface conversion, are specified in Chapter 3.

Mineral dust

The temperature dependence of the dust contribution is taken from the INAS parameterisation of Ullrich et al. (Ullrich et al. 2017), obtained by reanalysing eleven years of immersion-freezing experiments on desert dust in the AIDA cloud chamber. The fit is exponential in temperature,

$$n_{s,\mathrm{dust}}(T) = \exp\left(150.577 - 0.517\,T\right)\ \mathrm{m}^{-2}, \qquad T \in [243, 259]\ \mathrm{K}, \tag{2.11}$$

with T the absolute temperature and the coefficient 0.517 having the unit K^{-1}. The INAS density of desert dust increases continuously as the temperature decreases and does not reach a plateau within the fitted temperature interval.

Snomax

The biological contribution is described using the Snomax characterisation of Wex et al. (Wex et al. 2015), a seven-instrument intercomparison in which Snomax served as the common test substance. In contrast to the dust parameterisation, the ice nucleation activity is reported per unit mass. The number of ice nucleation active macromolecules was found to scale with the volume, and hence with the mass, of the material present in a droplet rather than with its surface area (Wex et al. 2015).

The published parameterisation is

$$n_{m,\mathrm{snomax}}(T_\mathrm{C}) = 1.4 \times 10^{9}\ \mathrm{mg}^{-1} \left[1 - \exp\left(-2 \times 10^{-10}\, e^{-\left(2.34\,^\circ\mathrm{C}^{-1}\right) T_\mathrm{C}}\right)\right], \tag{2.12}$$

where T_C is the temperature in degrees Celsius. The mass-based ice nucleation activity increases rapidly as the temperature decreases and approaches a plateau of $1.4 \times 10^{9}\ \mathrm{mg}^{-1}$ below about $-12\,^\circ\mathrm{C}$ (Wex et al. 2015). This behaviour differs from the INAS density of desert dust, which increases continuously with decreasing temperature within its fitted interval.

2.8 The INP Composition-Attribution Problem

Although the total INP concentration can be measured, identifying which particles act as INPs at a given temperature, and which chemical component of a particle carries the activity, remains one of the open problems of the field (Kanji et al. 2017; Burrows et al. 2022).

Internal mixing compounds the difficulty. Powders sold as individual minerals may contain other mineral components, so that even in a nominally pure laboratory sample the component responsible for nucleation may not be established unambiguously (Murray et al. 2012), and in the atmosphere the degree of internal and external mixing is itself usually unknown (Kanji et al. 2017). Direct chemical characterisation of ice-crystal residuals does not remove

the ambiguity either: imperfect separation of residuals from interstitial particles means that non-INPs can outnumber INPs among the particles that are analysed, and where several particle types are present together in the same sample, the activity cannot be assigned to any one of them (Burrows et al. 2022).

Different primary aerosol types can moreover produce closely similar mass-spectral signatures, and secondary material is often hard to distinguish from primary material (Christopoulos et al. 2018).

The particles responsible for a freezing spectrum therefore often cannot be identified uniquely from the freezing measurement alone. The mixed dust-Snomax spectra studied here are a controlled, synthetic version of this attribution problem, in which the composition of every sample is known by construction and can be used to train and evaluate a retrieval method.

2.9 Machine Learning Background

Machine learning provides methods for estimating relationships between input variables and an output from labelled examples. These methods range from relatively simple linear models to flexible nonlinear models such as tree ensembles and neural networks (Hastie et al. 2009; Bishop 2006). Machine-learning methods have become widely used in aerosol research because they can represent relationships in which several measured variables contribute jointly and nonlinearly to the quantity being predicted (Y. Wang et al. 2025).

Supervised Learning and Regression

In supervised learning, the available dataset consists of N labelled examples,

$$\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}, \tag{2.13}$$

where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$ contains the p input features of sample i, and y_i is its known target value. A model $f(\boldsymbol{x}; \boldsymbol{\theta})$, with parameters $\boldsymbol{\theta}$, is fitted to these examples and then used to predict the target for samples not used during fitting (Hastie et al. 2009; Bishop 2006).

If the target represents a discrete category, the task is classification. If the target is continuous, the task is regression. For regression, the model produces a numerical prediction

$$\hat{y}_i = f(\boldsymbol{x}_i; \boldsymbol{\theta}). \tag{2.14}$$

The parameters are commonly estimated by minimising a loss function that measures the difference between the observed and predicted targets. Under squared-error loss, the empirical loss is

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \left[y_i - f(\boldsymbol{x}_i; \boldsymbol{\theta}) \right]^2 . \tag{2.15}$$

A model should not only reproduce the training data but also generalise to new samples. A model that is too simple may fail to represent the underlying relationship, resulting in high bias. A model that is too flexible may fit random variation in the training data, resulting in high variance and overfitting. The balance between these effects is the bias-variance trade-off (Hastie et al. 2009; Bishop 2006).

Linear and Regularised Regression

A multiple linear regression model represents the prediction as

$$\hat{y}_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \tag{2.16}$$

where β_0 is the intercept and β_j is the coefficient of feature j. Ordinary least-squares regression estimates these coefficients by minimising the sum of the squared residuals (Hastie et al. 2009; Bishop 2006).

Linear regression can have a high variance when the input features are strongly correlated or when the number of features is large relative to the number of training samples. Regularisation addresses this problem by adding a penalty on the magnitude of the coefficients. Ridge regression, lasso regression and elastic-net regression can be written in the common form

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \left[(1-\eta) \sum_{j=1}^{p} \beta_j^2 + \eta \sum_{j=1}^{p} |\beta_j| \right] \right\}. \tag{2.17}$$

Here, $\lambda \geq 0$ controls the overall regularisation strength and $0 \leq \eta \leq 1$ controls the balance between the two penalties. When $\lambda = 0$, Equation (2.17) reduces to ordinary least squares. Ridge regression uses the squared-coefficient, or ℓ_2, penalty obtained with $\eta = 0$. It reduces the magnitudes of the coefficients but does not normally set them exactly to zero. Lasso regression uses the absolute-value, or ℓ_1, penalty obtained with $\eta = 1$. It can set some coefficients exactly to zero and can therefore perform feature selection. Elastic-net regression combines the two penalties using an intermediate value of η (Hastie et al. 2009).

Polynomial regression extends a linear model by replacing the original features with polynomial terms, such as squared terms and interactions between features. The prediction remains linear in its fitted coefficients, but it can represent a nonlinear relationship between the original features and the target. Increasing the polynomial degree increases the flexibility of the model but also increases its risk of overfitting (Bishop 2006; Hastie et al. 2009).

The penalties used by regularised regression depend on the numerical scales of the input features. Consequently, feature scaling or a clear justification for not applying it is important when regularised linear models are fitted (Hastie et al. 2009).

k-Nearest-Neighbours Regression

The k-nearest-neighbours method is an instance-based method. It does not fit an explicit global function. For a new input $\boldsymbol{x}$, it identifies the k training samples closest to $\boldsymbol{x}$ according to a selected distance measure and predicts the target from their observed values (Hastie et al. 2009; Bishop 2006).

For regression, the prediction can be written as

$$\hat{y}(\boldsymbol{x}) = \frac{\sum_{i \in \mathcal{N}_k(\boldsymbol{x})} w_i(\boldsymbol{x})\, y_i}{\sum_{i \in \mathcal{N}_k(\boldsymbol{x})} w_i(\boldsymbol{x})}, \tag{2.18}$$

where $\mathcal{N}_k(\boldsymbol{x})$ is the set of the k nearest training samples and w_i is the weight assigned to neighbour i. Uniform weighting assigns the same weight to every neighbour, whereas distance weighting gives more influence to closer neighbours.

A small value of k produces a flexible model that responds strongly to local variation and may have high variance. A large value of k averages over a wider region of the feature space and gives a smoother model, but it may fail to represent local structure. Because the neighbours are determined from distances between feature vectors, the result also depends on the relative numerical scales of the features and on the selected distance measure (Hastie et al. 2009).

Tree-Based Ensemble Models

Decision trees and random forests. A regression tree divides the feature space recursively into regions. At each internal node, a feature and a threshold are selected to divide the training samples into two child nodes. For regression, the split is commonly chosen to reduce the squared-error impurity, and the prediction in each terminal node is the mean target value of the training samples assigned to that node (Hastie et al. 2009; Louppe 2014).

A deep individual tree can represent nonlinear relationships and interactions, but its structure can change substantially when the training sample changes. Random forests reduce this instability by combining many trees. Each tree is trained on a bootstrap sample drawn with replacement from the original training set. In addition, only a randomly selected subset of the features is considered at each split. The regression prediction is the average of the individual tree predictions,

$$\hat{y}(\boldsymbol{x}) = \frac{1}{M} \sum_{m=1}^{M} T_m(\boldsymbol{x}), \tag{2.19}$$

where M is the number of trees and $T_m(\boldsymbol{x})$ is the prediction of tree m. Bootstrap sampling and random feature selection reduce the similarity between the trees, while averaging reduces the variance of the ensemble (Breiman 2001; Hastie et al. 2009).

Tree-based models divide the data using ordered feature thresholds. Their predictions are therefore not affected by monotonic rescaling of an individual feature. They can also represent nonlinear relationships and interactions without requiring these terms to be specified in advance (Hastie et al. 2009).

Gradient boosting. Random forests build trees largely independently and average their predictions. Gradient boosting instead builds an ensemble sequentially. The model after M steps has the additive form

$$F_M(\boldsymbol{x}) = F_0(\boldsymbol{x}) + \sum_{m=1}^{M} \nu \, \gamma_m T_m(\boldsymbol{x}), \tag{2.20}$$

where T_m is the tree added at step m, γ_m determines its contribution and ν is the learning rate.

At each step, the new tree is fitted to the negative gradient of the loss function evaluated using the current model. For squared-error loss, this negative gradient corresponds to the residuals between the observed and currently predicted targets. Each added tree therefore attempts to correct errors left by the existing ensemble. The number and depth of the trees control model complexity, while a smaller learning rate reduces the contribution of each new tree and normally requires more trees (Friedman 2001).

XGBoost and LightGBM. XGBoost and LightGBM are implementations of gradient-boosted decision trees. XGBoost uses a regularised objective that combines the training loss with penalties on the complexity and leaf weights of the trees. It also supports row and column subsampling and uses efficient procedures for identifying candidate split points (Chen and Guestrin 2016).

LightGBM uses gradient-based one-side sampling, which retains samples with large gradients while sampling from those with small gradients, and exclusive feature bundling, which combines features that rarely take non-zero values at the same time. These procedures reduce the computational cost of fitting gradient-boosted trees, particularly for large or high-dimensional datasets (Ke et al. 2017). XGBoost and LightGBM retain the basic sequential structure of gradient boosting but differ in their regularisation, tree-construction and computational strategies.

Neural Networks

A feed-forward neural network represents a prediction as a sequence of transformations arranged in layers. If $\boldsymbol{h}^{(0)} = \boldsymbol{x}$ is the input vector, the output of hidden layer ℓ is

$$\boldsymbol{h}^{(\ell)} = \sigma_\ell \left(\boldsymbol{W}^{(\ell)} \boldsymbol{h}^{(\ell-1)} + \boldsymbol{b}^{(\ell)} \right), \tag{2.21}$$

where $\boldsymbol{W}^{(\ell)}$ contains the layer weights, $\boldsymbol{b}^{(\ell)}$ contains the biases and σ_ℓ is a nonlinear activation function. Common activation functions include the hyperbolic tangent and the rectified linear unit. For regression, the final layer normally has a linear output so that it can return a continuous value (Bishop 2006; Goodfellow et al. 2016).

A multilayer perceptron is a feed-forward network containing one or more hidden layers. The nonlinear activation functions allow it to represent relationships that cannot be expressed by a linear model. Feed-forward networks with a sufficiently large hidden layer can approximate a broad class of continuous functions, although this theoretical result does not determine the number of units required or guarantee that training will find the best parameters (Hornik et al. 1989).

The weights are estimated by minimising a loss function. Backpropagation computes the gradient of the loss with respect to the parameters by propagating derivatives backwards through the network, after which a gradient-based optimisation method updates the weights. Networks with several hidden layers are generally described as deep neural networks (Goodfellow et al. 2016).

Increasing the number and width of the hidden layers increases the model capacity but also increases the risk of overfitting. Weight penalties limit the magnitude of the fitted weights. Dropout provides another form of regularisation by temporarily removing randomly selected units during training, preventing the fitted network from relying too strongly on individual connections (Goodfellow et al. 2016).

Model Assessment and Selection

The available data are commonly divided into training, validation and test subsets. The training subset is used to fit model parameters. The validation subset is used to select hyperparameters, such as regularisation strength, tree depth or the number of neighbours. The test subset is kept separate during training and model selection and is used only for the final assessment of generalisation performance (Hastie et al. 2009; Bishop 2006).

Cross-validation provides a more stable estimate of generalisation error from the available training data. In k-fold cross-validation, the data are divided into k approximately equal folds. The model is fitted k times, each time using $k - 1$ folds for training and the remaining fold for validation. Every sample is therefore used once for validation, and the validation scores are averaged across the folds (Hastie et al. 2009).

For a regression dataset containing N target values, the root mean square error is

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}. \tag{2.22}$$

RMSE is expressed in the same unit as the target. Because the residuals are squared, larger prediction errors have a stronger influence on RMSE than smaller errors.

The coefficient of determination is

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \tag{2.23}$$

where $\bar{y}$ is the mean observed target. A value of $R^2 = 1$ indicates exact predictions, while $R^2 = 0$ indicates the same squared error as predicting the mean target for every sample. A negative value can occur when the model performs worse than this mean-value prediction (Hastie et al. 2009; Bishop 2006).

Overfitting can be assessed by comparing performance on the training data with performance on validation or test data. A substantially lower training error than validation error indicates that the model has fitted patterns that do not generalise. Cross-validation estimates also contain sampling uncertainty, and small differences between mean scores do not necessarily indicate meaningful differences between models.

The one-standard-error rule accounts for this uncertainty. It first identifies the model with the lowest mean cross-validation error and then selects the least complex model whose error is no more than one estimated standard error above that minimum. The rule therefore favours a simpler model when its estimated performance cannot be clearly distinguished from that of the best-scoring model (Hastie et al. 2009).

Feature Importance and SHAP

Accurate predictions alone do not show how a model uses its input features. Feature-importance and feature-attribution methods provide complementary ways of examining the fitted model. Impurity-based importance gives a global summary derived from the structure of a tree ensemble, whereas SHAP assigns a contribution to each feature for each individual prediction.

Impurity-based feature importance. When a regression tree splits a node, the split reduces the squared-error impurity of the samples in that node. The impurity-based importance of feature j in an ensemble can be written as

$$I_j = \frac{1}{M} \sum_{m=1}^{M} \sum_{\substack{t \in T_m \\ v(t)=j}} p(t)\, \Delta i(t), \tag{2.24}$$

where M is the number of trees, $v(t)$ is the feature used at node t, $p(t)$ is the proportion of training samples reaching that node, and $\Delta i(t)$ is the reduction in impurity produced by the split. The importance values are usually normalised so that their sum over all features is one (Louppe 2014).

A high value means that the feature was frequently used to produce large reductions in impurity. It does not show whether increasing the feature increases or decreases the prediction, and it should not be interpreted as a causal effect. The values can also be affected by correlated features and by masking, where one of several related features is selected for a split while the others receive lower importance (Louppe 2014).

SHAP values. SHAP, or SHapley Additive exPlanations, is based on the Shapley value from cooperative game theory. For a particular sample $\boldsymbol{x}$, the model prediction is represented as

$$f(\boldsymbol{x}) = \phi_0 + \sum_{j=1}^{p} \phi_j(\boldsymbol{x}), \tag{2.25}$$

where ϕ_0 is the expected model prediction over a reference population and $\phi_j(\boldsymbol{x})$ is the contribution of feature j to the prediction for sample $\boldsymbol{x}$. A positive SHAP value moves the prediction above the reference value, while a negative SHAP value moves it below the reference value (Lundberg and Lee 2017).

Let $\mathcal{F}$ denote the set of all p input features and let $v_{\boldsymbol{x}}(\mathcal{S})$ denote the expected model output when the features in subset $\mathcal{S}$ are known for sample $\boldsymbol{x}$. The SHAP value of feature j is

$$\phi_j(\boldsymbol{x}) = \sum_{\mathcal{S} \subseteq \mathcal{F} \setminus \{j\}} \frac{|\mathcal{S}|!\ (p - |\mathcal{S}| - 1)!}{p!} \left[v_{\boldsymbol{x}}\left(\mathcal{S} \cup \{j\}\right) - v_{\boldsymbol{x}}(\mathcal{S}) \right]. \tag{2.26}$$

Equation (2.26) averages the marginal contribution of feature j over all possible subsets of the other features. The weighting accounts for the different orders in which features can be added. The resulting attributions satisfy local accuracy, missingness and consistency within the class of additive feature-attribution methods (Lundberg and Lee 2017).

Direct evaluation of Equation (2.26) becomes expensive as the number of features increases because it considers all possible feature subsets. TreeSHAP uses the internal structure of

decision trees to calculate exact Shapley-value explanations for trees and tree ensembles in polynomial rather than exponential time (Lundberg et al. 2020). It can therefore be applied efficiently to random forests and gradient-boosted trees. SHAP values are local because they explain one prediction. They can be combined across samples to describe the model globally. The mean absolute SHAP value,

$$I_j^{\mathrm{SHAP}} = \frac{1}{N} \sum_{i=1}^{N} \left| \phi_j(\boldsymbol{x}_i) \right|, \tag{2.27}$$

summarises the average magnitude of the contribution of feature j. A SHAP summary plot retains more information by showing the distribution, magnitude and direction of the feature contributions across the samples (Lundberg et al. 2020).

Machine Learning in Aerosol and Ice-Nucleation Research

Early machine-learning applications in aerosol research included supervised prediction of pollutant concentrations and the classification of single-particle mass spectra. Applications now also include source estimation, instrument correction and prediction of aerosol physicochemical properties (Y. Wang et al. 2025).

Christopoulos et al. (Christopoulos et al. 2018) used random forests to classify single-particle mass spectra into aerosol composition classes. G. Wang et al. (G. Wang et al. 2024) compared k-nearest neighbours, decision trees, random forests, support-vector machines and multilayer perceptrons for automatic classification of single-particle mass-spectrometry data.

Regression models have also been used for continuous aerosol properties. May and Li (May and Li 2022) applied ordinary least squares, stepwise regression, lasso, support-vector regression and neural-network models to predict the mass absorption cross-section of black carbon from aerosol measurements. In an application directly related to INP instrumentation,

Garimella et al. (Garimella et al. 2017) used random-forest regression to characterise the non-ideal behaviour of a continuous-flow diffusion chamber. These studies demonstrate the use of supervised learning for both aerosol classification and the prediction of continuous atmospheric quantities.

3 Methodology

3.1 Overview of the modelling pipeline

This chapter presents the end-to-end workflow used to generate a synthetic ensemble of INP spectra for two-component dust + Snomax mixtures, extracting a set of scale-invariant descriptors from each spectrum, and training regression models on these features for the purpose of recovering the INP spectrum composition.

The objective of this study is to establish the foundation of a solid machine learning workflow to recover the fraction of each component of a single perturbed synthetic spectrum. The pipeline (Figure 3.1) proceeds in the following stages, each of which is detailed in the sections indicated:

1. Computation of the species-specific INAS parameterisations for mineral dust and Snomax on a common temperature grid (Section 3.2).
2. Conversion of the Snomax parameterisation to a surface-area basis and formulation of the two-species mixed-spectrum equation (Sections 3.2 and 3.3).
3. Generation of a per-sample surface-area ensemble by Latin-hypercube sampling of the composition R and the number-median diameter d_{pm}, which controls the total surface-area concentration (Section 3.4).
4. Synthesis of the mixed INP spectra, including a per-sample INAS parameterisation perturbation and a multiplicative pointwise spectrum-perturbation model (Section 3.5).
5. Classification of each sample into a detectability regime and restriction of the dataset to the mixed regime (Section 3.6).
6. Extraction of a set of scale-invariant spectral features (Section 3.7).
7. Selection of the candidate features through variance and correlation filtering, followed by SHAP-based feature ranking (Section 3.8).
8. Stratified partitioning of the dataset into training, validation and test subsets (Section 3.9).
9. Training, tuning and evaluation of a suite of regression models under a common protocol (Section 3.10).
10. Selection of a single model by a cross-validation-based one-standard-error rule (Section 3.11).

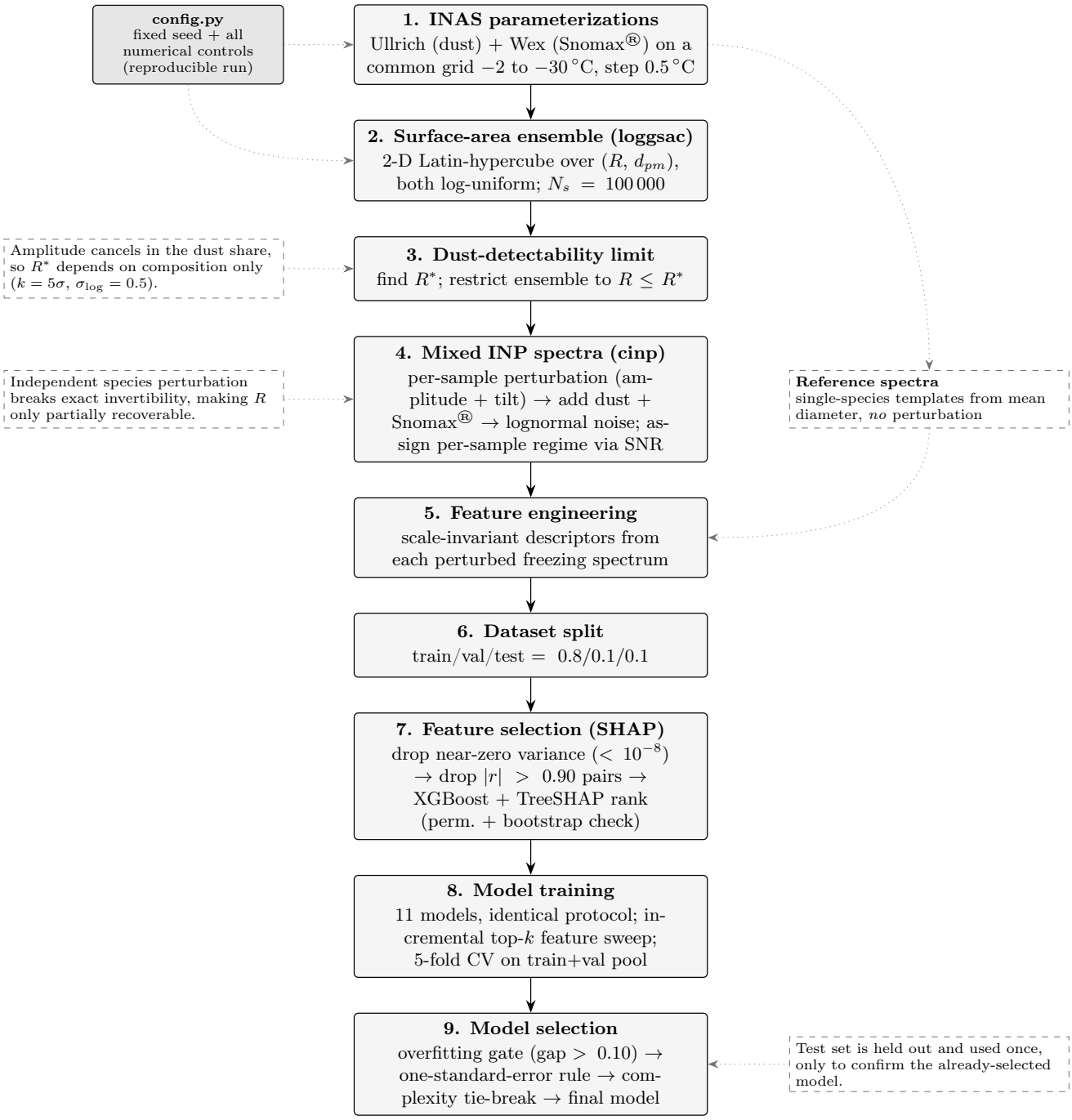


Figure 3.1: Overview of the modelling pipeline. The INAS parameterisations of the two species and the surface-area ensemble are combined, and a mixed spectrum is synthesised for every sample with a per-sample INAS parameterisation perturbation and a pointwise spectrum-perturbation model; the detectability regime of each sample is assigned during the synthesis. Scale-invariant features are then extracted from the perturbed spectra of the mixed-regime samples only, the dataset is split, the feature set is reduced by variance and correlation filtering and ranked by SHAP, the regression models are trained under a common protocol, and a single model is finally selected. All numerical controls and the random seed are fixed in `config.py`.

The two source parameterisations use different temperature units: the dust parameterisation is expressed in kelvin, whereas the Snomax parameterisation is expressed in degrees Celsius. Therefore, the equations and their validity ranges are reported below using the units of the original publications. Temperatures are converted explicitly when the two parameterisations are evaluated on the common grid and combined during spectrum synthesis. The complete set of numerical control values is fixed in a single configuration module, and all stochastic steps use a recorded random seed to ensure reproducibility.

3.2 Ice nucleation active surface site density parameterisations

The mixed spectrum is constructed from two INP species using the parameterisations introduced in Section 2.7:

- Desert dust is represented by the parameterisation of (Ullrich et al. 2017), given in Equation (2.11) as

$$n_{s,\mathrm{dust}}(T) = \exp\left(150.577 - 0.517\,T\right)\ \mathrm{m}^{-2}, \qquad T \in [243, 259]\ \mathrm{K}.$$

- Snomax is represented by the parameterisation of (Wex et al. 2015), given in Equation (2.12) as

$$n_{m,\mathrm{snomax}}(T_\mathrm{C}) = 1.4 \times 10^{9} \left[1 - \exp\left(-2 \times 10^{-10} e^{-2.34\,T_\mathrm{C}}\right)\right]\ \mathrm{mg}^{-1},$$

  where T_C is the temperature in degrees Celsius.

Each parameterisation is valid in its own temperature domain; in this study both are evaluated on one common temperature grid. The grid is generated from 243 K to 271 K in steps of 0.5 K, which gives 57 points and corresponds to −30.15 °C to −2.15 °C after conversion to degrees Celsius. The higher-temperature bound is set by the Snomax parameterisation and the lower-temperature bound by the temperature range over which the dust parameterisation was fitted.

The common grid uses a temperature interval of 0.5 K. This matches the 0.5 K temperature bins used to derive the Snomax parameterisation in its high temperature active regime above −12.5 °C (Wex et al. 2015). The dust parameterisation is based on AIDA chamber measurements, for which a temperature uncertainty of approximately ±0.3 K has been reported (Möhler et al. 2006).

A temperature interval of 1 K could under-sample the steep change in Snomax activity near its freezing onset (Wex et al. 2015). Conversely, an interval finer than 0.5 K would increase the numerical resolution without adding independent information beyond the resolution of the underlying measurements. The 0.5 K interval is therefore retained as a compromise between resolving the temperature dependence and remaining consistent with the source data.

Mineral dust

The dust INAS density follows the log-linear fit of (Ullrich et al. 2017), Equation (2.11), obtained from the AIDA cloud-chamber immersion-freezing experiments. The fit is applied unchanged over its fitted validity range 243-259 K, with the temperature in kelvin.

The common temperature grid extends to 271 K to include the high-temperature activity of Snomax. Because the mixed spectrum is calculated from the sum of the dust and Snomax contributions at every point on this grid, $n_{s,\mathrm{dust}}$ must also be defined between 259 and 271 K. However, the published dust parameterisation is valid only up to 259 K. Rather than applying the original fit outside its validity range or setting the dust contribution abruptly to zero, a continuous exponential decay is prescribed above this threshold. This extension is a modelling choice of the present study and is not part of the parameterisation of (Ullrich et al. 2017):

$$n_{s,\mathrm{dust}}(T) = n_{\mathrm{ref}} \exp\left(-k_{\mathrm{decay}}\,\Delta T\right), \qquad T > T_{\mathrm{high}}, \tag{3.1}$$

where $n_{\mathrm{ref}} = n_{s,\mathrm{dust}}(259\,\mathrm{K})$ is the value at the threshold, $\Delta T = T - 259\,\mathrm{K}$, and $k_{\mathrm{decay}} = 0.517 \times f_{\mathrm{slope}} = 1.551\,\mathrm{K}^{-1}$ is the amplified decay slope obtained with the amplification factor $f_{\mathrm{slope}} = 3.0$. It is deliberate: suppressing the dust activity on the high temperature side ensures that dust becomes undetectable in mixed dust and Snomax compositions at the high temperature end of the grid, where Snomax dominates the ice nucleation activity.

Snomax

The Snomax INAS density is taken from the intercomparison study of (Wex et al. 2015), Equation (2.12), in which the ice nucleation activity of Snomax is expressed per unit *mass* of substance, in mg^{-1}, with the temperature in degrees Celsius. In the implementation, the parameterisation is evaluated on the common temperature grid from 243 K to 271 K.

Figure 3.2 shows the unperturbed fitted dust and Snomax curves. Separate vertical axes are used because the dust activity is expressed per unit surface area, while the Snomax activity is expressed per unit mass. The dust INAS density decreases continuously as temperature increases. In contrast, the Snomax density remains close to its maximum value below about $-12\,^{\circ}\mathrm{C}$ and decreases sharply at higher temperatures. This difference in temperature dependence allows the retrieval method to estimate the composition from the shape of a mixed spectrum.

Surface-to-mass conversion for Snomax

Because the dust parameterisation is surface-area based (Equation (2.11)), whereas the Snomax parameterisation is mass based (Equation (2.12)), the Snomax INAS density is converted to a surface-area equivalent using

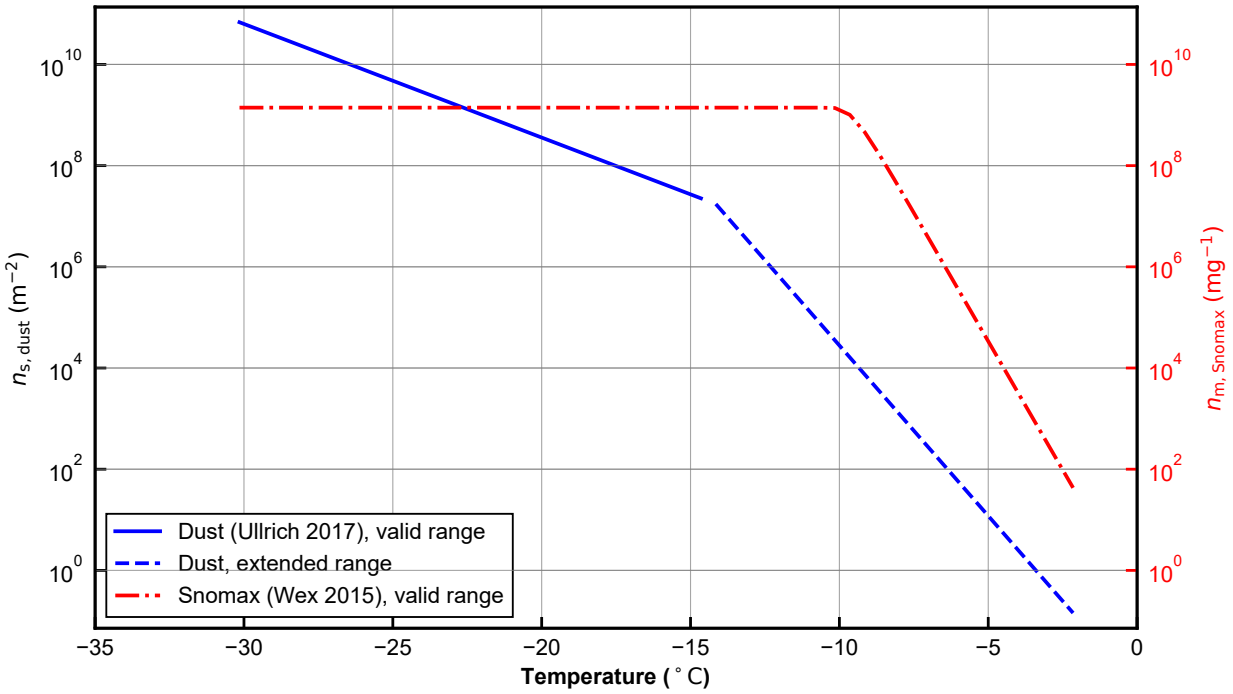


Figure 3.2: Unperturbed fitted ice-nucleation active-site densities of mineral dust (Ullrich et al. 2017) and of Snomax (Wex et al. 2015) on the common temperature grid. The two parameterisations are reported on different bases and are therefore drawn against separate axes, with no conversion applied: the dust density $n_{s,\mathrm{dust}}$ is per unit surface area (left axis, m^{-2}) and the Snomax density $n_{m,\mathrm{snomax}}$ is per unit mass (right axis, mg^{-1}). Only the shapes are comparable here, not the levels. The dust curve is the log-linear Ullrich fit inside its fitted range (solid) and the accelerated high temperature decay of Equation (3.1) above 259 K (dashed); the Snomax curve is the saturating Wex form, sitting on its ceiling lower than about −12 °C and falling steeply towards higher temperatures. The conversion of the Snomax density to a common surface-area basis by the factor of Equation (3.3) is applied only when the two species contributions are combined into a spectrum (Section 3.2).

$$n_{s,\mathrm{snomax}}(T) = f_m \, n_{m,\mathrm{snomax}}(T), \tag{3.2}$$

where f_m is the mass associated with a unit particle surface area. Treating the Snomax particles as spheres of diameter d and bulk density ρ, the volume-to-surface ratio is

$$\frac{V}{A} = \frac{d}{6}.$$

Therefore, the mass associated with a surface area concentration S is

$$m = \rho \left(\frac{d}{6} \right) S.$$

The corresponding conversion factor is

$$f_m = \frac{\rho d}{6} = 2.25 \times 10^{-4}\ \mathrm{kg\,m^{-2}} = 225\ \mathrm{mg\,m^{-2}}, \tag{3.3}$$

evaluated with $\rho = 1350\ \mathrm{kg\,m^{-3}}$ and $d = 1 \times 10^{-6}$ m (Wex et al. 2015). The value $225\ \mathrm{mg\,m^{-2}}$ is used with the mass-based Snomax density in $\mathrm{mg^{-1}}$. Their product has units of $\mathrm{m^{-2}}$, placing the Snomax activity on the same surface-area basis as the dust activity. The numerically equivalent kilogram-based quantities may be used in the implementation without changing the result.

3.3 Mixed aerosol populations and composition recovery

When two types of INP are present in the same sample as an external mixture, their ice-nucleation active sites are assumed to act independently. Each component is treated as a separate population with its own INAS density and mean particle surface area. Under this assumption, the species-specific INP concentrations are added to obtain the total INP concentration (Murray et al. 2012; Wex et al. 2015; Phillips et al. 2008):

$$\begin{aligned} c_{\mathrm{INP}}(T) = {} & \frac{S_{\mathrm{dust}}}{\bar{a}_{\mathrm{dust}}} \left[1 - \exp\left(-n_{s,\mathrm{dust}}(T)\, \bar{a}_{\mathrm{dust}}\right)\right] \\ & + \frac{S_{\mathrm{snomax}}}{\bar{a}_{\mathrm{snomax}}} \left[1 - \exp\left(-n_{s,\mathrm{snomax}}(T)\, \bar{a}_{\mathrm{snomax}}\right)\right]. \end{aligned} \tag{3.4}$$

Here, S_{dust} and S_{snomax} are the component particle surface area concentrations, expressed in $\mathrm{m^2\,m^{-3}}$, and $\bar{a}_{\mathrm{dust}}$ and $\bar{a}_{\mathrm{snomax}}$ are the corresponding mean particle surface areas, expressed in $\mathrm{m^2\ particle^{-1}}$. The same additivity justifies subtracting a background spectrum from a sample spectrum in the analysis of real measurements (David et al. 2019).

The total particle surface area concentration is

$$S = S_{\mathrm{dust}} + S_{\mathrm{snomax}}. \tag{3.5}$$

The Snomax surface fraction is defined as

$$R = \frac{S_{\mathrm{snomax}}}{S_{\mathrm{dust}} + S_{\mathrm{snomax}}}. \tag{3.6}$$

The component surface area concentrations can therefore be written as $S_{\mathrm{snomax}} = R\,S$ and $S_{\mathrm{dust}} = (1 - R)\,S$. The parameter R is a fraction between zero and one and appears in the code and output files under the name `ssf`. It describes the composition of the mixture, while S describes its total particle loading.

For each species i, its INP response per unit particle surface-area concentration is defined as

$$q_i(T) = \frac{1}{\bar{a}_i}\left[1 - \exp\left(-n_{s,i}(T)\,\bar{a}_i\right)\right]. \tag{3.7}$$

Substituting the component surface area concentrations into Equation (3.4) gives

$$c_{\mathrm{INP}}(T) = S\left[(1-R)\,q_{\mathrm{dust}}(T) + R\,q_{\mathrm{snomax}}(T)\right]. \tag{3.8}$$

Equation (3.8) separates the effects of total loading and composition. For fixed species responses, S controls the overall magnitude of the spectrum, while R controls the relative contributions of dust and Snomax. Because their INAS densities have different temperature dependences, changing R also changes the shape of the mixed spectrum. Recovering R from a single perturbed mixed spectrum is the inverse problem considered in this work. The retrieval is limited by two effects. First, independent perturbations are applied around the unperturbed fitted INAS density of each component to represent the reported variability of the parameterisations. Consequently, samples with the same value of R can produce different spectra. This parameterisation variability is treated separately from the pointwise spectrum perturbation. Second, when one component contributes only a small share of the mixture, its contribution can fall below the adopted detectability threshold.

The retrieval is therefore evaluated only over the composition range in which both components remain detectable.

3.4 Surface-area concentration ensemble

The composition of each synthetic sample determines how the total aerosol surface area is divided between dust and Snomax. The total surface area concentration S and the average surface area per particle $\bar{a}$ are calculated using Equations (2.9) and (2.10), respectively, introduced in Section 2.6.

For both species, the particle number concentration is fixed at $N_0 = 1000\,\mathrm{m}^{-3}$, and the geometric standard deviation is fixed at $\sigma_g = 1.5$. The value of σ_g is a prescribed modelling choice rather than a material-specific measurement. It is kept fixed so that variation in the distribution width does not introduce an additional input variable. The same value is used for both species so that an assumed difference in distribution width cannot provide information about the composition. The median diameter of the number-size distribution, d_{pm}, is sampled as described in Section 3.4. Within each sample, both species use the same d_{pm} and σ_g, so $\bar{a}_{\mathrm{dust}} = \bar{a}_{\mathrm{snomax}}$.

Size range and the choice of the amplitude bounds

With N_0 and σ_g fixed, the sampled d_{pm} changes both the total surface area concentration S and the mean particle surface area $\bar{a}$. In the linear limit $n_s(T)\bar{a} \ll 1$, Equation (2.6) reduces to $c_{\mathrm{INP}}(T) \approx S\, n_s(T)$, so changing d_{pm} primarily changes the overall concentration level of the spectrum. As activation approaches saturation, however, the associated change in $\bar{a}$ can also affect the spectral shape. The bounds must therefore be selected to limit saturation across the informative temperature and composition ranges. In this study, d_{pm} is used as a numerical control variable for the particle loading and is not claimed to represent the physical particle size of either dust or Snomax. The input data for the machine-learning models are constructed to be independent of the total surface-area concentration (Section 3.7).

Operative bounds. The diameter range was selected to keep the spectra calculated from the unperturbed fitted parameterisations within the partially activated regime. If the mean particle surface area is too large, the activation term in Equation (2.6) approaches unity over part of the temperature range. The INP concentration then approaches the particle number concentration $S/\bar{a}$, and the spectrum becomes less responsive to changes in the INAS curve. This reduces the composition information carried by the spectral shape.

The amplitude bounds were therefore selected so that, for the unperturbed fitted parameterisations, the activation term remains well below unity across the informative temperature and composition ranges. The same values were used for both species:

$$d_{pm} \in [1.0 \times 10^{-8},\ 5.0 \times 10^{-8}]\ \mathrm{m}. \tag{3.9}$$

This interval retains a factor of five in number-median diameter and therefore a factor of twenty-five in surface area. Using Equation (2.9), the corresponding total surface area concentration ranges from 4.36×10^{-13} m^2 m^{-3} to 1.09×10^{-11} m^2 m^{-3}. At the upper diameter bound, the activation terms calculated from the unperturbed fitted parameterisations remain well below unity across the temperature grid, keeping the corresponding reference spectra within the partially activated regime.

The same diameter bounds are assigned to both species as a controlled modelling choice. Together with the common values of N_0 and σ_g, this prevents the assumed particle loading from introducing species-dependent differences and allows the influence of the two INAS parameterisations on the spectral shape to be compared directly. This choice does not imply that atmospheric dust and Snomax populations have identical size distributions.

The ensemble consists of $N_s = 100\,000$ samples. Each sample is defined by two sampled variables: the Snomax surface fraction R, defined in Equation (3.6), and an amplitude number-median diameter d_{pm} within the range of Equation (3.9). The diameter determines the overall scale of the spectrum and is varied so that the machine-learning models cannot rely on absolute concentration alone. A two-dimensional Latin hypercube sampling (LHS) design (McKay et al. 1979) was used to sample these variables. The construction is summarised in Figure 3.3.

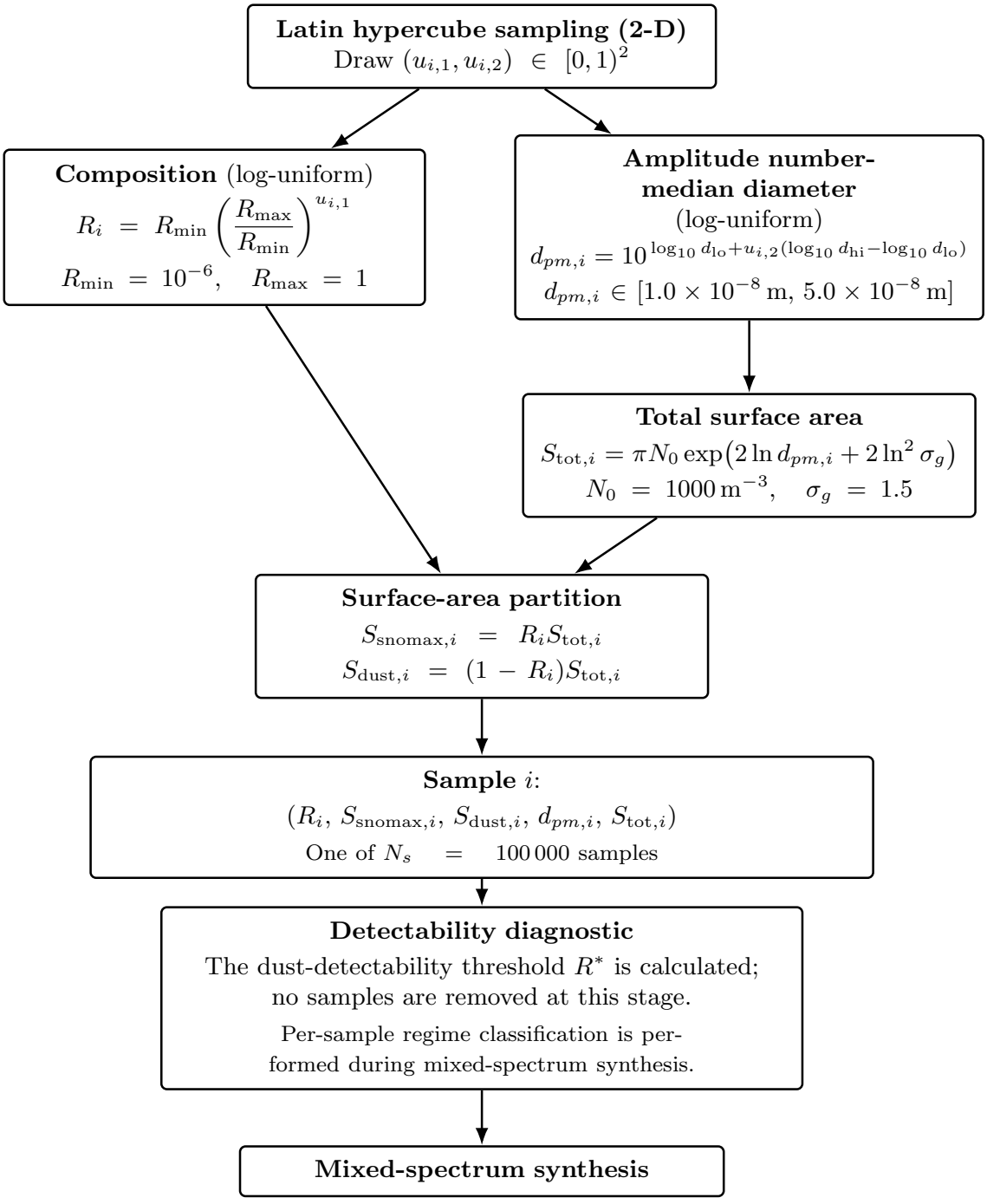


Figure 3.3: Construction of the surface-area concentration ensemble: two Latin hypercube coordinates determine the Snomax surface fraction R_i and the amplitude number-median diameter $d_{pm,i}$. The diameter determines the total surface area $S_{\mathrm{tot},i}$, which is partitioned by R_i into the dust and Snomax contributions.

The LHS design stratifies each variable across its range, reducing the clustering and gaps that can occur with independent random sampling without requiring a full Cartesian product. Figure 3.4 illustrates this sampling design.

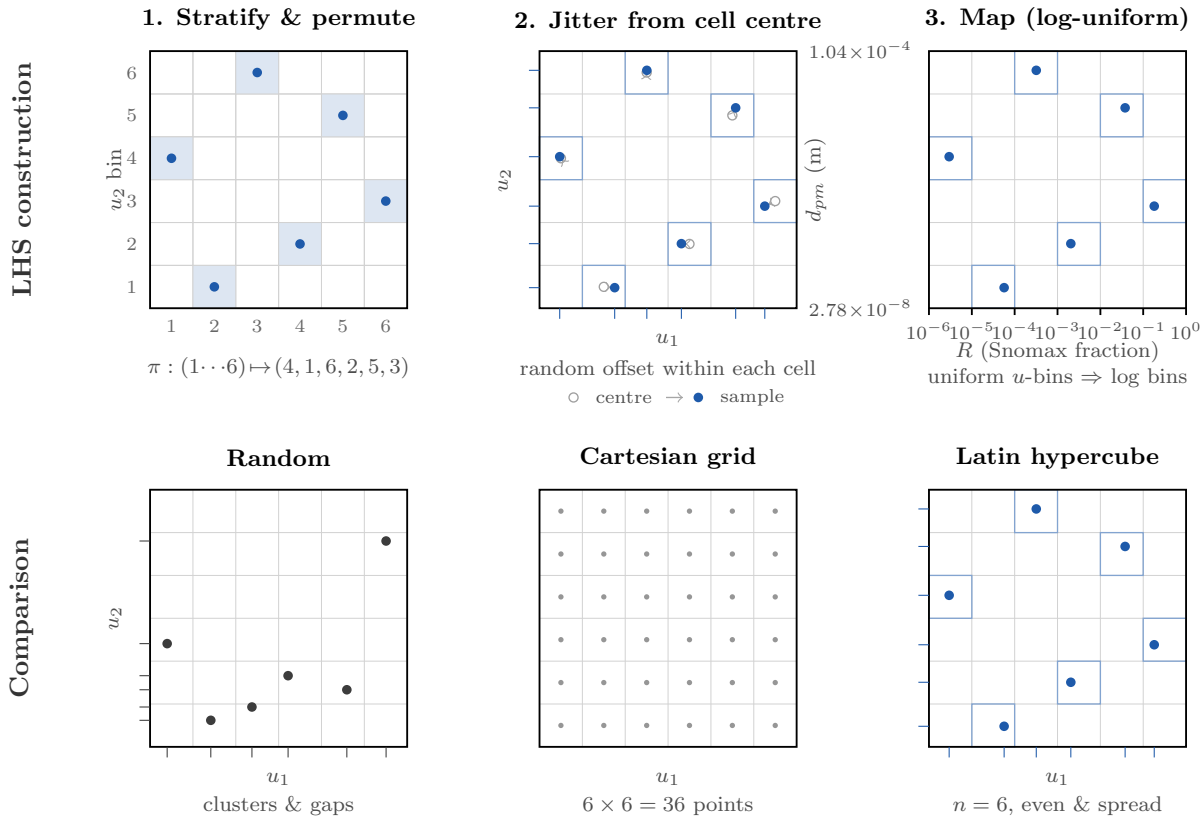


Figure 3.4: Two-dimensional Latin hypercube sampling of the two ensemble variables. **Top row:** each variable is divided into equal intervals, a random permutation selects one cell in every row and column, and one point is sampled within each selected cell. The log-uniform transformations then map the unit-square coordinates to the Snomax surface fraction R and the amplitude number-median diameter d_{pm}. **Bottom row:** comparison with independent random sampling and a full Cartesian grid at the same illustrative sampling size. The illustration uses six samples for legibility, whereas the ensemble consists of $N_s = 100\,000$ samples.

Let $u_{i,1}, u_{i,2} \in [0, 1)$ denote the two Latin hypercube coordinates of sample i. The Snomax surface fraction is drawn log-uniformly as

$$R_i = R_{\min} \left(\frac{R_{\max}}{R_{\min}} \right)^{u_{i,1}}, \qquad R_{\min} = 10^{-6}, \quad R_{\max} = 1, \tag{3.10}$$

so that small Snomax fractions, where the dust signature dominates and the transition between regimes occurs, are sampled more densely than they would be under linear-uniform sampling.

The number-median diameter is also drawn log-uniformly over the shared amplitude interval of Equation (3.9):

$$d_{pm,i} = 10^{\log_{10} d_{\mathrm{lo}} + u_{i,2}\left(\log_{10} d_{\mathrm{hi}} - \log_{10} d_{\mathrm{lo}}\right)}, \tag{3.11}$$

where $d_{\mathrm{lo}} = 1.0 \times 10^{-8}\,\mathrm{m}$ and $d_{\mathrm{hi}} = 5.0 \times 10^{-8}\,\mathrm{m}$ are the number-median-diameter limits used for both species.

The total surface area $S_{\mathrm{tot},i}$ of sample i is calculated using Equation (2.9), with $d_{pm} = d_{pm,i}$, and is partitioned according to the composition:

$$S_{\mathrm{snomax},i} = R_i S_{\mathrm{tot},i}, \qquad S_{\mathrm{dust},i} = (1 - R_i) S_{\mathrm{tot},i}. \tag{3.12}$$

Consequently,

$$R_i = \frac{S_{\mathrm{snomax},i}}{S_{\mathrm{snomax},i} + S_{\mathrm{dust},i}}$$

holds exactly. The resulting parameter tuple

$$(R_i,\ S_{\mathrm{snomax},i},\ S_{\mathrm{dust},i},\ d_{pm,i},\ S_{\mathrm{tot},i})$$

defines the input to the spectrum synthesis for sample i.

A fixed reference number-median diameter is calculated from the midpoint of the shared diameter range:

$$d_{pm}^{\mathrm{ref}} = \frac{1}{2}\left(d_{\mathrm{lo}} + d_{\mathrm{hi}}\right) = 3.0 \times 10^{-8}\,\mathrm{m}.$$

Using Equation (2.9), this gives the same reference surface area, $3.93 \times 10^{-12}\,\mathrm{m^2\,m^{-3}}$, for both species. The resulting single-species reference spectra therefore differ through their INAS parameterisations rather than their assumed loading. They define the reference shapes used during feature extraction (Sections 3.5 and 3.7).

A representative lognormal surface-area distribution is shown in Figure 3.5. Because both species use the same geometric standard deviation and diameter bounds, the distribution applies to both dust and Snomax.

Table 3.1: Summary of the parameterisations and particle-population parameters used for dust and Snomax. The conversion density and diameter apply only to Snomax because its published parameterisation is mass based.

Parameter	Dust	Snomax
INAS parameterisation	Ullrich et al. (Ullrich et al. 2017)	Wex et al. (Wex et al. 2015)
Original parameterisation basis	Surface area (m^{-2})	Mass (mg^{-1})
Particle number concentration, N_0	$1000\,m^{-3}$	$1000\,m^{-3}$
Geometric standard deviation, σ_g	1.5	1.5
Sampled number-median diameter, d_{pm}	1.0×10^{-8} m to 5.0×10^{-8} m	1.0×10^{-8} m to 5.0×10^{-8} m
Reference number-median diameter, d_{pm}^{ref}	3.0×10^{-8} m	3.0×10^{-8} m
Bulk density used for conversion, ρ	Not required	$1350\,kg\,m^{-3}$ (Wex et al. 2015)
Particle diameter used for conversion, d	Not required	1.0×10^{-6} m (Wex et al. 2015)

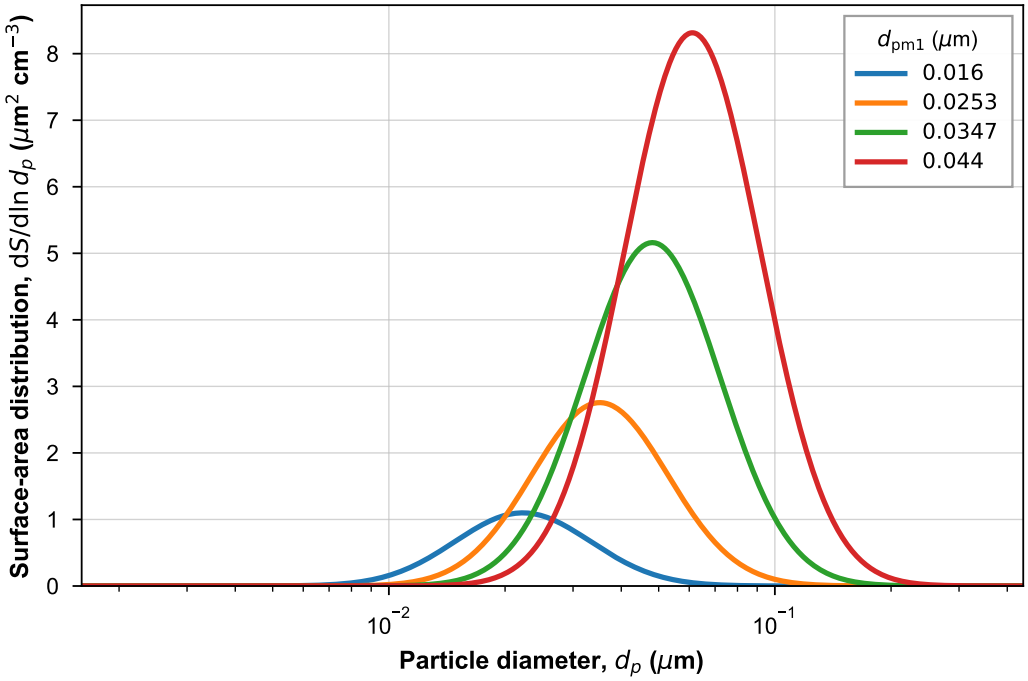


Figure 3.5: Representative lognormal surface-area distributions for the sampled number-median-diameter range. A single distribution family applies to both dust and Snomax because the same particle number concentration N_0, geometric standard deviation σ_g, and d_{pm} bounds are used for both species. The total surface area concentration is the integral of the distribution and is given by Equation (2.9).

3.5 Synthetic mixed INP spectra

The generation of a mixed INP spectrum for each sample consists of three steps:

1. application of a per-sample perturbation to the INAS parameterisations;
2. addition of the two species contributions;
3. application of a pointwise spectrum perturbation.

The first perturbation changes the complete INAS curve coherently across the temperature range, whereas the pointwise perturbation is applied independently at each temperature after the species contributions have been combined.

Throughout this thesis, the term *INAS parameterisation perturbation* refers to the coherent per-sample amplitude and temperature-dependence modification applied separately to dust and Snomax. The term *pointwise spectrum perturbation* refers to the independent multiplicative factor applied at each temperature after the two species contributions have been combined. Measurement uncertainty is used only when discussing measured data, whereas prediction error refers to the difference between the predicted and true Snomax surface fractions.

Per-sample INAS parameterisation perturbation

The published INAS parameterisations provide the unperturbed fitted behaviour of each species. The measurements underlying these parameterisations nevertheless show variability between samples and experimental methods. For mineral dust, Ullrich et al. (Ullrich et al. 2017) report a relative uncertainty in the measured INAS density, while the Snomax intercomparison of Wex et al. (Wex et al. 2015) shows an inter-method spread in the mass-based ice nucleation activity. A controlled parameterisation perturbation is therefore applied so that the generated samples are not exact copies of the unperturbed fitted parameterisations.

The origin of the variability around the dust parameterisation is illustrated in Figure 3.6. The unperturbed fitted parameterisation represents the fitted central relationship but not the spread of the individual measurements.

Let $\bar{n}_{s,\mathrm{dust}}(T)$ denote the unperturbed fitted surface-area-based dust density of Equation (2.11), and let $\bar{n}_{m,\mathrm{snomax}}(T)$ denote the unperturbed fitted mass-based Snomax density of Equation (2.12). Following the general use of stochastic parameterisations to represent variability around a deterministic reference curve (Palmer et al. 2009; Berner et al. 2017), a strictly positive multiplicative perturbation is applied in natural-log space:

$$n^{(i)}_{s,\mathrm{dust}}(T) = \bar{n}_{s,\mathrm{dust}}(T) \exp\left[a_{\mathrm{dust},i} + b_{\mathrm{dust},i}(T - T_{\mathrm{ref}})\right], \tag{3.13a}$$

$$n^{(i)}_{m,\mathrm{snomax}}(T) = \bar{n}_{m,\mathrm{snomax}}(T) \exp\left[a_{\mathrm{snomax},i} + b_{\mathrm{snomax},i}(T - T_{\mathrm{ref}})\right]. \tag{3.13b}$$

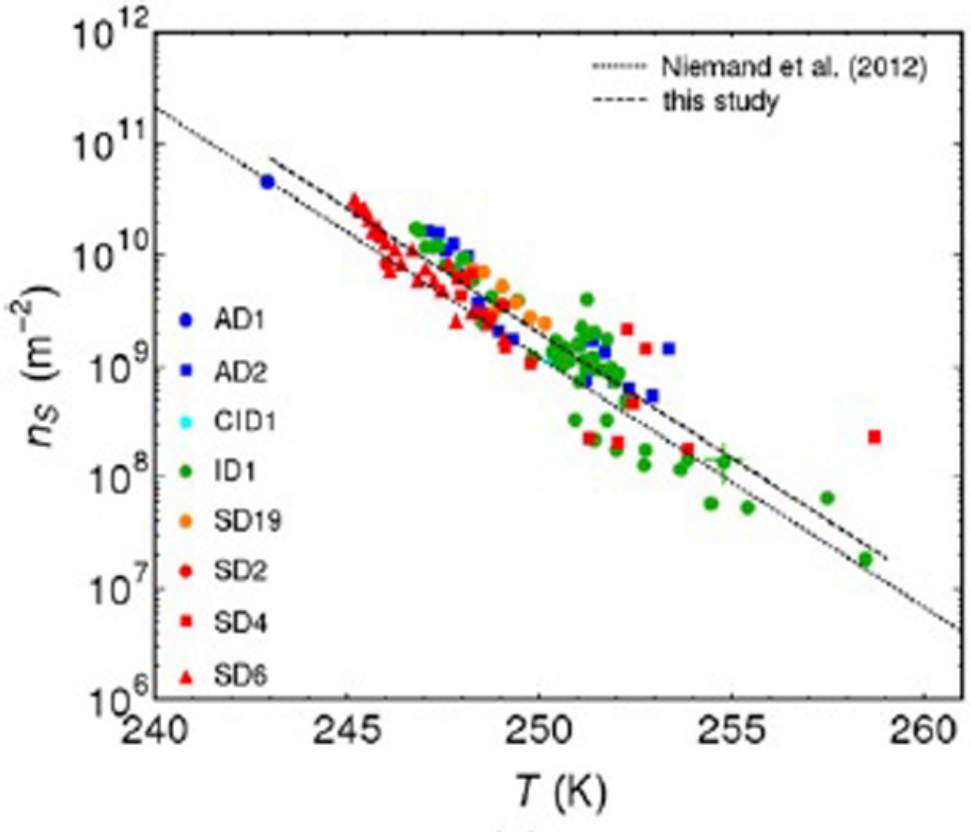


Figure 3.6: Dust INAS measurements and the unperturbed fitted parameterisation $\bar{n}_{s,\mathrm{dust}}(T)$. The displayed panel illustrates the spread of the measurements around the fitted curve. Adapted from Ullrich et al. (Ullrich et al. 2017).

Here, $a_{x,i}$ is the temperature-independent log-amplitude perturbation and $b_{x,i}$ controls how the multiplicative amplitude changes with temperature, where $x \in \{\mathrm{dust}, \mathrm{snomax}\}$. They are sampled as

$$a_{x,i} \sim \mathcal{N}(0, \sigma_{a,x}^2), \qquad b_{x,i} \sim \mathcal{N}(0, \sigma_{b,x}^2).$$

The standard deviation $\sigma_{a,x}$ is dimensionless because $a_{x,i}$ acts in natural-log space, whereas $b_{x,i}$ and $\sigma_{b,x}$ have units of K^{-1}. In Equations (3.13a) and (3.13b), T denotes the temperature on the common grid in kelvin and T_{ref} denotes the reference temperature in kelvin. Consequently, the product $b_{x,i}(T - T_{\mathrm{ref}})$ is dimensionless, as required for the argument of the exponential.

Because these perturbations have zero mean in natural-log space, their multiplicative factor has a median of one at each temperature. The unperturbed fitted parameterisation is therefore the pointwise median of the perturbed ensemble rather than its arithmetic mean.

The reference temperature is set to $T_{\mathrm{ref}} = -20\,^{\circ}\mathrm{C}$. It lies near the centre of the modelled temperature interval and within the temperature ranges used for both parameterisations. At this temperature, the contribution of the temperature-dependence term is zero by construction. The reference temperature therefore provides a fixed point around which the temperature-dependent modification is applied; it does not imply that ice nucleation activity and temperature dependence are physically independent.

For dust, Equation (3.13a) is algebraically equivalent to modifying the intercept and slope of its log-linear parameterisation. For Snomax, Equation (3.13b) modifies the mass-based density multiplicatively but does not correspond directly to a perturbation of the coefficients of the nonlinear Wex parameterisation. Its application to Snomax is therefore a modelling choice.

The log-amplitude standard deviations are set to

$$\sigma_{a,\mathrm{dust}} = \sigma_{a,\mathrm{snomax}} = 0.4.$$

Because the perturbation is applied in natural-log space, a positive one-standard-deviation amplitude draw multiplies the density by $e^{0.4} \approx 1.49$, while a negative one-standard-deviation draw multiplies it by $e^{-0.4} \approx 0.67$. This setting is of the same order as the approximately 40 % relative uncertainty reported for the dust INAS measurements by Ullrich et al. (Ullrich et al. 2017) and remains within the approximately factor-of-two inter-method spread reported for Snomax by Wex et al. (Wex et al. 2015).

The temperature-dependence standard deviations are set to

$$\sigma_{b,\mathrm{dust}} = 0.01\,\mathrm{K}^{-1}, \qquad \sigma_{b,\mathrm{snomax}} = 0.02\,\mathrm{K}^{-1}.$$

No published uncertainty is available for the temperature dependence of either parameterisation, so these values are modelling choices. For dust, a temperature difference of 10 K from the reference temperature corresponds to a one-standard-deviation contribution of 0.1 in natural-log space. In combination with the amplitude term, this changes the total log-space standard deviation from 0.4 to $\sqrt{0.4^2 + 0.1^2} \approx 0.41$. The larger value used for Snomax allows more variation in the temperature dependence of its steeper curve.

The effect of the two terms is illustrated in Figure 3.7. The log-amplitude term shifts the complete curve, while the temperature-dependence term changes its slope around T_{ref}. Applying both terms gives the complete parameterisation perturbation.

The draws are sampled independently for dust and Snomax. The two parameterisations are derived from separate experimental datasets (Ullrich et al. 2017; Wex et al. 2015), and neither study provides information about covariance between the variability of the two species. Treating the draws as independent is therefore a modelling assumption. These perturbations do not change the prescribed surface fraction R; they modify only the realised ice-nucleation activities of the two components. Consequently, samples with the same value of R can produce different apparent species contributions. This prevents an exact

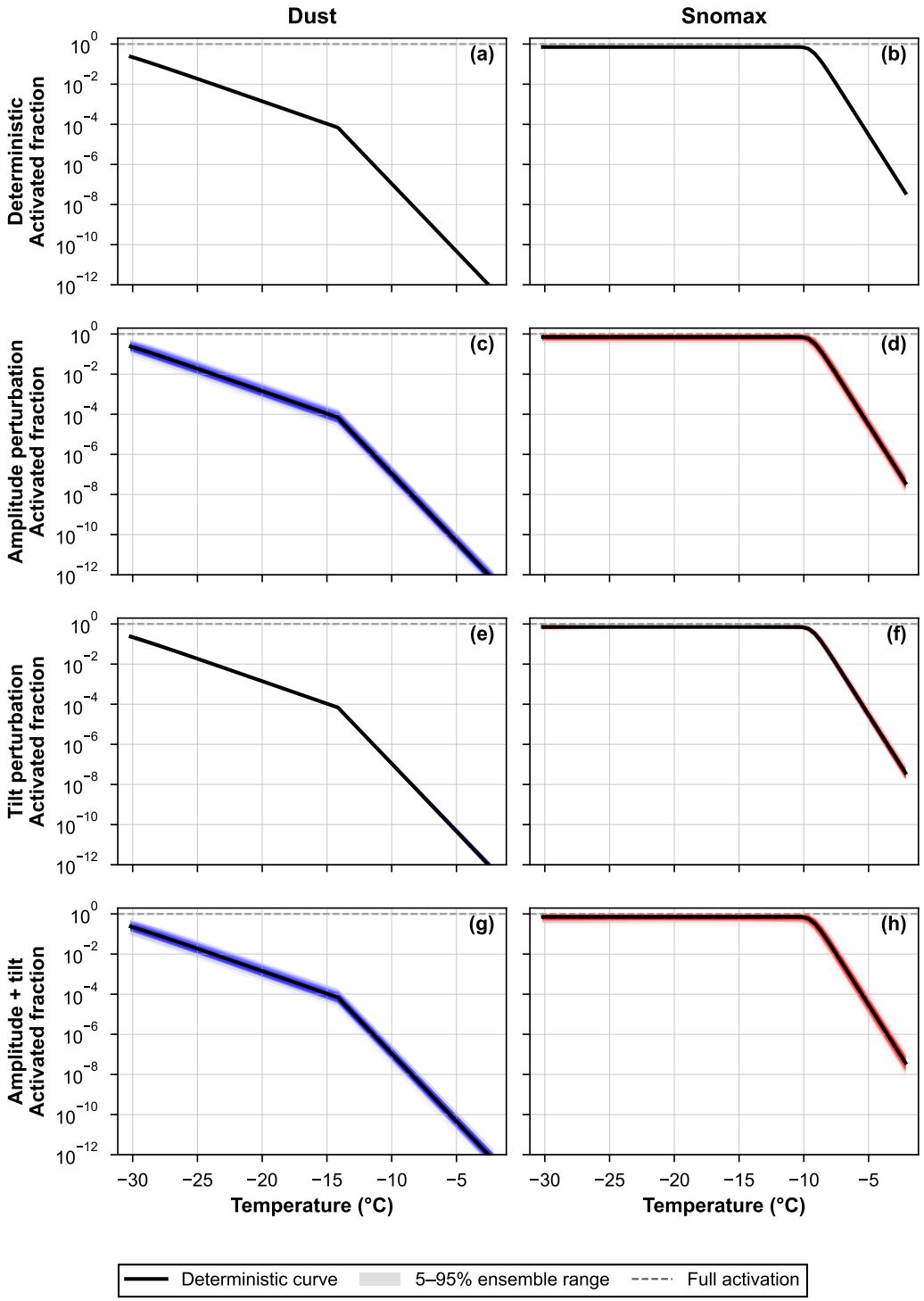


Figure 3.7: Decomposition of the per-sample INAS parameterisation perturbation for dust and Snomax. From top to bottom, the rows show the unperturbed fitted curve, the log-amplitude term, the temperature-dependence term, and their combined effect. The temperature-dependence term is zero at $T_{\mathrm{ref}} = -20\,^{\circ}\mathrm{C}$.

inversion of the mixture and makes R only partially recoverable from the spectral shape. The parameterisation-perturbed spectra are used to generate the input data for the machine-learning models. Spectra calculated from the unperturbed fitted parameterisations are retained as diagnostic outputs.

Mixed spectrum

The mixed INP concentration of sample i, before applying the pointwise spectrum perturbation, is the sum of the dust and Snomax contributions:

$$\begin{aligned} c^{(i)}(T) = &\frac{S_{\mathrm{dust},i}}{\bar{a}_{\mathrm{dust},i}} \left[1 - \exp\left(-n^{(i)}_{\mathrm{s,dust}}(T)\bar{a}_{\mathrm{dust},i}\right)\right] \\ &+ \frac{S_{\mathrm{snomax},i}}{\bar{a}_{\mathrm{snomax},i}} \left[1 - \exp\left(-f_m n^{(i)}_{m,\mathrm{snomax}}(T)\bar{a}_{\mathrm{snomax},i}\right)\right]. \end{aligned} \tag{3.14}$$

Here, f_m is the mass-to-surface conversion factor of Equation (3.3), and $\bar{a}_{\mathrm{dust},i}$ and $\bar{a}_{\mathrm{snomax},i}$ are the mean surface areas per particle. The two contributions have the same units, m^{-3}, and can therefore be added directly.

The frozen fractions and the additivity of the two species contributions are introduced in Chapter 2. The fixed particle number concentration N_0 controls the concentration scale, while the input descriptors used by the machine-learning models are constructed to be independent of this scale (Section 3.7).

Pointwise spectrum perturbation

After the two species contributions have been combined, a separate pointwise perturbation is applied independently at each temperature. This prescribed variability model is not fitted to a particular measurement instrument. It is used to represent unresolved temperature-by-temperature variability in a synthetic observed spectrum.

The perturbation is strictly positive, multiplicative and mean preserving:

$$\begin{aligned} \tilde{c}^{(i)}(T) &= c^{(i)}(T)\exp[\varepsilon_i(T)], \\ \varepsilon_i(T) &\sim \mathcal{N}\left(-\frac{1}{2}\sigma^2_{\log}, \sigma^2_{\log}\right), \qquad \sigma_{\log} = 0.5. \end{aligned} \tag{3.15}$$

The value $\sigma_{\log} = 0.5$ is a study-specific modelling choice rather than an uncertainty derived from a particular instrument or measurement dataset. It corresponds to a geometric multiplicative spread of $e^{0.5} \approx 1.65$ per log-space standard deviation. The mean $-\frac{1}{2}\sigma^2_{\log}$ ensures that

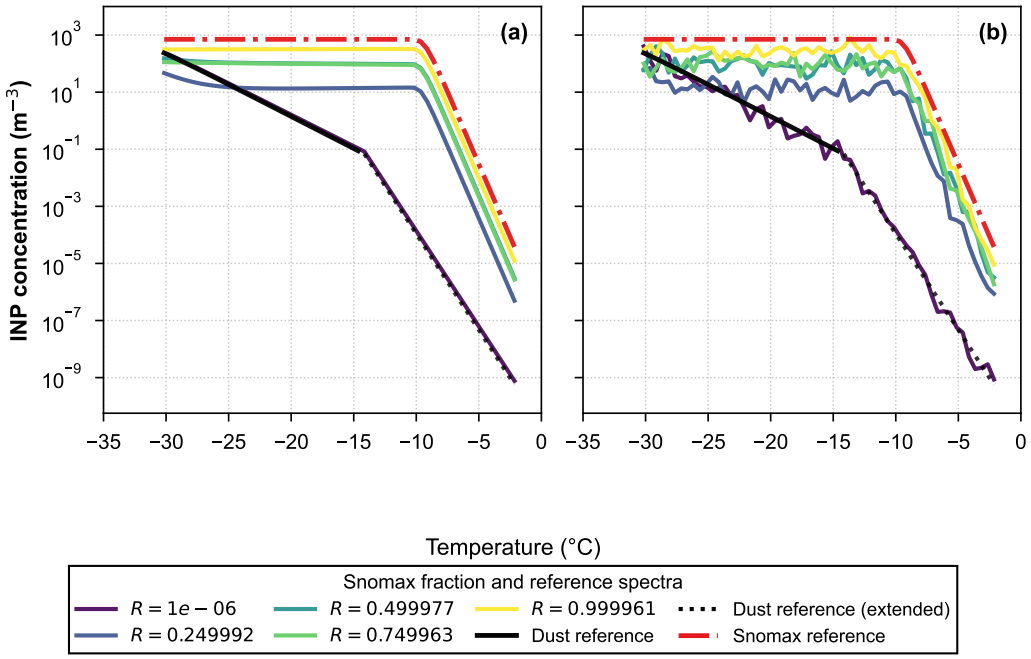


Figure 3.8: Representative mixed INP spectra before and after application of the pointwise spectrum perturbation. The per-sample INAS parameterisation perturbation is included in both panels. Panel (a) shows the spectra before the pointwise perturbation, whereas panel (b) shows the same spectra after the perturbation. The parameterisation perturbation changes each complete curve coherently across temperature, whereas the pointwise perturbation changes the individual temperature points.

$$\mathbb{E}[\exp(\varepsilon_i)] = 1,$$

so the perturbation does not change the expected concentration and every generated value remains positive. The resulting spectrum $\tilde{c}^{(i)}(T)$ is used for feature extraction.

For each sample, a second spectrum is calculated from the unperturbed fitted INAS parameterisations. The same pointwise factor $\exp[\varepsilon_i(T)]$ is applied to both the unperturbed-parameterisation spectrum and the parameterisation-perturbed spectrum. Both versions, the pointwise factor and the draws $a_{x,i}$ and $b_{x,i}$ are stored as diagnostic outputs. The detectability statistic of Section 3.6 is evaluated using the parameterisation-perturbed spectrum before the pointwise perturbation is applied.

The two perturbations serve different purposes. The parameterisation perturbation is described by two values per species and acts coherently across the temperature axis. The pointwise spectrum perturbation uses one independent draw at each temperature. Their separate effects are illustrated in Figure 3.8.

Reference spectra

The reference surface areas are defined in Section 3.4. Using these values, two single-species reference spectra are calculated once from the unperturbed fitted INAS parameterisations, without either the parameterisation perturbation or the pointwise spectrum perturbation:

$$c^{\mathrm{ref}}_{\mathrm{dust}}(T) = \frac{S^{\mathrm{ref}}_{\mathrm{dust}}}{\bar{a}^{\mathrm{ref}}_{\mathrm{dust}}} \left[1 - \exp\left(-\bar{n}_{s,\mathrm{dust}}(T)\,\bar{a}^{\mathrm{ref}}_{\mathrm{dust}}\right)\right], \tag{3.16}$$

$$c^{\mathrm{ref}}_{\mathrm{snomax}}(T) = \frac{S^{\mathrm{ref}}_{\mathrm{snomax}}}{\bar{a}^{\mathrm{ref}}_{\mathrm{snomax}}} \left[1 - \exp\left(-f_m \bar{n}_{m,\mathrm{snomax}}(T)\,\bar{a}^{\mathrm{ref}}_{\mathrm{snomax}}\right)\right]. \tag{3.17}$$

These spectra provide the fixed reference shapes used when the synthetic spectra are compared during feature extraction.

3.6 Detectability and regime classification

The parameterisation-perturbed mixed INP spectrum contains contributions from dust and Snomax. When one species contributes only a small fraction of the total signal, its contribution may fall below the prescribed pointwise variability of Section 3.5. A signal-to-noise detection statistic is therefore used to determine whether each species is detectable within a given synthetic spectrum.

Throughout this section, the overbar in $\bar{n}_{s,\mathrm{dust}}(T)$ and $\bar{n}_{m,\mathrm{snomax}}(T)$ denotes the unperturbed fitted parameterisation value at temperature T. It does not denote an average over temperature.

Per-sample detection statistic and regimes

The fraction of the mixed INP spectrum contributed by species x at temperature T is

$$g_x(T) = \frac{c_x(T)}{c(T)}, \tag{3.18}$$

where $x \in \{\mathrm{dust}, \mathrm{snomax}\}$, $c_x(T)$ is the contribution of species x, and $c(T)$ is the total mixed INP concentration. Removing species x from the spectrum changes the logarithmic concentration by

$$\delta_x(T) = \ln c(T) - \ln[c(T) - c_x(T)] = -\ln[1 - g_x(T)]\,. \tag{3.19}$$

Table 3.2: Criteria used to assign each generated spectrum to one of four detectability regimes. The classification is evaluated on the parameterisation-perturbed spectrum before the pointwise spectrum perturbation is applied, using $k = 5$.

Regime	$\mathrm{SNR}_{\mathrm{dust}}$	$\mathrm{SNR}_{\mathrm{snomax}}$
mixed	≥ 5	≥ 5
dust-dominated	≥ 5	< 5
Snomax-dominated	< 5	≥ 5
undetectable	< 5	< 5

The pointwise spectrum perturbation has the log-space standard deviation $\sigma_{\log}$ at each temperature (Equation (3.15)). The detection statistic is therefore defined as

$$\mathrm{SNR}_x = \sqrt{\sum_T \left(\frac{\delta_x(T)}{\sigma_{\log}} \right)^2}. \tag{3.20}$$

The sum combines the standardized logarithmic contribution of the species across the temperature grid. A species is classified as detectable when

$$\mathrm{SNR}_x \geq k. \tag{3.21}$$

The threshold is set to $k = 5$ in this study. This is a conservative modelling choice rather than a value derived from a specific instrument or an external standard. It requires the combined contribution across the temperature grid to reach five times the prescribed log-space variability scale.

No sensitivity analysis was performed for alternative values of k. A lower threshold would classify more weak species contributions as detectable and would enlarge the mixed regime. A higher threshold would retain fewer samples with more clearly detectable contributions. Changing k would therefore change the regime counts, the composition range included in the modelling dataset, and potentially the reported model performance.

Each sample is assigned to one of four regimes according to the two detectability criteria. The criteria are summarised in Table 3.2.

Figure 3.9 shows how the two detection statistics change with composition, while Figure 4.4 shows the corresponding classification in the plane of the two SNR values.

Representative spectra from the different regimes are shown in Figure 4.5.

The regime label, the two detection statistics and the two detectability flags are retained as per-sample metadata. The label defines which samples are carried into the modelling dataset, as described in Section 3.6. It is not written into the feature table and is not used as a predictor. This prevents direct use of the regime label by the machine-learning models,

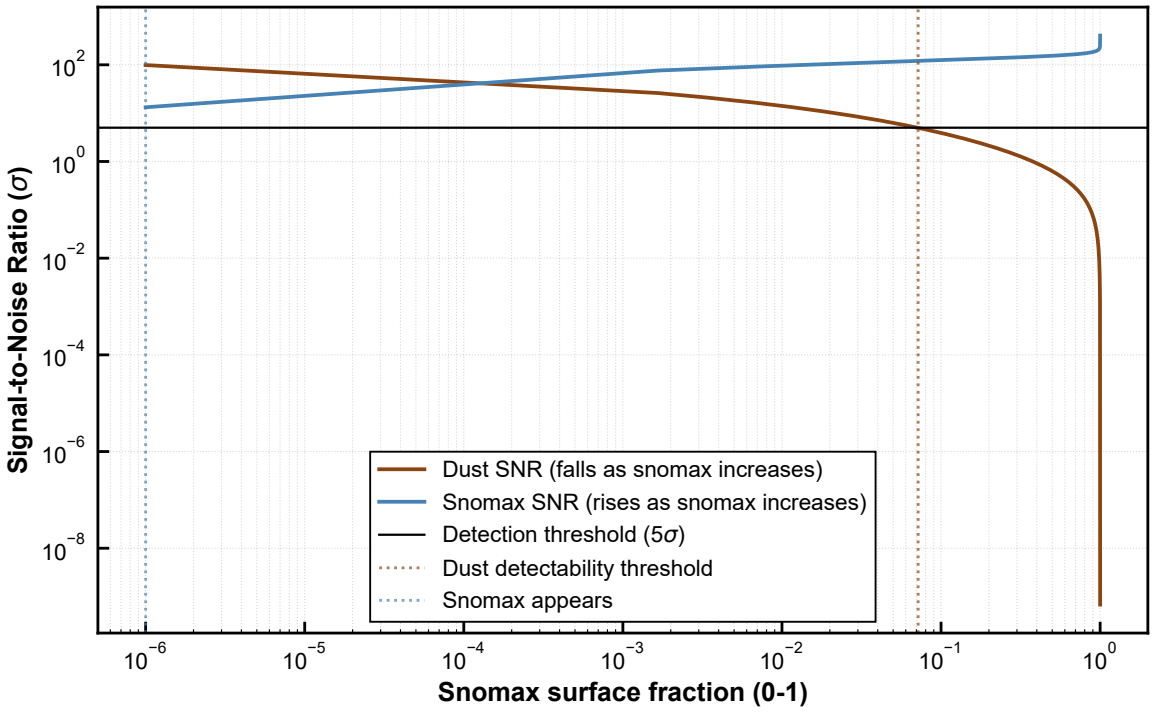


Figure 3.9: Detection statistic of each species as a function of the Snomax surface fraction, evaluated using the unperturbed fitted INAS parameterisation curves. The dust statistic decreases and the Snomax statistic increases as the composition shifts towards Snomax. The horizontal line shows the threshold $k = 5$.

although filtering the dataset by regime still restricts the population on which the models are trained and evaluated.

Dust-detectability limit and dataset restriction

The same detection statistic is used to determine the maximum Snomax surface fraction at which the dust contribution remains detectable. For this diagnostic, the dilute-activation approximation

$$1 - \exp\left[-n_s(T)\bar{a}\right] \simeq n_s(T)\bar{a}, \qquad n_s(T)\bar{a} \ll 1,$$

is applied to the unperturbed fitted parameterisations. At the largest sampled number-median diameter, the mean particle surface area is

$$\bar{a}_{\max} = \pi(5.0 \times 10^{-8}\,\mathrm{m})^2 \exp\left[2\ln^2(1.5)\right] \approx 1.09 \times 10^{-14}\,\mathrm{m}^2.$$

The corresponding maximum dimensionless activation terms of the unperturbed fitted parameterisations are approximately

$$\max_T \left[\bar{n}_{s,\mathrm{dust}}(T)\bar{a}_{\max}\right] \approx 7.4 \times 10^{-4},$$

and

$$\max_T \left[f_m \bar{n}_{m,\mathrm{snomax}}(T)\bar{a}_{\max}\right] \approx 3.4 \times 10^{-3}.$$

Both values are much smaller than one, supporting the use of the dilute-activation approximation for this diagnostic. Under this approximation, the dust fractional contribution at Snomax surface fraction R is

$$g^{\mathrm{lin}}_{\mathrm{dust}}(T;R) \simeq \frac{(1-R)\bar{n}_{s,\mathrm{dust}}(T)}{(1-R)\bar{n}_{s,\mathrm{dust}}(T) + f_m R \bar{n}_{m,\mathrm{snomax}}(T)}. \tag{3.22}$$

where $\bar{n}_{s,\mathrm{dust}}(T)$ and $\bar{n}_{m,\mathrm{snomax}}(T)$ are the unperturbed fitted parameterisation values of Ullrich et al. (Ullrich et al. 2017) and Wex et al. (Wex et al. 2015), respectively. The factor f_m is the Snomax mass-to-surface conversion factor defined in Equation (3.3).

The cancellation in Equation (3.22) concerns the common total surface-area concentration, not the two INAS densities. Substituting

$$S_{\mathrm{dust}} = (1-R)S_{\mathrm{tot}}, \qquad S_{\mathrm{snomax}} = RS_{\mathrm{tot}}$$

into the two species contributions introduces the common factor S_{tot} in the numerator and denominator of their ratio. Dividing both by this factor gives Equation (3.22). Under this formulation, the dust fraction depends on R and the two parameterisation values but not on the numerical value of S_{tot}.

Using the dilute-limit estimate of Equation (3.22) in Equation (3.20) gives the dust detection statistic as a function of composition:

$$\mathrm{SNR}_{\mathrm{dust}}(R) = \sqrt{\sum_T \left[\frac{-\ln\left(1 - g^{\mathrm{lin}}_{\mathrm{dust}}(T;R)\right)}{\sigma_{\log}}\right]^2}. \tag{3.23}$$

The statistic decreases as R increases because the contribution of Snomax becomes larger relative to that of dust.

The dust-detectability threshold R^* is defined as the largest Snomax surface fraction on a grid of 600 equally spaced values over $[0, 1]$ for which the dust detection criterion remains satisfied:

$$\mathrm{SNR}_{\mathrm{dust}}(R^*) \geq k, \qquad k = 5. \tag{3.24}$$

The grid spacing is 1/599, and R^* is reported at this numerical resolution. For $R > R^*$, the dust contribution does not satisfy the adopted detectability criterion. The behaviour around this limit is shown in Figure 4.2.

The dust-detectability threshold and the per-sample regime classification have different roles. The dust-detectability threshold is a diagnostic calculated once using the unperturbed fitted parameterisations. It identifies the composition above which the dust contribution does not satisfy the adopted criterion for the specified value of $\sigma_{\log}$. It is not used directly to remove samples.

Every one of the $N_s = 100\,000$ generated samples is passed to the spectrum synthesis of Section 3.5 and is classified using its own parameterisation-perturbed spectrum. The classification is therefore applied to the complete ensemble before any dataset restriction.

The dataset restriction is applied later, during feature extraction. Only samples classified as *mixed*, for which both dust and Snomax satisfy the $k = 5$ criterion, are carried into the feature table. The dust-dominated, Snomax-dominated and undetectable samples are excluded.

This restriction makes the regression conditional on spectra in which both species satisfy the adopted detectability criterion. It does not mean that the excluded spectra contain only one species; a second component can be present but remain below the threshold. The filtering therefore introduces selection bias by removing spectra with weak species contributions. It may make the regression problem easier and prevents the reported performance from being interpreted as performance over the full ensemble or over unrestricted atmospheric observations.

The reported model metrics must consequently be interpreted as conditional on the mixed regime and on the choice $k = 5$. The present model does not establish whether subthreshold biological or dust contributions can be estimated reliably from dust-dominated or Snomax-dominated spectra. Applying this approach to unrestricted measurements would require either a preceding detectability assessment or training and evaluation over the other regimes as well.

Because every retained sample has the same regime label, the label is constant within the modelling dataset and is not included as a predictor. This prevents direct regime-label leakage into the regression but does not remove the selection bias caused by the restriction itself. The regime-resolved diagnostics described in Sections 3.10 and 3.11 are therefore inactive in the configuration used here.

3.7 Feature engineering

A feature is an individual measurable property or characteristic of a data sample (Bishop 2006). In this study, it is a scalar descriptor calculated from one synthetic observed INP

spectrum $\tilde{c}(T)$ after the parameterisation perturbation and the pointwise spectrum perturbation have been applied. Feature extraction reduces each spectrum to a fixed set of scalar descriptors.

The synthetic INP spectrum has two relevant properties:

- The overall magnitude depends on the total surface-area concentration of the aerosol population. Increasing the total surface-area concentration scales the spectrum upwards, while decreasing it scales the spectrum downwards. This quantity is not the prediction target and varies independently between samples.
- The spectral shape contains information about the composition of the aerosol population, which is the prediction target.

The features are therefore constructed to be scale invariant. Multiplying the complete spectrum by a positive constant does not change a scale-invariant feature. Such features describe the spectral shape rather than the absolute concentration.

The features are divided into four groups:

1. shape similarity to the reference spectra;
2. non-negative decomposition onto the reference shapes;
3. freezing-density distribution;
4. log-spectrum slopes and onset.

Shape similarity to the reference spectra. Cosine similarity measures the similarity between two vectors independently of their magnitudes (Manning et al. 2008):

$$\cos\theta(\mathbf{a},\mathbf{b}) = \frac{\mathbf{a}\cdot\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}. \tag{3.25}$$

The cosine similarity between each synthetic spectrum and each of the two reference spectra of Equations (3.16) and (3.17) is calculated. The calculation is performed both on the linear spectrum and on the mean-centred base-ten logarithm of the spectrum. Mean centring removes a constant shift in log space, so these features capture similarity in spectral shape rather than differences in magnitude.

Non-negative decomposition onto the reference shapes. Each synthetic spectrum can also be represented as a non-negative linear combination of the two reference shapes. Let

$$\mathbf{A} = \begin{bmatrix} c^{\mathrm{ref}}_{\mathrm{dust}} & c^{\mathrm{ref}}_{\mathrm{snomax}} \end{bmatrix}$$

be the matrix containing the reference spectra. The non-negative coefficients $\boldsymbol{\gamma}$ are obtained using non-negative least squares (Lawson and Hanson 1995):

$$\boldsymbol{\gamma}^{*} = \arg\min_{\boldsymbol{\gamma} \geq 0} \left\| \mathbf{A}\boldsymbol{\gamma} - \tilde{\mathbf{c}} \right\|_2 . \tag{3.26}$$

The two fitted coefficients are multiplied by the integrated area of their corresponding reference shapes and normalised so that the resulting dust and Snomax contribution fractions sum to one. This normalisation makes the fractions scale invariant. The relative reconstruction residual

$$r_{\mathrm{NNLS}} = \frac{\left\| \tilde{\mathbf{c}} - \mathbf{A}\boldsymbol{\gamma}^{*} \right\|_2}{\left\| \tilde{\mathbf{c}} \right\|_2} \tag{3.27}$$

is also retained as a feature.

Freezing-density distribution. The temperatures at which increments in the cumulative INP concentration occur may contain information about the composition. To describe this behaviour independently of the overall concentration, the spectrum is converted into a normalised per-degree freezing-increment distribution.

The differential form of a cumulative freezing spectrum describes the number of freezing events occurring over a temperature interval (Vali 1971). Based on this relation, the normalised freezing-density feature used here is defined as

$$p(T) = \frac{\max(-\mathrm{d}\tilde{c}/\mathrm{d}T,\ 0)}{\sum_{T} \max(-\mathrm{d}\tilde{c}/\mathrm{d}T,\ 0)}. \tag{3.28}$$

Only positive freezing increments are retained, and the distribution is normalised so that it sums to one. It therefore describes where freezing occurs along the temperature axis rather than the total number of activated particles.

The following descriptors are calculated from $p(T)$:

- weighted mean and standard deviation of temperature;
- Shannon entropy;
- the 10th, 50th and 90th temperature quantiles;

- the temperature of the distribution peak;
- the fraction of the distribution at temperatures higher than the fixed split at −15°C.

The Shannon entropy is calculated as (Bishop 2006)

$$H[p] = -\sum_{T} p(T) \ln p(T). \tag{3.29}$$

It measures how broadly the freezing increments are distributed across temperature. A distribution concentrated in a small number of temperature intervals has low entropy, whereas a distribution spread more evenly across the temperature range has high entropy.

Log-spectrum slopes and onset. The slope of $\log_{10} \tilde{c}(T)$ against temperature is used to describe the temperature dependence of the spectrum. The logarithm is applied because the INP concentration spans several orders of magnitude across the temperature range. A slope fitted directly to the untransformed concentration would be dominated by the largest values, whereas the approximately exponential temperature dependence of the parameterisations becomes closer to linear in log space (Ullrich et al. 2017; Wex et al. 2015).

A linear regression is fitted to the log-transformed spectrum over the complete temperature grid (Bishop 2006). Separate slopes are also fitted to the high temperature and low temperature parts of the spectrum. The division at −15°C is a methodological choice based on the different temperature ranges in which dust and Snomax are active, as discussed in Section 2.7. The ratio of the high temperature segment slope to the low temperature segment slope is retained as a descriptor of the change in spectral slope between these two temperature regions.

The onset temperature is defined as the highest temperature at which the spectrum exceeds 10^{-3} of its maximum value. The value 10^{-3} is a methodological threshold used consistently for all samples. Because Snomax is active at higher temperatures than mineral dust, the onset temperature provides information about the presence of a high temperature active contribution (Ullrich et al. 2017; Wex et al. 2015).

In total, a pool of eighteen features is generated, as summarised in Table 3.3.

3.8 Feature selection

Before the selection, the reference-shape features are removed from the candidate pool. These are the cosine similarities to the dust and Snomax references (in linear and in log space) and the non-negative decomposition features (the dust and Snomax contribution fractions and the reconstruction residual). These features are computed against the exact dust and Snomax reference shapes which are used to generate the spectra, so they read the composition almost directly, like an analytic inversion of the synthetic spectrum generation framework, and the surface fraction becomes almost directly readable from them. This is a

Table 3.3: Engineered spectral features. All features are scale invariant.

Group	Feature
Shape similarity	cosine similarity to dust reference
	cosine similarity to Snomax reference
	log-space cosine similarity to dust reference
	log-space cosine similarity to Snomax reference
NNLS decomposition	dust contribution fraction
	Snomax contribution fraction
	relative reconstruction residual
Freezing density	weighted mean temperature
	weighted standard deviation of temperature
	Shannon entropy of freezing density
	10th temperature quantile (T_{10})
	50th temperature quantile
	90th temperature quantile
	peak temperature
	high temperature side fraction ($T \geq -15°C$)
Slopes and onset	log-spectrum slope
	high temperature to low temperature slope ratio
	onset temperature

conceptual leakage: on a real spectrum the exact reference shapes are not available, so the performance gained from these features would not transfer to real data. For this reason these seven features are excluded, and only the shape-intrinsic features are kept as candidates, which leaves eleven features (the freezing-density and the slope/onset groups).

The eleven remaining candidates and their feature labels are defined in Section 3.7 and listed in Table 3.4. Unlike the excluded reference-shape features, they are calculated directly from a single INP spectrum and do not require the pure dust or Snomax reference curves. The freezing-density features are calculated from a normalised distribution, while the slope and onset features describe changes in the spectrum with temperature. They are therefore independent of the absolute aerosol loading and can, in principle, be calculated from measured spectra.

However, their transferability to measurements is not demonstrated by the synthetic dataset. The slope features depend on the available temperature range and require sufficient data on both sides of the split at −15 °C. The freezing-density features may be affected by temperature resolution, measurement uncertainty, detection limits and the numerical derivative used to calculate the freezing increments. Their performance must therefore be tested using measured spectra, as discussed further in Section 5.2.

In the purpose of selecting the best set of features among these eleven features, these steps are taken. For the selection process, only the training set is being used, so the validation and test splits are not looked at during the selection.

Table 3.4: Meanings and dataset labels of the eleven shape-intrinsic candidate features.

Meaning	Feature label
Weighted mean freezing temperature	`density_mean_T`
Weighted standard deviation of freezing temperature	`density_std_T`
Shannon entropy of the freezing-density distribution	`density_entropy`
10th temperature quantile	`t10`
Median temperature	`t50`
90th temperature quantile	`t90`
Temperature of the freezing-density peak	`peak_T`
Fraction of freezing occurring on the high temperature side	`warm_temperature_fraction`
Slope of the full log-spectrum	`slope_log`
Ratio of the high temperature- and low temperature-side slopes	`slope_warm_cold_ratio`
Onset temperature	`onset_temperature`

- First, the features with near-zero variance are dropped. A feature that is almost constant over the samples carries almost no information to separate one sample from another, so it cannot help the model to distinguish different target values and is removed (variance below 10^{-8}).
- Second, the remaining features are tested for correlation. All pairs whose absolute Pearson correlation exceeds 0.90 are collected and processed in order of decreasing correlation. For each pair still under consideration, the member less correlated with the target is removed and the other is kept; a pair is skipped if one of its members has already been removed by an earlier, more strongly correlated pair. This greedy ordering matters, because a feature that is kept in one pair can still be removed later in a more weakly correlated pair.
- Third, the remaining features are used to train an XGBoost model (Chen and Guestrin 2016). XGBoost is chosen because potentially nonlinear relationships between the features and the target can be represented by a gradient-boosted tree model. In addition, TreeSHAP provides an exact and computationally efficient method for calculating SHAP values for tree-based models (Lundberg et al. 2020).
- Fourth, the remaining features are ranked by their mean absolute SHAP value (Lundberg and Lee 2017). SHAP measures the contribution of each feature to the model predictions. The ranking is checked using permutation importance and bootstrap resampling. In bootstrap resampling, repeated datasets are created by randomly sampling observations from the training set with replacement (Efron 1979). The SHAP ranking is recalculated for each bootstrap sample, and the variation in the rank of each feature is recorded. A small variation indicates that the feature ranking is not strongly dependent on the particular training samples used. The results of these checks are presented in Figure 4.9.

The SHAP step by itself outputs the full list of the cleaned features ranked by SHAP, without selecting a fixed number of top-ranked features. The reduction of the number of features is done in a later step: the ranked features are added one by one in SHAP order, and the models are trained and evaluated for each feature count. A cheap pair of proxy models, one linear and one gradient-boosted, is used for this sweep, so that the curve of error against feature count is obtained without training the whole model suite at every count. The feature count is then fixed in the configuration and the full model suite is trained only at that count. The selected feature count and the corresponding performance curve are reported in Section 4.4.

3.9 Dataset partitioning

From the raw data, the feature tables are generated and split into training, validation, and test subsets with proportions of 0.8, 0.1, and 0.1 respectively. Due to the fact that the target is continuous and cannot be stratified on exact values, to ensure that the three subsets carry the same distribution of composition, the target R is cut into quantile bins (twenty bins, each holding a comparable number of samples), and the split is performed proportionally *within* each bin, so validation and test performance are representative across the whole range of R. All random operations use a fixed seed for reproducibility.

Only the target column and the feature columns listed in the configuration are written into the split files; any other column of the feature table, including the detectability metadata, is dropped at this stage. The candidate pool carried forward is therefore exactly the eleven shape-intrinsic features of Section 3.8.

3.10 Regression models and training protocol

From classification to regression

Most existing applications of machine learning in aerosol science are classification problems, in which particles are assigned to discrete types, typically using single-particle mass spectra (Christopoulos et al. 2018; G. Wang et al. 2024). The problem considered here is instead a regression problem: the continuous Snomax surface fraction must be recovered from a mixed synthetic INP spectrum. Although the forward relation of Equation (3.4) is specified, the mapping from the resulting spectrum back to the composition is made ill conditioned by the per-sample INAS parameterisation perturbation and the pointwise spectrum perturbation. A direct analytic inversion is therefore not used.

Shared training pipeline

The prediction target is the Snomax surface fraction R. The regression models do not use the raw spectrum directly; they use the selected shape-intrinsic features extracted from each perturbed synthetic mixed spectrum. These features represent both linear and nonlinear relationships with the target. All candidate models are trained and evaluated under the common protocol described below.

On the stratified split of Section 3.9, the hyperparameters are tuned using grid search: each configuration is fitted on the training subset and ranked based on RMSE on the validation subset. After the best hyperparameter configuration for each model family is fixed, the models are compared and selected using cross-validation on the combined training-and-validation pool and the three-step rule of Section 3.11. The test subset is not used during hyperparameter tuning or model selection. After the selection decision is fixed, full test-set metrics are calculated. The test RMSE values of the models closest to the one-standard-error band are also calculated as a post-selection diagnostic, but they are not used to revise the selection decision.

The feature-count sweep uses the two proxy models described in Section 3.8. Features are added sequentially according to their SHAP ranking, and the performance of the linear and gradient-boosted proxy models is evaluated at each feature count. After the feature count is fixed, the complete model suite is trained using only the selected feature subset.

Feature scaling is applied to the scale-sensitive estimators, including the linear and regularised-linear models, polynomial ridge regression, k-nearest neighbours, MLP and DNN. The mean and standard deviation used for standardisation are estimated from the relevant training data only. During cross-validation, the transformation is refitted independently within every training fold and then applied to the corresponding held-out fold, preventing information leakage. Tree-based models use the unscaled features because their threshold-based splits are invariant to monotonic rescaling.

Model families

For the purpose of examining different supervised learning algorithms, including simple linear, neighbour-based, tree-based, and neural network models, a total of 11 models were examined in this study. These models and their corresponding hyperparameters are presented in Table 3.5. The models were selected to compare complementary model families with different assumptions and levels of flexibility. Linear and regularised linear models (ordinary least squares, ridge, lasso, and elastic net) provide interpretable baselines, while polynomial ridge regression and k-nearest neighbours allow nonlinear relationships to be represented. Tree ensembles (random forest (Breiman 2001), gradient boosting (Friedman 2001), XGBoost (Chen and Guestrin 2016), and LightGBM (Ke et al. 2017)) capture nonlinear structure and interactions between features. The neural network family includes multilayer perceptrons and deep neural networks (DNN), which provide flexible function approximators (Hornik et al. 1989; Goodfellow et al. 2016).

Table 3.5: Regression models and the principal hyperparameters searched on the validation subset.

Model	Principal hyperparameters searched
Linear / Ridge	regularisation strength $\alpha \in \{0, 10^{-3}, 10^{-2}, 0.1, 1, 10, 100\}$ (ordinary least squares when $\alpha = 0$)
Lasso	$\alpha \in \{10^{-4}, 10^{-3}, 10^{-2}, 0.1, 0.5, 1, 5, 10\}$
Elastic net	$\alpha \in \{10^{-4}, 10^{-3}, 10^{-2}, 0.1, 1, 10\}$; ℓ_1 ratio $\in \{0.1, 0.3, 0.5, 0.7, 0.9\}$
Polynomial (ridge)	degree $\in \{2, 3\}$; $\alpha \in \{10^{-3}, 10^{-2}, 0.1, 1, 10, 100\}$
k-NN	neighbours $\in \{3, 5, 7, 10, 15, 25\}$; weights $\in$ {uniform, distance}; $p \in \{1, 2\}$
Gradient boosting	trees $\in \{100, 200, 300\}$; learning rate $\in \{0.01, 0.05, 0.1\}$; depth $\in \{2, 3, 4\}$; subsample $\in \{0.8, 1.0\}$
Random forest	trees $\in \{100, 200, 300, 500, 1000\}$; max. depth $\in \{10, 20, 30, 50, 100\}$; min. samples per leaf $\in \{5, 10, 15, 20, 50\}$
XGBoost	trees $\in \{100, 200, 300\}$; learning rate $\in \{0.01, 0.05, 0.1\}$; depth $\in \{2, 3, 4\}$; subsample $\in \{0.8, 1.0\}$; column subsample $\in \{0.8, 1.0\}$; $\lambda \in \{1.0, 10.0\}$
LightGBM	trees $\in \{100, 200, 300\}$; learning rate $\in \{0.01, 0.05, 0.1\}$; leaves $\in \{15, 31, 63\}$; depth $\in \{-1, 5, 10\}$; subsample $\in \{0.8, 1.0\}$; column subsample $\in \{0.8, 1.0\}$; $\lambda \in \{0, 1.0, 10.0\}$
MLP	hidden-layer widths $\in \{(32), (64), (128), (64, 32), (128, 64)\}$; activation $\in$ {ReLU, tanh}; $\alpha \in \{10^{-5}, 10^{-4}, 10^{-3}\}$; initial learning rate $\in \{10^{-3}, 10^{-2}\}$
Deep NN (Keras)	hidden-layer widths $\in \{(512, 256, 128, 64), (1024, 512, 256, 128, 64, 32)\}$; dropout rate $\in \{0.1, 0.2\}$; learning rate $\in \{0.001, 0.0005\}$; batch size $\in \{32, 64, 128\}$

The DNN hyperparameter search considered two fully connected architectures: one with four hidden layers containing 512, 256, 128, and 64 units, and one with six hidden layers containing 1024, 512, 256, 128, 64, and 32 units. Dropout rates of 0.1 and 0.2 were tested, meaning that 10% or 20% of the units were randomly deactivated during training to reduce overfitting. The general principles of neural networks and dropout regularisation are described in Chapter 2. The DNN was implemented in TensorFlow (Abadi et al. 2016).

Evaluation metrics

To evaluate model performance in reproducing the target, the following metrics are used: mean absolute error (MAE), root-mean-square error (RMSE), coefficient of determination (R^2), mean signed error (bias), maximum absolute error, and the fraction of samples whose absolute error falls within a tolerance of 0.05.

A composition-resolved breakdown supplements the overall metrics: the test samples are binned into twenty equal-width bins over the observed range of R and the error is reported per bin, which reveals how the accuracy varies with composition rather than only on average.

The evaluation code additionally supports a regime-resolved breakdown, which would report the same errors separately per detectability regime and flag a model as suspect if its error were unexpectedly low in a regime where the composition is not recoverable in principle. It is not produced in the configuration used here, for the reason given in Section 3.6. The composition-resolved breakdown plays the corresponding role instead: within the mixed regime it shows how the error develops as the composition approaches the boundary at which the dust contribution falls below the adopted detectability criterion.

Cross-validation and stability

The robustness of each fitted model is evaluated by k-fold cross-validation on the combined training-and-validation pool, from which the mean and standard deviation of the cross-validated RMSE and R^2 are recorded. The same folds also supply the overfitting gap used as the first selection criterion, which is defined in Section 3.11.

The $k = 5$ used here, which is a standard choice for regression datasets that balances bias and variance (Abedin et al. 2026). In our case, the combined training-and-validation pool comprises 90% of the data, so each fold holds out approximately 20% of this pool for testing, leaving the remaining 80% for training within the fold; this is large enough for a stable error estimate. The folds are taken as contiguous blocks of the pool in its stored order, without shuffling, and identically for every model. This keeps the comparison between models exact, since each model sees the same five partitions, and it makes the cross-validation sensitive to any dependence of a model on the row ordering of the data, which is used as a diagnostic in the discussion of the polynomial model. The deep neural network instead uses a lightweight 3-fold cross-validation with reduced training epochs (150 versus 500 for the final fit) to manage its computational cost; because model selection relies on the one-standard-error rule (Section 3.11) rather than on point estimates, the slightly noisier estimate is acceptable.

3.11 Model selection

The test subset is kept hidden during hyperparameter tuning and during the three-step model-selection procedure. The final model is selected using the cross-validation results and the predefined simplicity ranking, without reference to test-set performance. Only after the selection decision has been fixed is the test subset evaluated. Full test metrics are reported for the selected model, while the test RMSE values of the models closest to the one-standard-error band are shown only as a post-selection diagnostic and do not affect the selected model. This procedure is illustrated in Figure 3.10.

This procedure includes these steps:

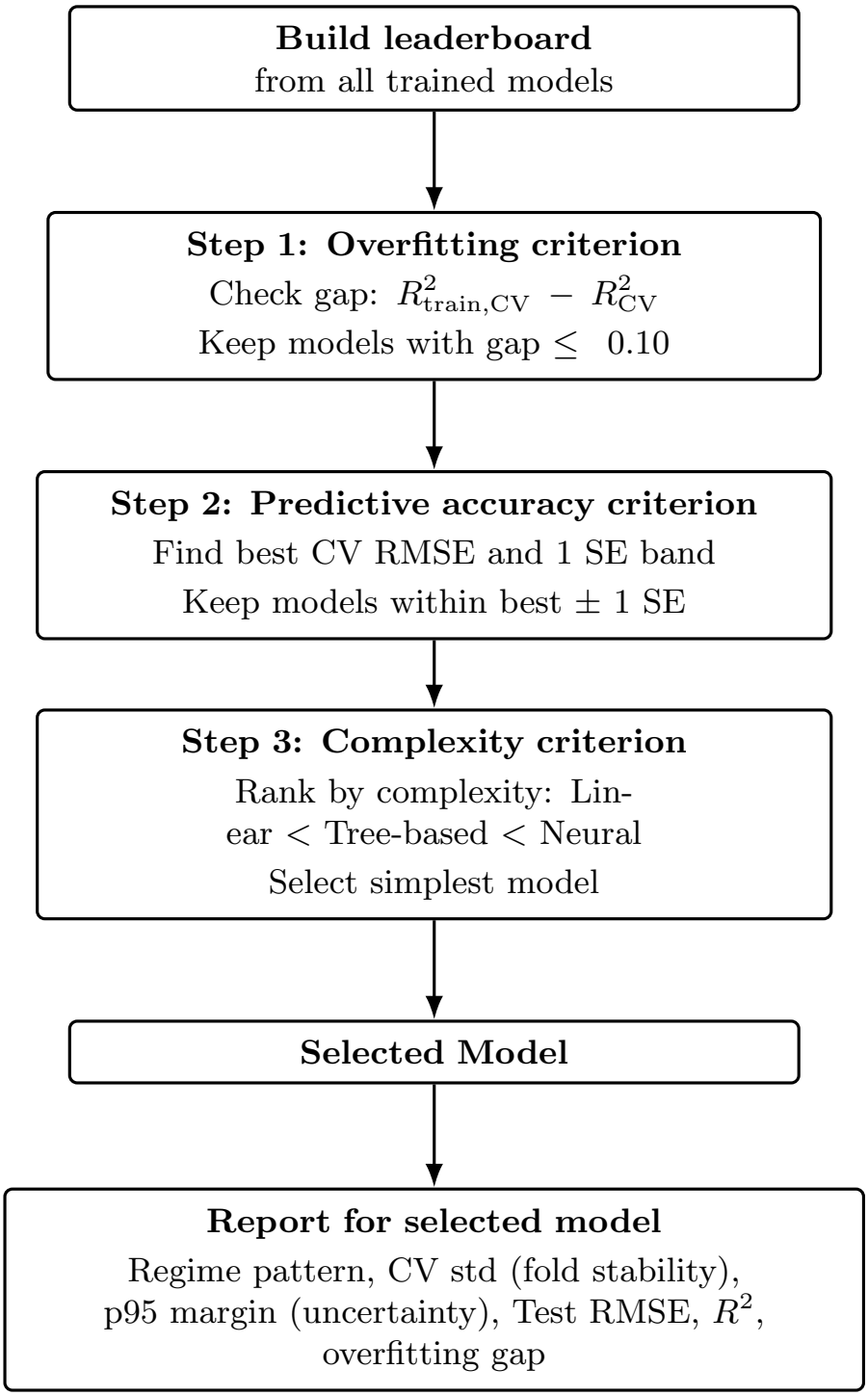


Figure 3.10: Procedure used to select the final regression model.

Step 1: Overfitting criterion. In this step, the overfitting behaviour of each model is checked. The overfitting gap is computed as the difference in R^2 between the training data and the held-out evaluation, which is a standard diagnostic for the generalisation gap (Hastie et al. 2009). Here, the held-out evaluation is not the static validation split, but the held-out fold of the five-fold cross-validation performed on the pooled training and validation data: in each fold, four parts are used for training, and the remaining part is held out to estimate performance. The gap is therefore $R^2_{\text{train,CV}} - R^2_{\text{CV}}$, where $R^2_{\text{train,CV}}$ is the mean training-fold score and R^2_{CV} is the mean held-out-fold score. In this study, a difference of 0.10 is adopted as a study-specific screening threshold rather than as a universal definition of overfitting. A model satisfying

$$R^2_{\text{train,CV}} - R^2_{\text{CV}} > 0.10$$

is removed from the model-selection pool.

Step 2: Predictive accuracy criterion. For each model remaining in the selection pool, the mean cross-validation RMSE and its standard error are calculated. If s_m is the sample standard deviation of the fold RMSE values for model m and K_m is the number of folds, the standard error is

$$\text{SE}_m = \frac{s_m}{\sqrt{K_m}}.$$

Let m^* denote the model with the lowest mean cross-validation RMSE. The one-standard-error ceiling is defined as

$$\overline{\text{RMSE}}_{m^*} + \text{SE}_{m^*}.$$

Every model whose mean cross-validation RMSE is at or below this ceiling remains in the selection pool. RMSE values were rounded to four decimal places for model selection to avoid interpreting differences at a numerical precision that is not meaningful for the intended physical application.

Step 3: Simplicity criterion. Among the models remaining in the selection pool, the final model is chosen using the predefined simplicity ranking in Table 3.6. This ordinal ranking is a study-specific tie-breaker based on model structure and interpretability; it is not a quantitative measure of computational cost, parameter count or training time. It is applied only after the overfitting and predictive-accuracy criteria. The surviving model with the lowest simplicity rank is selected, with any tie broken by the lower mean cross-validation RMSE.

For the winner, these parameters are reported:

Table 3.6: Study-specific simplicity ranking used for model selection.

Category	Model	Simplicity rank
Linear / regularised	Linear regression / ridge	1
	Lasso	2
	Elastic net	3
	Polynomial ridge	4
Neighbour / tree-based	k-nearest neighbours	5
	Random forest	7
	Gradient boosting	8
	XGBoost	8
	LightGBM	8
Neural networks	MLP	9
	Deep neural network	10

1. Composition-resolved error, which shows how the error develops across the range of R retained in the mixed regime, and in particular whether it grows towards the boundary at which the dust contribution falls below the adopted detectability criterion.
2. Fold-to-fold stability, which is the cross-validation standard deviation. It shows how much the RMSE varies across the cross-validation folds; a smaller value indicates that the model's performance is consistent and is not an artefact of a particular fold split.
3. The test-subset metrics, computed once at this stage only, as the final confirmation of the selected model.

The selection procedure also contains a regime-consistency check, which compares the error in the recoverable regimes with the error in the Snomax-dominated regime and flags the model if the latter is not the larger of the two. Like the regime-resolved breakdown, it is inactive in the configuration used here (Section 3.6).

4 Results

This chapter reports the outcome of the pipeline defined in Chapter 3 and follows the same order as the pipeline itself: the generated ensemble and the realised perturbation, the detectability limit and the dataset that survives it, the feature selection and the choice of the number of features, and finally the model comparison and the performance of the selected model.

4.1 The generated ensemble

The main sampling settings are briefly recalled here to provide context for the ensemble results. Following the procedure described in Section 3.4, the production run generated $N_s = 100\,000$ samples using two-dimensional LHS, using two variables: the Snomax surface fraction R and the median particle diameter d_{pm}. The R was sampled log-uniformly over $[10^{-6}, 1]$ according to Equation (3.10). The d_{pm} was sampled log-uniformly over 1.0×10^{-8} m to 5.0×10^{-8} m, using Equation (3.11) and the limits in Equation (3.9).

Under this log-uniform distribution, approximately 80.8 % of the sampled compositions are expected to lie below the dust-detectability threshold $R^* = 0.0701$, which is presented in Section 4.2. By comparison, a linear-uniform distribution would place only approximately 7.0 % of the samples below this value. The log-uniform distribution therefore provides much better coverage of the low-fraction region around the dust detectability limit.

Over the sampled diameter range, the total surface-area concentration varies from 4.36×10^{-13} to 1.09×10^{-11} $\mathrm{m^2\,m^{-3}}$, corresponding to a factor of twenty-five. Since the particle diameter was sampled independently of the surface fraction, the aerosol loading varies independently of the composition. This allows the spectrum amplitude to vary across the ensemble without providing direct information about the target composition, which is why the subsequent feature extraction uses scale-invariant descriptors.

The realised stochastic parameterisation

The effect of the perturbation on the observable quantity is shown in Figure 4.1, which gives the activated fraction of each pure species for the unperturbed fitted parameterisation and for an ensemble of 200 perturbed realisations. The perturbed ensemble occupies a band instead of a single line, and the shaded region shows the 5th–95th percentile range of the

activated fraction produced by the prescribed parameterisation perturbation. The perturbation band becomes wider as the temperature moves farther from the pivot temperature, $T_{\mathrm{ref}} = -20\,^{\circ}\mathrm{C}$.

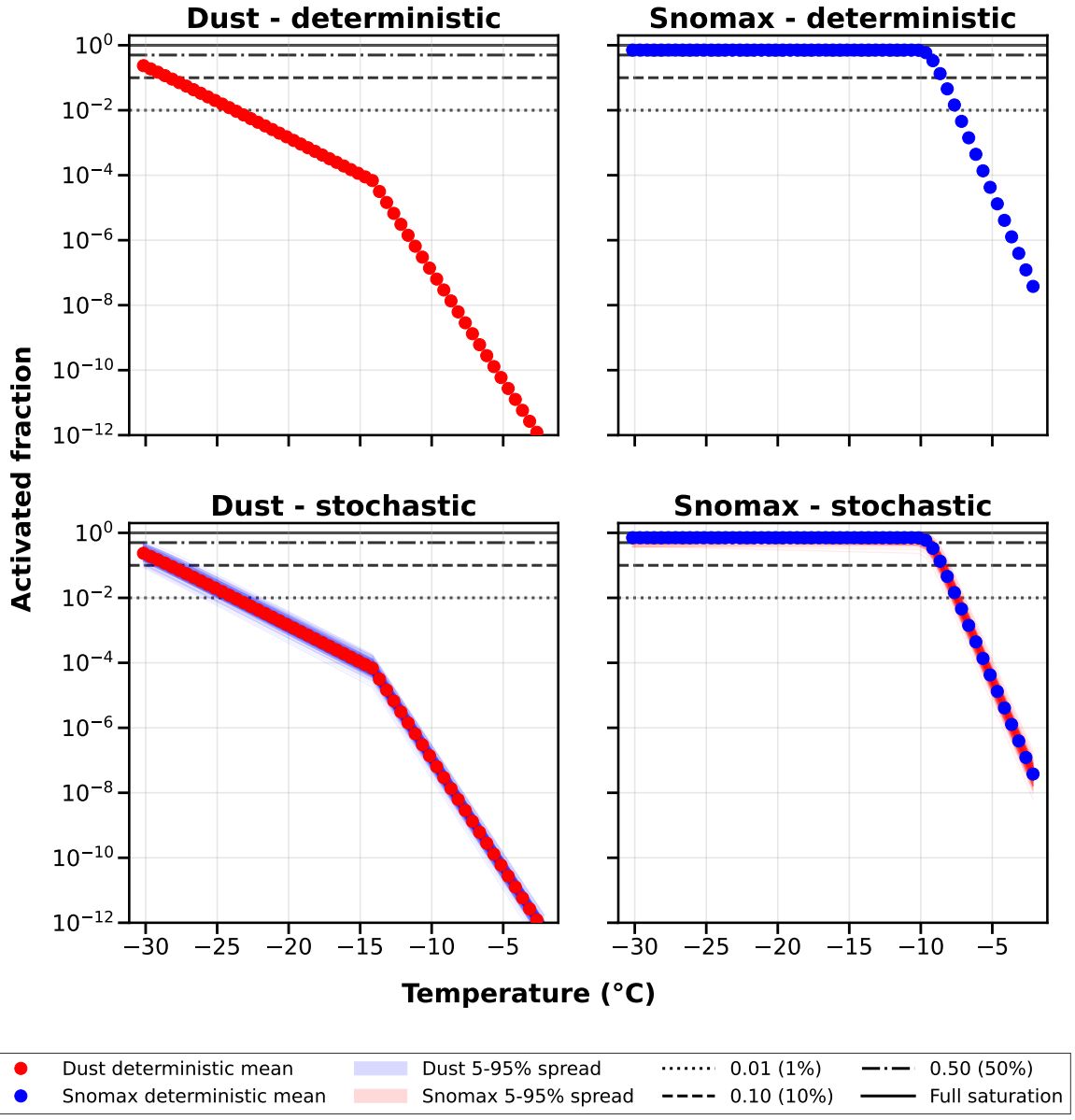


Figure 4.1: Activated fraction of the two pure species on the common temperature grid. The left and right columns show dust and Snomax, respectively. The top row shows the unperturbed fitted parameterisations without perturbation. The bottom row shows an ensemble of 200 realisations with the per-sample amplitude and tilt draw applied, with the the $5th$-$95th$ percentile range shaded and the unperturbed fitted curve overlaid. The spread closes towards the pivot temperature and opens towards both ends of the grid.

Because the draws are independent for the two species, the difference between the dust draw and the Snomax draw of a given sample moves the apparent contribution ratio away from the value that R alone would produce. The forward map from R to the spectrum is therefore

not exactly invertible even before the pointwise spectrum perturbation of Equation (3.15) is applied. This contributes to the retrieval error reported in Section 4.6, and it is one reason why a perfect score should not be expected on this dataset. The second layer, the pointwise spectrum perturbation, adds one independent draw per temperature point and contains no prescribed composition-dependent information; the two layers are shown together on representative spectra in Figure 3.8.

4.2 Detectability and the retained dataset

The dust-detectability threshold

Evaluated on the unperturbed fitted INAS parameterisations with $\sigma_{\log} = 0.5$ and the 5σ criterion, the detection statistic of Section 3.6 gives the dust-detectability threshold

$$R^* = 0.0701. \tag{4.1}$$

The dust contribution on the unperturbed fitted parameterisations satisfies the adopted detection criterion up to a Snomax surface fraction of approximately seven per cent. The behaviour of the detection statistic around this threshold is shown in Figure 4.2. Within the dilute-limit diagnostic of Equation (3.22), the overall amplitude cancels from the fractional contribution, so this threshold is independent of the assumed total loading.

What this limit means for the spectrum is illustrated in Figure 4.3. The figure shows the dust and Snomax contributions at half, at one and at twice R^*, calculated from the unperturbed fitted parameterisations without perturbations or pointwise spectrum perturbation. At half R^*, dust controls the low temperature part of the spectrum. At R^* and even at twice R^*, the influence of dust remains visible at the lowest temperatures. This occurs because the dust activity continues to increase towards lower temperatures, while the Snomax activity reaches a plateau. Therefore, R^* does not indicate that the dust contribution becomes zero or visually disappears from a spectrum without the pointwise perturbation. Instead, it is the limit defined by the detection statistic of Equation (3.20). At R^*, the dust statistic is approximately equal to the 5σ threshold, and above R^* it falls below this threshold under the assumed prescribed pointwise variability level of $\sigma_{\log} = 0.5$. The remaining effect of dust at low temperatures should therefore still be considered when interpreting this limit.

Per-sample regimes and the retained dataset

Using the detection statistic of Equation (3.20) and the four regime definitions introduced in Section 3.6, each of the 100 000 samples is classified based on its parameterisation-perturbed spectrum before the pointwise spectrum perturbation is applied. The resulting counts are given in Table 4.1: 45 661 samples (45.7 %) are classified as mixed, 43 405 (43.4 %) as dust-dominated and 10 934 (10.9 %) as Snomax-dominated. No sample falls into the undetectable

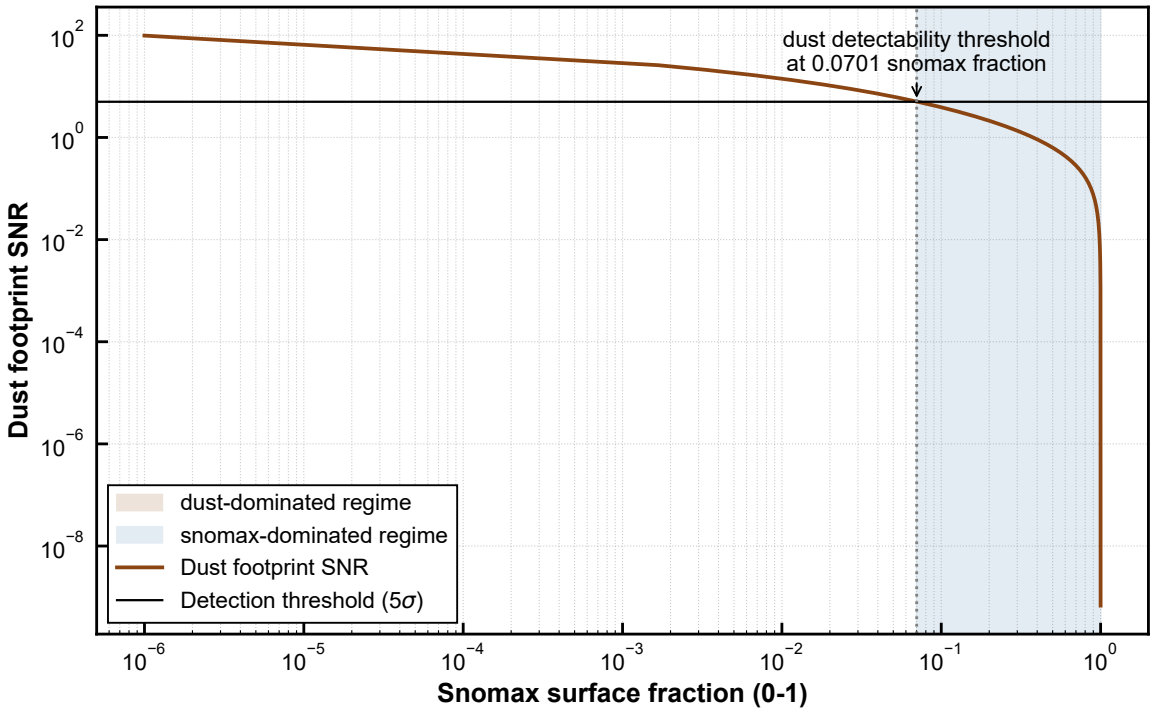


Figure 4.2: Dust detection statistic as a function of the Snomax surface fraction, calculated using the unperturbed fitted INAS parameterisations. The dashed horizontal line shows the threshold $k = 5$, and the vertical line shows $R^* = 0.0701$. The common total surface-area concentration S_{tot} cancels from the fractional contribution in Equation (3.22), so the curve depends on composition rather than on the value of S_{tot}.

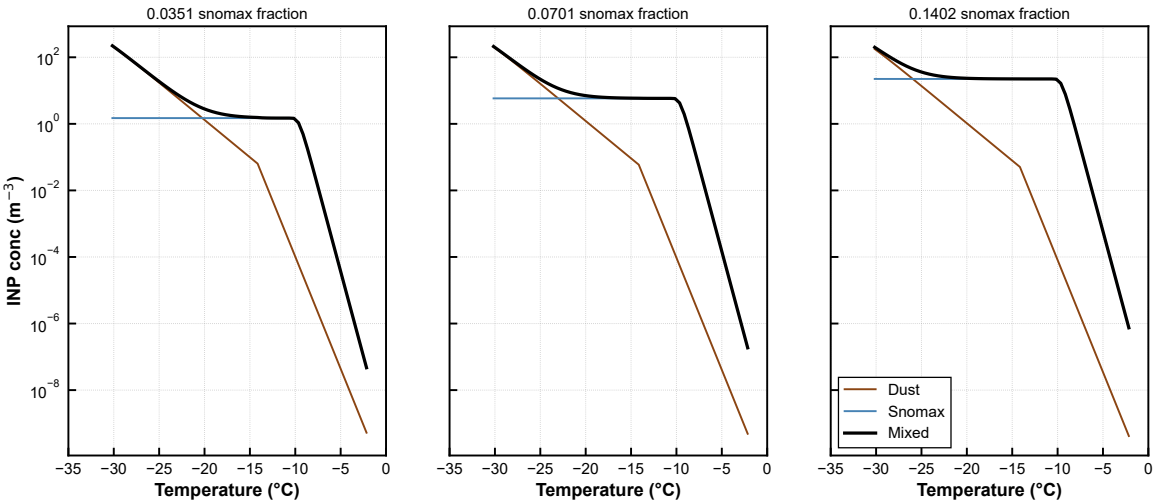


Figure 4.3: Dust and Snomax contributions and their sum at half, at one and at twice the dust-detectability threshold, $R^* = 0.0701$, calculated using the unperturbed fitted parameterisations without applying the pointwise spectrum perturbation. The dust contribution remains visible at the lowest temperatures, including at twice R^* in panel (c). The value R^* represents the limit defined by the 5σ detection criterion under the assumed prescribed pointwise variability level, rather than the point at which the dust contribution disappears from the noise-free spectrum.

Table 4.1: Detectability criteria and sample counts for the 100 000 generated spectra. The classification is evaluated on the parameterisation-perturbed spectrum before the pointwise spectrum perturbation is applied, using $k = 5$.

Regime	$\mathrm{SNR}_{\mathrm{dust}}$	$\mathrm{SNR}_{\mathrm{snomax}}$	Samples
mixed	≥ 5	≥ 5	45 661
dust-dominated	≥ 5	< 5	43 405
Snomax-dominated	< 5	≥ 5	10 934
undetectable	< 5	< 5	0
total	–	–	100 000

regime, meaning that at least one species passes the selected 5σ criterion in every simulated spectrum under the prescribed multiplicative pointwise-variability model of Equation (3.15).

Because this model includes prescribed multiplicative pointwise-variability model but no absolute instrumental detection limit, this result does not show that every spectrum would be measurable in an experiment. Figure 4.4 shows the classification based on the two detection statistics, while Figure 4.5 shows representative spectra from the different regimes. In the dust-dominated and Snomax-dominated regimes, only one species passes the selected detection criterion, although both species still contribute to the simulated spectrum.

As specified in Section 3.6, only samples classified as mixed, for which both dust and Snomax pass the 5σ criterion of Equation (3.20), are carried into feature extraction. The detection thresholds and the resulting regimes are shown in Figure 4.4. This leaves 45 661 of the 100 000 generated spectra, corresponding to 45.7 % of the ensemble (Table 4.1). Under the log-uniform sampling, 80.8 % of the sampled compositions lie at or below the dust-detectability threshold calculated from the unperturbed fitted parameterisations $R^* = 0.0701$, shown in Figure 4.2. However, these percentages represent different conditions. The condition $R \leq R^*$ considers only whether the unperturbed dust contribution passes the detection threshold, whereas the mixed classification requires both species to pass the threshold in each parameterisation-perturbed spectrum.

The stratified split of Section 3.9 divides the retained set into 36 521 training, 4560 validation and 4580 test samples. All metrics reported in the remainder of this chapter are conditional on the mixed regime and must not be read as performance over the full composition range $[0, 1]$.

Size of the retained dataset

Following the definition of (Viering and Loog 2023), a learning curve describes the expected generalisation risk of a learning algorithm as a function of the training-set size (Viering and Loog 2023). In this study, this risk is estimated by the mean RMSE on the held-out cross-validation folds. To calculate the empirical learning curve, the selected model is

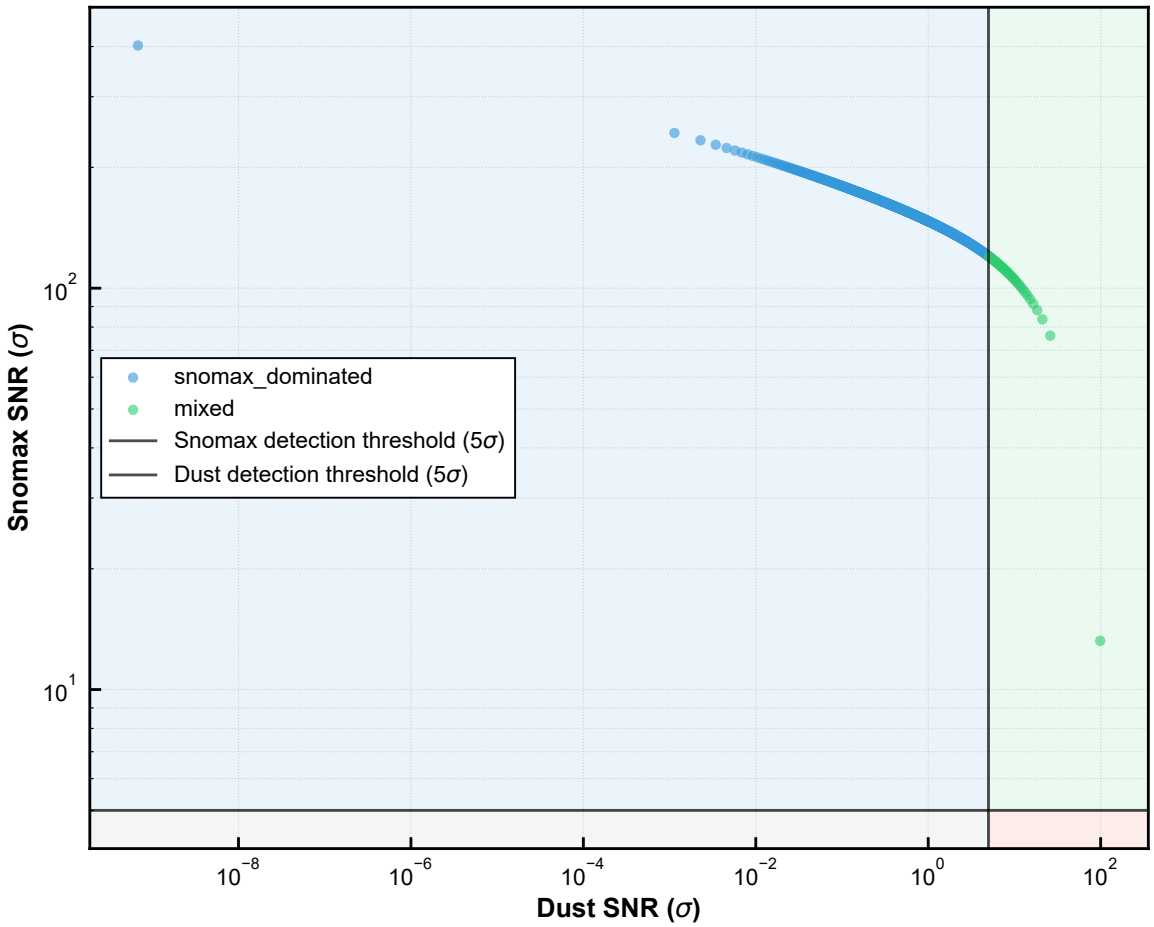


Figure 4.4: Detectability classification in the plane of the dust and Snomax detection statistics. The two thresholds divide the plane into four regimes. All generated samples contain the composition assigned during synthesis, but outside the mixed quadrant at least one species contribution falls below the adopted detectability threshold.

repeatedly fitted using training subsets of increasing size. The cross-validation RMSE is calculated for each subset size using the procedure described in Section 3.10. The training RMSE, calculated on the samples used to fit the model, is shown as an additional diagnostic. In both cases, a lower RMSE indicates that the predicted values of R are closer to their true values.

Figure 4.6 shows the empirical learning curve and the accompanying training-error curve. The training RMSE decreases from approximately 0.0181 for the smallest subset to 0.0165 for the full training set. Over the same range, the cross-validation RMSE changes only slightly, from approximately 0.0172 to 0.0171. The small difference between the training and cross-validation RMSE provides no clear indication of strong overfitting.

The cross-validation RMSE is nearly constant over the tested range. This indicates that approximately 5000 training samples already achieve a similar average cross-validation RMSE and that using additional samples provides little improvement in this metric. The

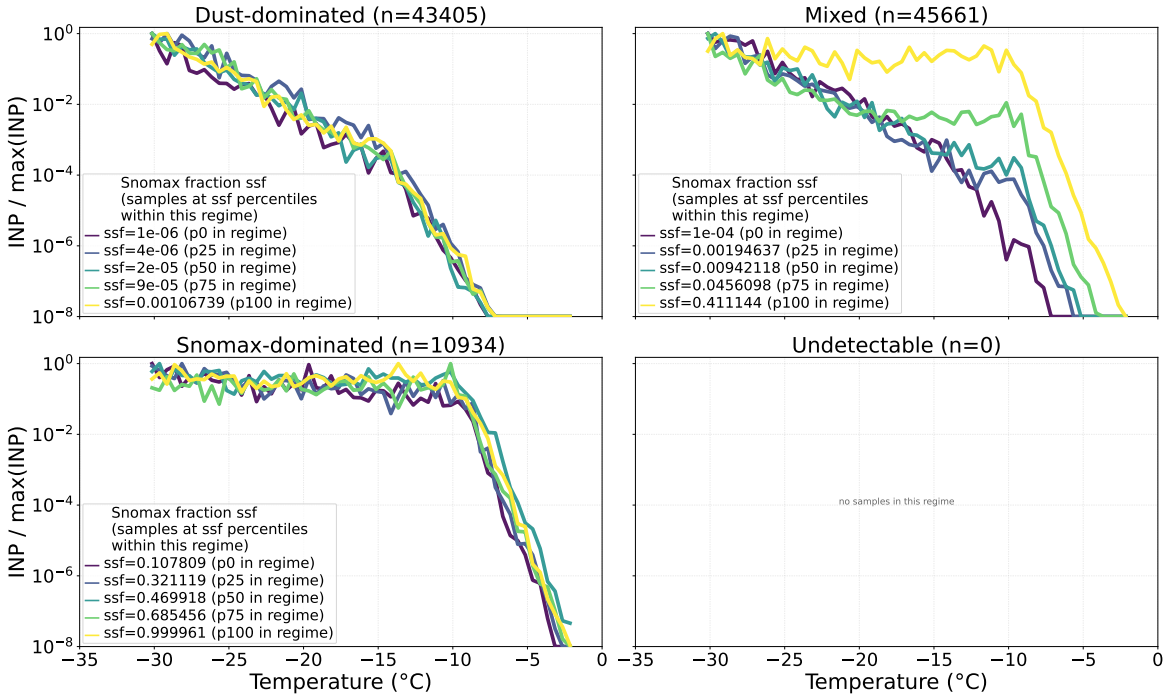


Figure 4.5: Representative synthetic spectra from the detectability regimes, including the per-sample INAS parameterisation perturbation and the pointwise spectrum perturbation. In the mixed regime, both the high temperature Snomax shoulder and the low temperature dust branch are detectable in the same spectrum. In the other regimes, one contribution is present by construction but falls below the adopted detectability threshold.

full training set is used for the final model because the samples are available and the computational cost is manageable.

4.3 Feature selection

The selection starts from the eleven shape-intrinsic candidates of Section 3.8, the seven reference-shape features having already been excluded as a conceptual leakage, and is performed on the training split only.

Variance filter. The near-zero-variance filter removes nothing. The smallest variance in the candidate pool is that of `warm_fraction`, at about 0.0035, which is more than five orders of magnitude above the threshold of 10^{-8}. On this dataset the filter is therefore only a formality, and the whole reduction is performed by the correlation step. This is not surprising: every candidate is derived from the same spectrum, and none of them is close to constant over an ensemble whose composition spans several decades.

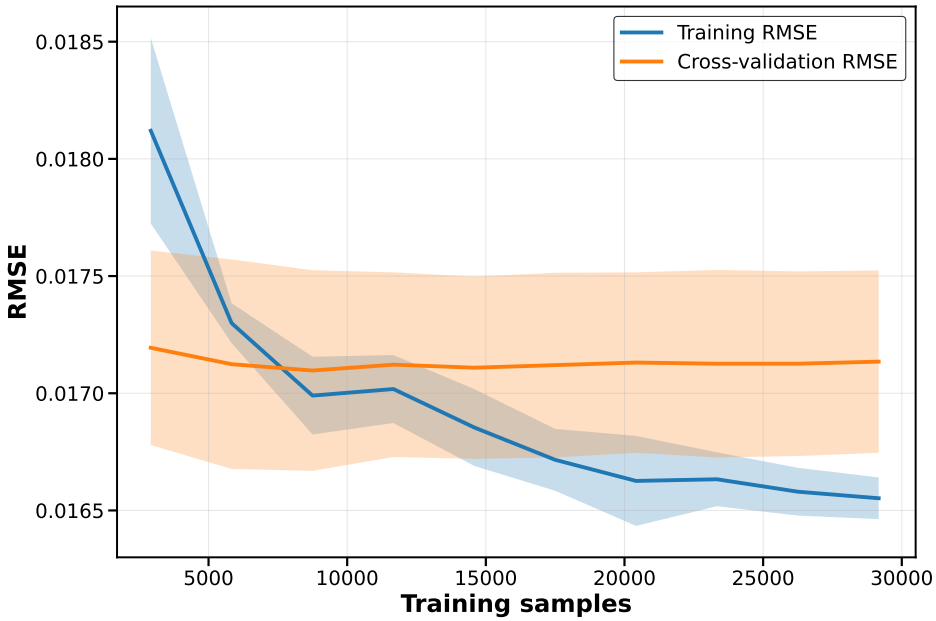


Figure 4.6: Learning curve of the selected model: training and cross-validation RMSE against the size of the training subset. The two curves converge to a small stable gap and the cross-validation error is nearly flat at the full training set.

Table 4.2: Correlated feature pairs above the threshold $|r| > 0.90$, in the order in which they are processed, with the member removed in each pair.

Removed	Retained	$\lvert r \rvert$
`t90`	`density_std_T`	0.95
`onset_temperature`	`slope_log`	0.92
`density_mean_T`	`warm_fraction`	0.92
`warm_fraction`	`density_std_T`	0.91

Correlation pruning. Correlation pruning removes redundant features that carry similar information. The procedure is described in Section 3.8. Feature pairs with an absolute Pearson correlation above 0.90 are processed from the strongest to the weakest correlation. For each pair, the feature less correlated with the target R is removed, while the other is retained. This procedure removes four of the eleven candidate features and leaves seven. The individual pruning decisions are listed in Table 4.2. The correlation structure before and after pruning is shown in Figure 4.7.

The removals are physically understandable rather than incidental. The quantile temperatures, the weighted mean and the weighted standard deviation of the freezing density all describe the position and the width of the same distribution, so they carry largely the same information; the onset temperature and the log-spectrum slope both respond to how early the high temperature active material begins to freeze. What survives is one descriptor of the position and width of the freezing distribution, one descriptor of its shape, and the two slope descriptors, which is a set with no redundant member left.

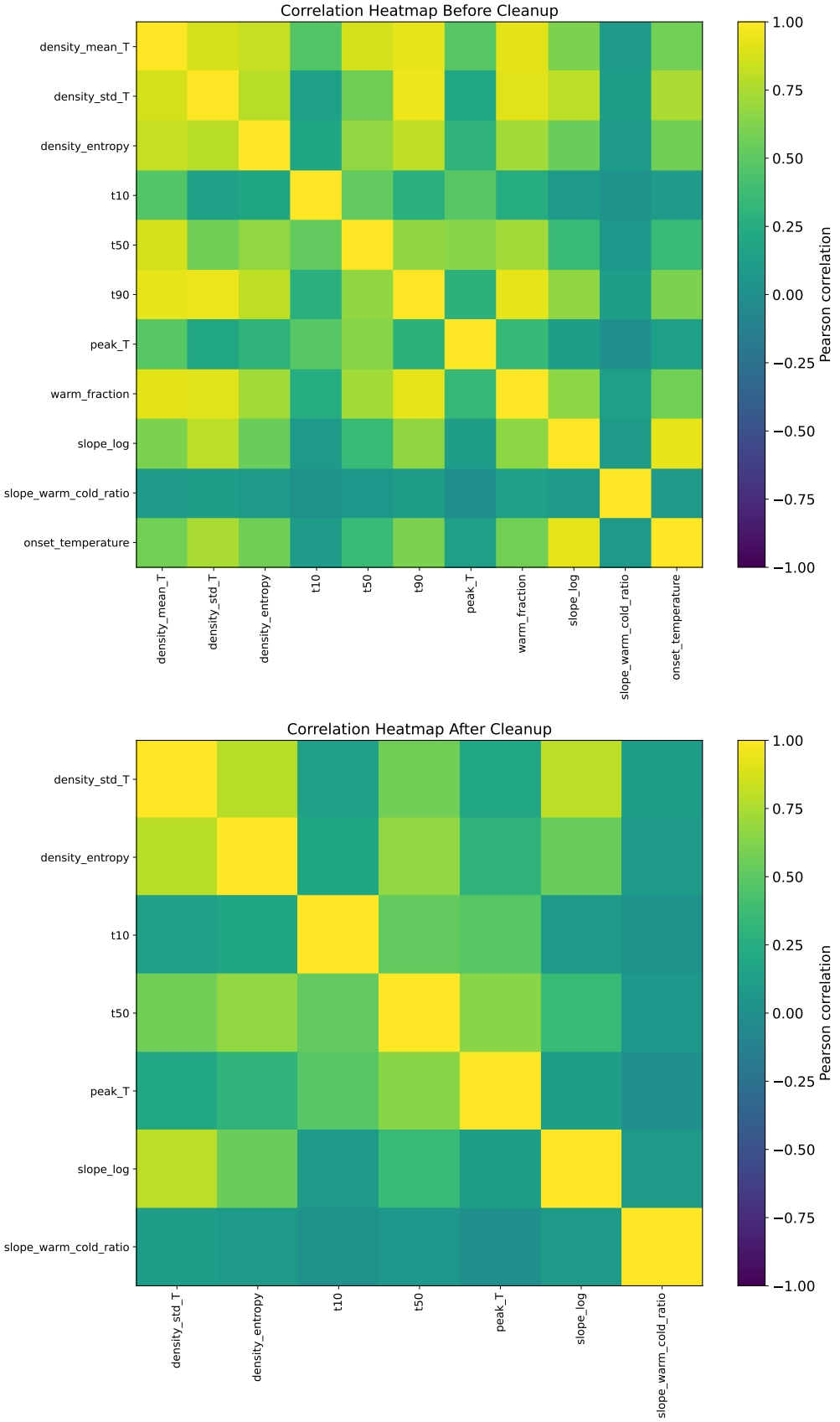


Figure 4.7: Absolute Pearson correlation between the candidate features, computed on the training subset only. **Top:** the eleven shape-intrinsic candidates, with several feature pairs above the threshold. **Bottom:** the seven features that remain after pruning, with no remaining pair above $|r| = 0.90$.

The four removals correspond to the correlation-pruning decisions in Table 4.2. The feature `t90` is removed in favour of `density_std_T`, and `onset_temperature` is removed in favour of `slope_log`. The feature `density_mean_T` is first removed in favour of `warm_fraction`, which is subsequently removed in favour of `density_std_T` by the greedy pruning procedure. Seven features remain. The feature `density_std_T` describes the width of the freezing-density distribution, while `density_entropy` describes how the freezing-density weights are distributed across temperature. The features `t10` and `t50` identify the 10th temperature quantile and the median temperature, respectively, and `peak_T` identifies the temperature of maximum freezing density. The feature `slope_log` describes the overall temperature dependence of the log-spectrum, whereas `slope_warm_cold_ratio` compares its high temperature and low temperature side slopes. The retained set therefore contains five freezing-density descriptors and two slope descriptors.

Importance ranking. The seven survivors are ranked by mean absolute SHAP value on the XGBoost reference model with the exact tree estimator, so the contributions are not approximated. The ranking is given in Table 4.3 and shown in Figure 4.8.

The ranking is strongly top-heavy. The three leading features, `slope_warm_cold_ratio`, `slope_log` and `density_std_T`, carry about 42 %, 31 % and 23 % of the total importance, so about 96 % together. Between the third and the fourth feature the importance drops by roughly a factor of fourteen, and the remaining four features account for only a few per cent in total. The gap between the third and the fourth feature is therefore much wider than any gap inside the leading group, which is a useful property: the cut can be made in a wide flat region of the ranking rather than between two nearly equal candidates.

Table 4.3: Ranking of the seven surviving features by mean absolute SHAP value on the XGBoost reference model.

Feature	mean |SHAP|	Rank
`slope_warm_cold_ratio`	0.0192	1
`slope_log`	0.0142	2
`density_std_T`	0.0107	3
`density_entropy`	0.0007	4
`t50`	0.0006	5
`t10`	0.0004	6
`peak_T`	0.0003	7

Robustness of the ranking. An ordering obtained from a single importance measure on a single fit is not by itself trustworthy, so the ranking is checked in two independent ways (Figure 4.9). First, the same features are ranked by permutation importance, which is model-agnostic and does not share the assumptions of the tree-based attribution. Second, the SHAP ranking is recomputed on bootstrap resamples of the training set and the spread of each feature's rank across the resamples is recorded. The leading group holds its position in both checks, so the ordering used for the feature-count sweep is not an artefact of one seed or one importance measure.

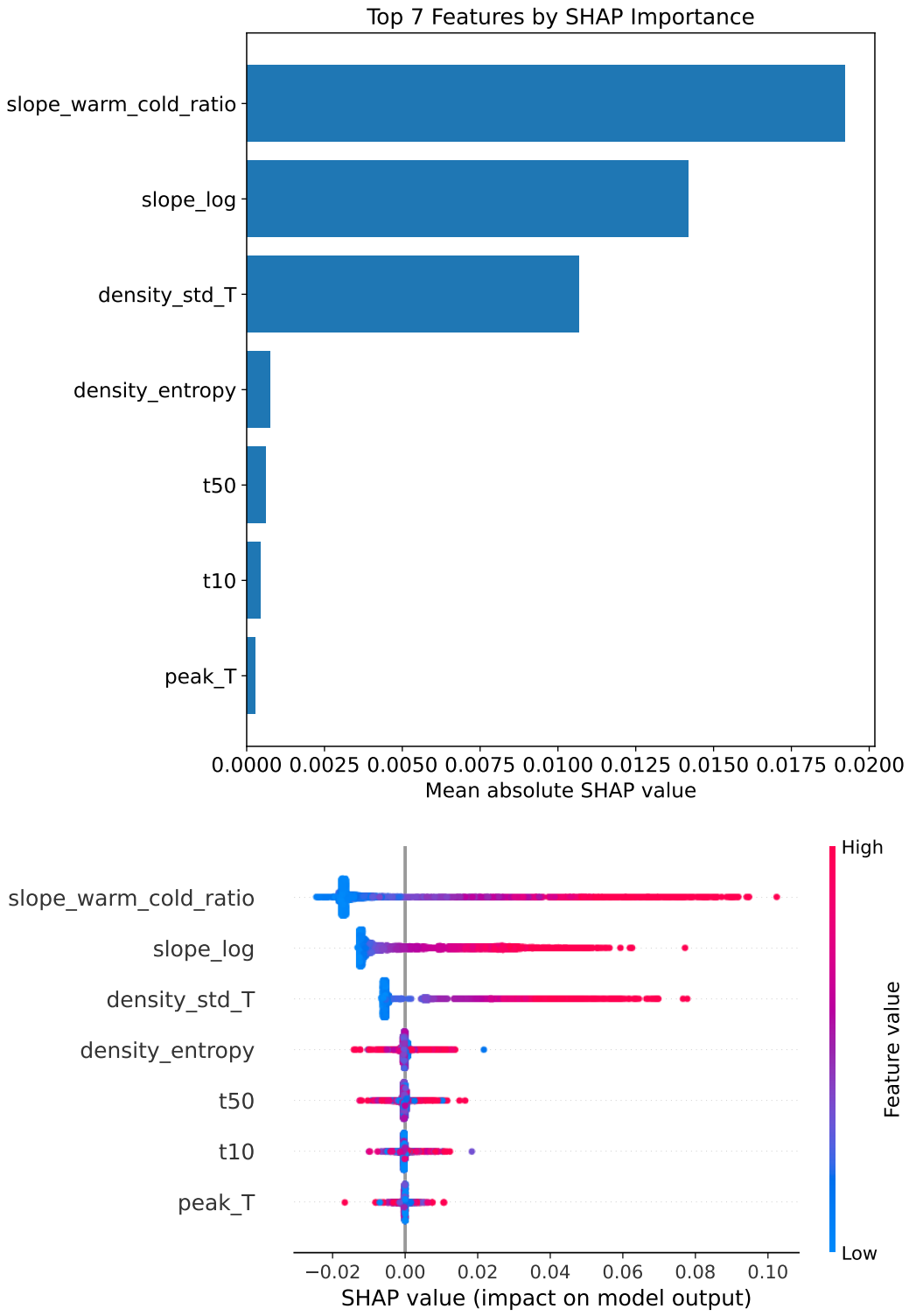


Figure 4.8: Top mean absolute SHAP value of each surviving feature. **Bottom** the distribution of the SHAP values of each feature over the training samples, which shows the direction as well as the size of the contribution.

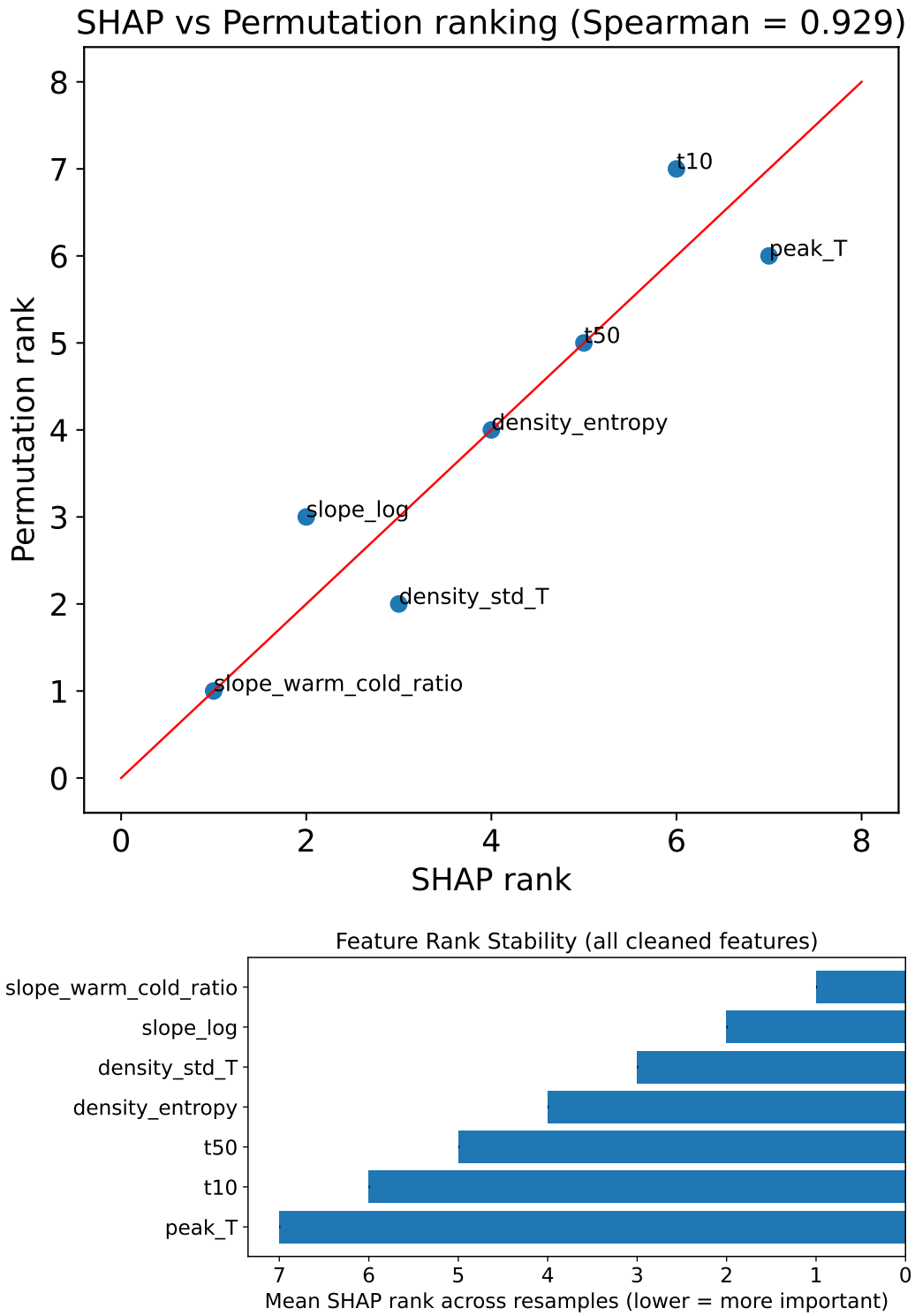


Figure 4.9: Top the SHAP ranking against the permutation-importance ranking of the surviving features. **Bottom** the stability of the SHAP rank of each feature across bootstrap resamples of the training set.

4.4 Feature interpretation and the number of features

Physical Interpretation of the Leading Features

All three leading features have a direct reading in terms of the activation behaviour of the two species.

`slope_warm_cold_ratio`, the ratio of the log-spectrum slope higher temperature than −15 °C to the slope lower temperature than it, is close to one for a single species, because a pure population has a similar steepness on both sides. A mixture departs from one as soon as the two contributions differ, because Snomax shapes the high temperature part of the spectrum and dust the low temperature part. It is a contrast between two parts of the same curve, which is precisely the kind of quantity a composition should control, and this is why it ranks first.

`slope_log`, the steepness of the whole log-spectrum, moves between the two pure values as the composition changes, because the two species have different characteristic steepnesses.

`density_std_T`, the weighted standard deviation of the normalised per-degree freezing density, measures how widely the freezing is spread over temperature. A pure population freezes over a comparatively narrow range, whereas two species active in different ranges freeze over a wider one, so the spread grows as the mixture becomes more balanced.

These three features belong to the eleven reference-free candidates introduced in Section 3.7. Their scale independence and their potential transfer to measured spectra, together with the limitations of this transfer, are discussed in Section 3.8. The remaining four features that survive correlation pruning are interpreted through the same freezing-density definitions given there.

Selection of the number of features

The number of features is evaluated using the incremental procedure described in Section 3.8. The features are added one at a time according to their SHAP ranking. The first subset contains only the highest-ranked feature, the second contains the first two features, and this continues until all seven features are included. Linear regression and XGBoost are fitted for every subset and evaluated using the five-fold cross-validation procedure described in Section 3.10. The resulting mean cross-validation R^2 values are shown in Figure 4.10.

For linear regression, the mean cross-validation R^2 increases from approximately −0.29 with one feature to 0.56 with two features and 0.86 with three features. Adding further features produces only small changes. This shows that the linear model needs at least three of the engineered features to describe the relationship between the spectral shape and the composition. XGBoost performs differently. Its mean cross-validation R^2 is already approximately 0.89 with the highest-ranked feature. Adding the second feature gives a

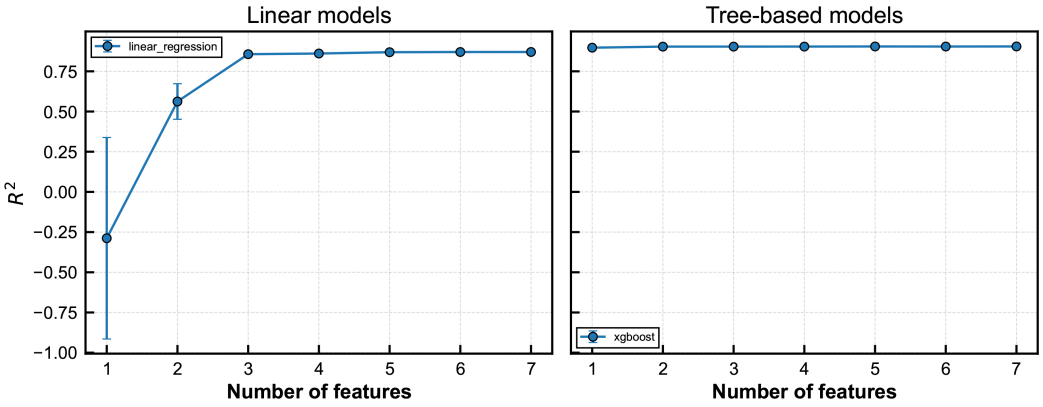


Figure 4.10: Mean five-fold cross-validation R^2 as the SHAP-ranked features are added one at a time. The error bars show the fold-to-fold standard deviation. **Left:** linear regression. **Right:** XGBoost. The XGBoost score changes only slightly after the second feature.

small improvement, while adding the third and later features produces no clear further improvement.

The first two SHAP-ranked features, `slope_warm_cold_ratio` and `slope_log`, are therefore retained for the full model comparison. The exclusion of `density_std_T` may appear inconsistent with its SHAP importance of approximately 23%. However, SHAP describes how a model containing all seven features uses each feature. It does not directly measure how much the prediction improves when that feature is added to a smaller subset. Feature ranking and feature-subset selection are different tasks, particularly when different features contain partly redundant information (Guyon and Elisseeff 2003). In this study, adding `density_std_T` as the third feature does not produce a clear improvement in the XGBoost cross-validation R^2. The two-feature subset is therefore used in the following model comparison.

4.5 Model comparison and selection

With the feature set fixed at the two leading features, the eleven candidate regressors of Table 3.5 are trained under the shared protocol of Section 3.10 and ranked by the three-step rule of Section 3.11. The outcome for every candidate is listed in Table 4.4, and Figure 4.11 shows which model leaves at which step and why.

Step 1, the overfitting gate. The overfitting gate is defined in Section 3.11. Three models exceed the gap threshold of 0.10 and are removed, as reported in Table 4.4 and shown in Figure 4.11. The polynomial model exceeds the threshold by a large margin, with a gap of the order of a thousand; the multilayer perceptron fails with a gap of 4.398 and the deep network with 0.297. The eight remaining models all have gaps below 0.06, and the four tree-based ones below 0.01. The gate is therefore doing real work on this dataset: it removes

Table 4.4: Model-selection results for the eleven candidates using the two selected features. The CV RMSE standard deviation describes the fold-to-fold variability. Step 1 removes models with an overfitting gap greater than 0.10. Step 2 retains models within the adopted cross-validation variability band. Step 3 selects the surviving model with the lowest simplicity rank, where a lower rank indicates a simpler model.

Model	Family	Mean CV RMSE	CV RMSE SD	Overfitting gap	Simplicity rank	Step 2 status	Final outcome
random forest	tree ensemble	0.0171	0.0003	0.0066	7	inside band	**selected**
gradient boosting	tree ensemble	0.0169	0.0002	0.0040	8	inside band	not selected: higher complexity
XGBoost	tree ensemble	0.0170	0.0002	0.0022	8	inside band	not selected: higher complexity
LightGBM	tree ensemble	0.0170	0.0003	0.0045	8	inside band	not selected: higher complexity
k-nearest neighbours	neighbour-based	0.0174	0.0003	0.0085	5	outside band	removed in Step 2
lasso	linear	0.0359	0.0041	0.0559	2	outside band	removed in Step 2
elastic net	linear	0.0359	0.0042	0.0570	3	outside band	removed in Step 2
linear regression	linear	0.0359	0.0042	0.0572	1	outside band	removed in Step 2
deep neural network	neural	0.0295	0.0172	0.2971	10	–	removed in Step 1
multilayer perceptron	neural	0.0766	0.0978	4.3984	9	–	removed in Step 1
polynomial ridge	linear, expanded basis	0.9774	1.8306	1439.9709	4	–	removed in Step 1

the three most flexible families, which are the ones whose flexibility is not matched by the amount of information that two shape features can carry.

Step 2, the one-standard-error band. The band ceiling lies at about 0.018 in cross-validation RMSE and four models fall inside it: gradient boosting, XGBoost, LightGBM and the random forest. The k-nearest neighbours model misses the band narrowly, while the three linear models miss it by a wide margin, at about 0.036, roughly twice the value of the tree ensembles. This confirms at the level of the full model comparison what the feature-count sweep had already shown within a single family, namely that a linear map from these features to the composition is not adequate.

Step 3, parsimony. Among the four models inside the band the simplest is selected. The random forest has simplicity rank 7 against 8 for the three boosted-tree models, so the random forest is chosen. The point of this step is that when four models have similar performance under the adopted one-standard-error rule, deciding between them on their score differences would amount to fitting the selection to noise, and it is better to spend the tie on a property that is not noisy. A random forest is a bagged average of independently grown trees with no sequential dependence between them (Breiman 2001), so it has fewer interacting hyperparameters and a more predictable behaviour than a boosted ensemble, which matters for a model that is meant to be interpreted and reused.

4.6 Final model

The selected model is a random forest regressor. It is trained using the two features selected in Section 4.4: `slope_warm_cold_ratio` and `slope_log`. The tuned hyperparameters of the selected model are listed in Table 4.5.

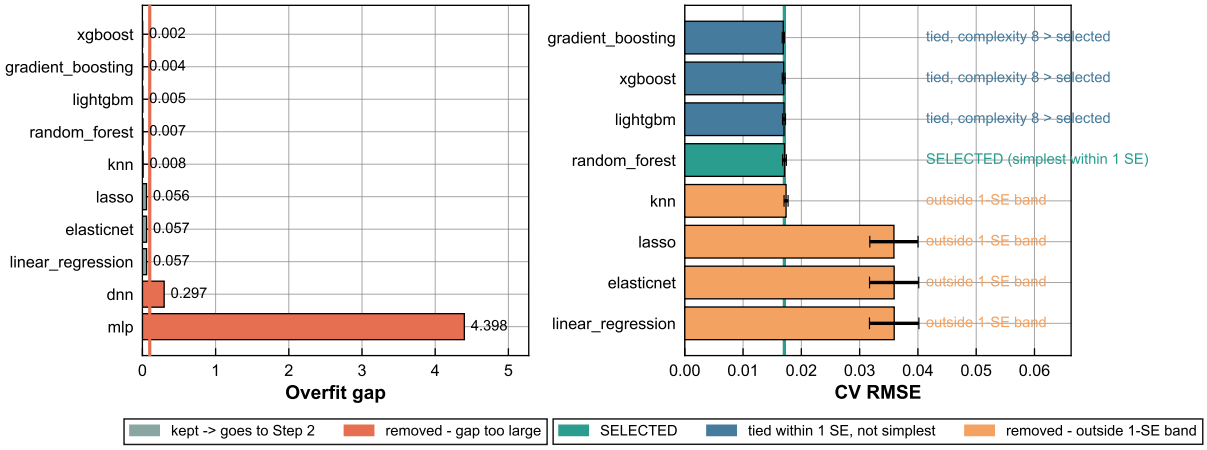


Figure 4.11: The selection rule applied. **Left** the overfitting gate at a threshold of 0.10, which removes the polynomial model, the multilayer perceptron and the deep network. **Right** the one-standard-error band and the parsimony tie-break among the survivors, with the random forest selected as the simplest model inside the band.

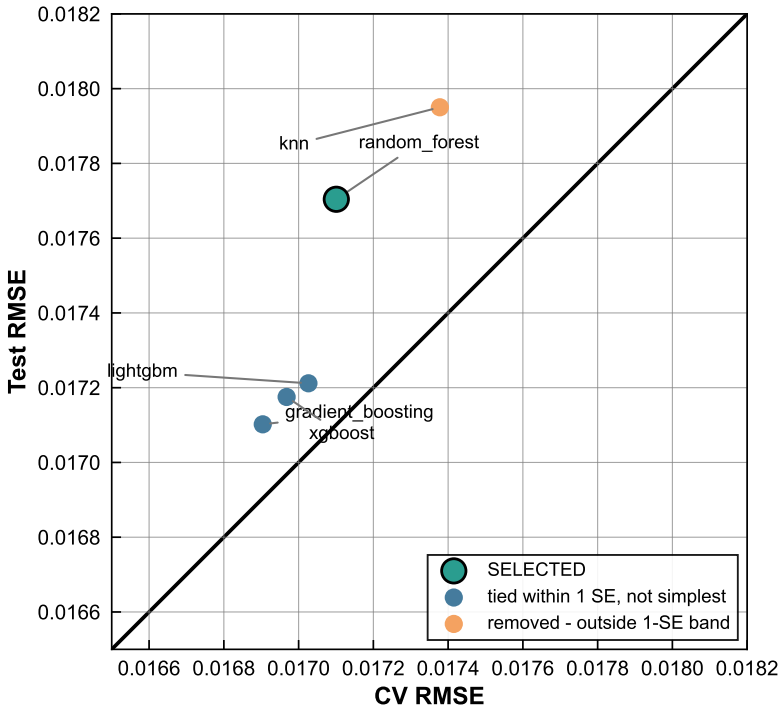


Figure 4.12: Cross-validation RMSE and held-out test RMSE for the five models closest to the one-standard-error selection band. The diagonal line indicates equal cross-validation and test RMSE. The colours show the outcome of the selection procedure: random forest is selected in Step 3, gradient boosting, XGBoost and LightGBM lie within the selection band but are not the simplest models, and k-nearest neighbours lies outside the band. The remaining candidate models are outside the displayed range and are reported in Table 4.4. The test RMSE is shown only for comparison and is not used for model selection.

Table 4.5: Tuned hyperparameters of the selected random forest, chosen by grid search on the validation split.

Hyperparameter	Value
`n_estimators`	1000
`max_depth`	10
`min_samples_leaf`	50

Overall performance

Table 4.6 summarises the performance on the three partitions. On the held-out test set the model reaches $R^2 = 0.90$ with an RMSE of 0.018 and a mean absolute error of 0.008 in Snomax surface fraction, and about 97 % of the test predictions lie within ±0.05 of the true value.

Three properties of this table are worth stating explicitly. First, the gap between the training and the test RMSE is only about 0.001, so the model generalises; this is consistent with the learning curve presented in Section 4.2. Second, the bias is negligible on all three partitions, so there is no systematic offset over the composition range as a whole. Third, the maximum absolute error is far larger than the RMSE on every partition. This indicates that a small number of samples have substantially larger errors than the typical prediction.

Table 4.6: Performance of the selected random forest on the three partitions. "Within 0.05" is the fraction of predictions falling within ±0.05 of the true Snomax surface fraction.

Split	MAE	RMSE	R^2	Bias	MaxErr	Within 0.05
train	0.008	0.017	0.91	0.000	0.286	97.3 %
validation	0.008	0.017	0.90	0.000	0.149	96.9 %
test	0.008	0.018	0.90	0.000	0.313	96.9 %

Distribution of Prediction Errors

Figure 4.13 locates the error. The predictions follow the diagonal with a bias of only +0.0002; the agreement is tightest at small compositions, where most of the retained samples lie, and the scatter widens as the composition grows, with errors appearing on both sides of the diagonal rather than as a shift of the whole cloud. The residual distribution is centred on zero and sharply peaked, and its width grows with the predicted composition, so the error is heteroscedastic, and large residuals are rare, which is consistent with a maximum error much larger than the RMSE.

The per-composition breakdown of Figure 4.14 makes the pattern quantitative. The bin-wise MAE and RMSE grow from close to zero at the smallest compositions to about 0.12 at the largest, and the bias stays near zero up to a composition of about 0.15 before turning

increasingly negative, reaching about -0.12, so the model under-predicts the largest Snomax fractions.

This is the point at which the retrieval result and the detectability analysis of Section 4.2 meet, and the behaviour is what the physics of the problem predicts. As the Snomax fraction grows the dust footprint on the mixed spectrum shrinks, the shape approaches that of a pure Snomax spectrum and saturates, and the shape features on which the model relies stop changing with the composition, exactly as Figure 4.3 shows at and above R^*. Two different high compositions then produce nearly the same shape, so the composition can no longer be read off, and the model responds by regressing towards the centre of the target distribution it has seen, which appears as a negative bias at the top of the range. The effect is reinforced by the sampling: because the composition was drawn log-uniformly, high compositions are also comparatively rare, so there is least data where the problem is hardest.

The relative feature importance of the final model, shown in Figure 4.15, is consistent with the SHAP ranking of Section 4.3: the high temperature-to-low temperature slope ratio dominates and the overall log-slope contributes the remainder, which is the same ordering the reference model produced on the full surviving feature set. That two independently computed attributions on two different model families agree is a sign that the model uses the composition-bearing part of the spectral shape and not an artefact of the synthesis.

4.7 Summary of the results

The results answer the research questions of this thesis as follows.

1. A continuous composition parameter can be recovered from a single perturbed synthetic mixed-species spectrum. On the retained mixed-regime dataset the selected model reaches $R^2 = 0.90$ on held-out data with an RMSE of 0.018 in surface fraction and a negligible bias, which is a useful retrieval even though the forward map was deliberately made non-invertible by the independent per-species perturbations.
2. The recovery needs neither many features nor a large model. Two scale-invariant shape features and a random forest are sufficient, and adding further features changes the score by less than 0.001 in R^2.
3. The best model among the eleven candidates is the random forest, chosen as the simplest model inside the one-standard-error band. The two features it uses, the high temperature to low temperature slope ratio and the overall steepness of the log-spectrum, are both reference-free and invariant to the total surface area.
4. The physical limit is set by the burial of the dust footprint under the Snomax signal. On the unperturbed fitted parameterisations the dust footprint drops below the 5σ detectability threshold defined using the prescribed pointwise variability at $R^* = 0.0701$, and the retrieval carries the signature of the same limit: the error grows and the bias turns negative towards the upper end of the composition range, because there the spectral shape saturates and stops encoding the composition. The limit is therefore

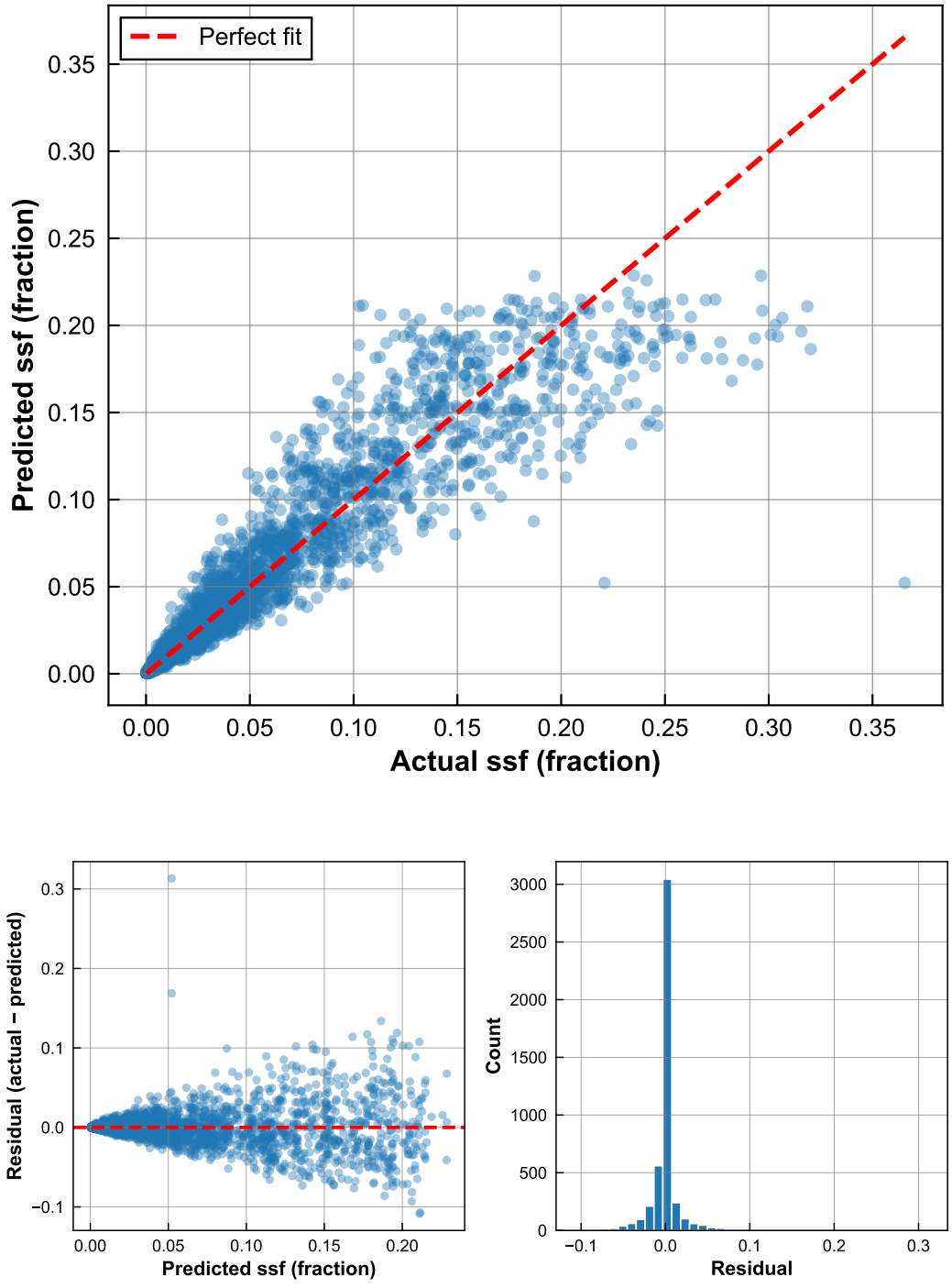


Figure 4.13: **Top** predicted against true Snomax surface fraction on the test set. **Bottom** the residual distribution, centred on zero with a width that grows with the predicted composition.

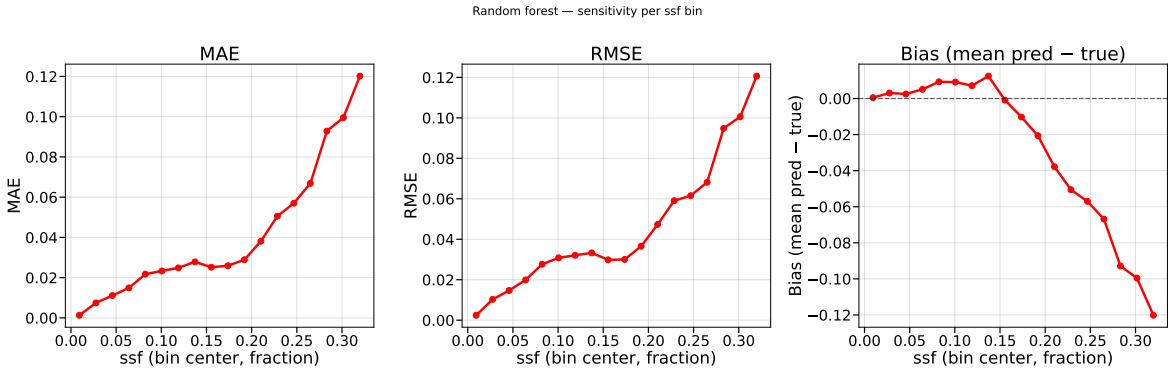


Figure 4.14: Test error per composition bin. The bin-wise MAE and RMSE grow towards the upper end of the retained range, and the bias turns increasingly negative above a composition of about 0.15, where the spectral shape saturates and stops encoding the composition.

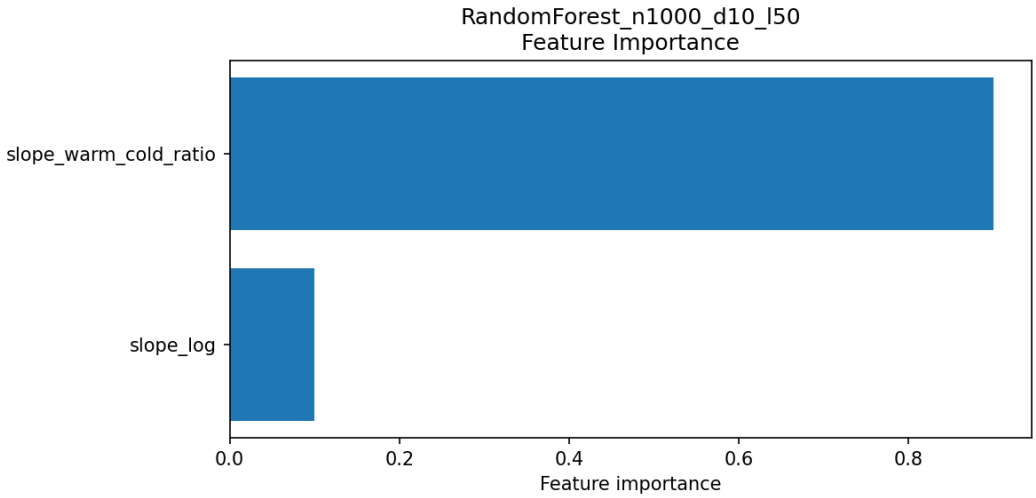


Figure 4.15: Impurity-based relative importance of the two features used by the selected random forest. The high temperature to low temperature slope ratio has the larger contribution, followed by the overall log-spectrum slope.

> not an artefact of the model but a property of the measurement problem, and any retrieval built on shape features will encounter it.

Two qualifications should be kept in mind when these numbers are read. All reported metrics are conditional on the mixed regime, so they describe the performance where the composition is recoverable in principle and not over the full range $[0, 1]$. And the whole evaluation rests on synthetic data, in which the noise and the parameterisation uncertainty follow the models imposed in Chapter 3; whether real spectra depart from those models in a way that matters can only be settled on measured mixtures.

5 Conclusions and Outlook

This chapter presents the main conclusions of the thesis. First, the research questions are answered based on the results of the study. The main limitations of the approach are then discussed, followed by recommendations for future research.

5.1 Answers to the research questions

Can a machine learning model recover the Snomax surface fraction, a continuous composition parameter, from a single perturbed synthetic INP spectrum of a mixture of dust and Snomax particles, and with what accuracy? Yes. The selected model reaches a coefficient of determination of $R^2 = 0.90$ on the held-out test set, with a mean absolute error of 0.008 and a root-mean-square error of 0.018 in surface fraction. The bias is negligible, about 2×10^{-4}, so the error is scatter and not a systematic offset, and about 97% of the test predictions lie within ±0.05 of the true fraction.

Which machine learning model gives the best result for this task, and by which criteria should it be chosen among the candidates? A random forest was selected. It is not the model with the single best cross-validated score, but it lies inside one standard error of the best score together with the three boosted-tree models, and it is the simplest member of that group, so the parsimony criterion selects it. The comparison also showed that the retrieval is nonlinear: the tree-based and neural models separate clearly from the linear family at small feature counts, where a nonlinear model already reaches $R^2 \approx 0.89$ from a single shape feature while the linear baseline needs about three features to approach the same level. The final model uses two features, the warm-to-cold slope ratio and the overall steepness of the log-spectrum. Both have a direct physical reading. The slope ratio compares the higher temperature side of the spectrum, where Snomax is active, with the cold side, where dust is active, so it is a composition-sensitive descriptor of the mixing ratio. The overall steepness reflects that the two species have different characteristic slopes in log space, so the slope of the mixture tracks the composition. Together with the spread of the freezing-density spectrum, these two features carry about 96% of the SHAP importance of the seven surviving candidates, with a drop of more than an order of magnitude to the remaining features. The model therefore uses the temperature regions where the two species are expected to differ, which is the behaviour required for the result to be physically meaningful and not an artefact of the dataset.

What detectability criterion determines when either component's contribution becomes too weak to detect relative to the prescribed pointwise spectrum variability, and how does the resulting threshold restrict the composition range over which the retrieval is meaningful? The adopted detectability criterion compares the contribution of each species across the temperature grid with the prescribed log-space pointwise variability scale, using the study-specific threshold $k = 5$. The criterion is evaluated separately for dust and Snomax, assigning every generated spectrum to one of four regimes: mixed, dust-dominated, Snomax-dominated or undetectable. Only the mixed regime, in which both species satisfy the criterion, is retained for model training and evaluation.

For the dust-specific diagnostic, evaluated using the unperturbed fitted parameterisations and the dilute-activation approximation, the total surface-area amplitude cancels from the fractional contribution. The resulting dust-detectability threshold is $R^* = 0.0701$. Above this Snomax surface fraction, the dust contribution no longer satisfies the adopted $k = 5$ criterion, although it does not become zero and may remain visible at low temperatures in an unperturbed spectrum.

This criterion restricts the interpretation of the reported performance: the model metrics apply only to spectra classified as mixed and not to the full composition range $[0, 1]$. The increase in error towards high Snomax fractions is consistent with reduced dust detectability and with the mixed-spectrum shape becoming less sensitive to R. Retrievals based only on comparable spectral-shape information may encounter a similar limitation, although the present study does not establish that every retrieval method would have exactly the same boundary.

5.2 Limitations

The results above are conditional on the assumptions of the synthetic framework, and the following limitations should be kept in mind when interpreting them.

- **Synthetic data only.** Every spectrum was generated using the synthetic spectrum generation framework. The retrieval was therefore trained and evaluated on spectra based on the same parameterisations and assumptions. The reported skill may overestimate the performance achievable on real measurements, for which the underlying INAS behaviour, perturbation structure and particle population may differ from those represented in the synthetic framework.
- **Experimental surface-area uncertainty.** The numerical precision achieved within the synthetic framework should not be interpreted as equivalent precision in a real measurement. The target variable R is defined from the component surface-area concentrations, $R = S_{\mathrm{snomax}}/S_{\mathrm{tot}}$. In the synthetic dataset these quantities, and therefore R, are known exactly by construction. In an experiment, however, aerosol surface-area concentration must be determined from measurements and is itself subject to substantial uncertainty. Consequently, very small differences in RMSE that can be resolved

numerically in the synthetic experiment should not be interpreted as physically resolvable differences in a real-world retrieval. The reported RMSE therefore describes the performance of the retrieval under the assumptions of the synthetic framework and not the precision with which the Snomax surface fraction could currently be determined experimentally.

- **Two components with well-separated regimes.** Dust and Snomax were chosen precisely because their ice nucleation activities have contrasting temperature dependences, which makes the composition legible in the shape of the spectrum. Real samples contain several ice-active components, some of which have similar spectral shapes, and the separability demonstrated here should not be assumed to carry over to them.
- **Representativeness of high Snomax fractions.** The synthetic sampling intentionally includes spectra in which the Snomax contribution dominates. These cases are useful for testing the retrieval and identifying its detectability limits, but they should not be interpreted as evidence that comparable biogenic dominance is common in ambient INP spectra. Snomax is a laboratory reference material, and the sampled surface fractions were selected to cover the prescribed composition range rather than to reproduce an observed atmospheric composition distribution. The results at high Snomax fractions therefore describe the synthetic experiment and not the expected frequency of such conditions in the atmosphere.
- **Additivity and external mixing.** The mixed spectrum is the sum of the two species' contributions, with no interaction between components and no internal mixing. Internal mixing is one of the main difficulties of the real attribution problem and is not represented.
- **Singular description.** The time dependence of freezing is neglected, so cooling-rate effects and the dependence of the apparent n_s on the experimental design are outside the model.
- **Idealised variability model.** The prescribed pointwise spectrum variability is independent and identically distributed in log space at every temperature point, and the INAS parameterisation perturbation is a two-parameter amplitude-plus-tilt form. Real measurements also carry systematic biases, temperature offsets, correlated errors and detection limits. The amplitude widths are grounded in reported errors, but the tilt widths are a modelling choice for which no published slope error exists, so the amount of structured variability is not fully constrained by measurements.
- **Conditional performance.** Only the mixed regime enters the training and evaluation, and within it the retained samples are concentrated at small surface fractions. The behaviour of the retrieval at high Snomax fractions is therefore poorly constrained, both because such samples are rare in the data and because the dust signal there is physically weak.
- **Idealised sampling of the spectrum.** Every sample is available on the same dense temperature grid over the full range. Real spectra are measured at fewer temperatures,

over narrower and instrument-specific ranges, and are often censored at the low-concentration end, which would affect the shape features directly.

5.3 Outlook

Several directions follow directly from these limitations.

- **Validation on real laboratory mixtures.** The most immediate next step is to apply the trained retrieval to measured dust and Snomax spectra mixed at controlled and known ratios. Such an experiment would test the transfer from synthetic to real spectra under conditions where the true composition is still known, and would show whether the two shape features remain informative when the variability structure of measured spectra replaces the modelled one.
- **More components.** Extending the target from a single fraction to a composition vector on the simplex would turn the retrieval into a genuine multi-component problem. This would also allow the harder case of components with overlapping temperature ranges to be studied, where the shape of the spectrum is expected to be far less informative.
- **Covering the full composition range with a two-stage approach.** Instead of restricting the dataset to one regime, a first-stage classifier could assign each spectrum to a detectability regime. A second-stage regression model trained for that regime could then estimate the composition when the relevant components are detectable. For regimes in which the composition is not recoverable, the method would report this limitation rather than return an unreliable estimate.
- **Reference shapes from real data.** The shape-similarity and decomposition features were excluded here because they used the exact synthesis references. If reference spectra were instead derived from independent measurements, these features could be reintroduced without the conceptual leakage, and could be evaluated as potential predictors.
- **Reporting uncertainty instead of a point estimate.** The error of the present model grows with the true fraction, so a single number is a poor summary of the retrieval near the detectability limit. Quantile regression or a conformal prediction interval would return a composition together with a range, which is what a retrieval product used in an atmospheric context should provide.
- **Application to field measurements.** Once validated in the laboratory, the approach could be applied to ambient spectra from field campaigns, where the biological and mineral contributions are exactly the quantities that are currently inferred qualitatively. A composition estimate with a stated uncertainty would be a useful input for INP parameterisations in cloud-resolving and climate models, where the split between biological and mineral ice nucleation is still one of the larger sources of uncertainty.

5.4 Concluding remarks

The central outcome of this work is that the composition of a two-component ice-nucleating particle mixture is contained in the shape of its immersion-freezing spectrum, and that a machine-learning model can read it back from a single pointwise-perturbed synthetic spectrum with a coefficient of determination of 0.90, using only two scale-invariant shape features and a model as simple as a random forest. Equally important is the second outcome, that the accuracy of the retrieval is bounded by a detectability limit which follows from the prescribed pointwise-variability level and the composition alone. The limit can be computed in advance, it explains where the retrieval degrades, and it separates what is a shortcoming of the method from what is a property of the measurement. Taken together, these two results suggest that composition attribution for ice-nucleating particles is a tractable regression problem, provided that the retrieval is stated together with the detectability regime in which it is valid.

Acknowledgements

I would like to thank Prof. Dr. Corinna Hoose and Prof. Dr. Thomas Leisner for giving me the opportunity to work on this topic and for supervising this thesis. Their questions and comments during our meetings helped me to keep the work focused and to see which parts of it were scientifically relevant.

I am especially grateful to my advisors, Dr. Gabriella Linnea Wallentin and Dr. Alexander Böhmländer, for their continuous support during the whole project. They took the time to discuss the details of the parameterizations, the noise model and the machine learning results with me, they read my drafts carefully, and they gave me clear and constructive feedback whenever I was stuck. I owe a particular debt to Dr. Alexander Böhmländer for his sharp mind and for the brilliant ideas he brought into our discussions, which more than once opened a way forward when I could not see one. This thesis benefited a lot from their guidance.

I would like to thank Dr. Ottmar Möhler for his insight into the field and for helping me to understand where the open gaps are, which shaped the direction of this work. I am also thankful to Dr. Kristina Höhler for helping me to see the parts of the problem that were not clear to me and for explaining them patiently.

I would also like to thank the members of the group of Dr. Ottmar Möhler at IMK-AAF and the members of the Cloud Physics group at IMK-TRO, Karlsruhe Institute of Technology, for the friendly working atmosphere and for the discussions that helped me to better understand ice nucleation and the experimental side of INP measurements.

Finally, I thank my family and my friends for their patience and their encouragement during my studies, especially Roya Pourshahbaz, whose support made this work much easier for me to carry out.

Bibliography

Abadi, M. et al. (2016). "TensorFlow: A System for Large-Scale Machine Learning". In: *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*. USA: USENIX Association, pp. 265–283. ISBN: 9781931971331.

Abedin, T., H. Xu, and S. Uddin (2026). "The impact of K selection in K-fold cross-validation on bias and variance in supervised learning models". In: *Scientific Reports* 16.1, p. 6084. DOI: 10.1038/s41598-026-37247-x.

Atkinson, J. D. et al. (2013). "The importance of feldspar for ice nucleation by mineral dust in mixed-phase clouds". In: *Nature* 498.7454, pp. 355–358. ISSN: 1476-4687. DOI: 10.1038/nature12278.

Berner, J. et al. (2017). "Stochastic Parameterization: Toward a New View of Weather and Climate Models". In: *Bulletin of the American Meteorological Society* 98.3, pp. 565–588. DOI: 10.1175/BAMS-D-15-00268.1.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Vol. 4. Springer New York.

Breiman, L. (2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32. DOI: 10.1023/A:1010933404324.

Burrows, S. M. et al. (2022). "Ice-Nucleating Particles That Impact Clouds and Climate: Observational and Modeling Research Needs". In: *Reviews of Geophysics* 60.2, e2021RG000745. DOI: 10.1029/2021RG000745.

Canzi, A. et al. (2025). "Unraveling Ice Nucleating Particle Concentration Variability: Insights Into Source Emissions Origin and Parameterizations". In: *Journal of Geophysical Research: Atmospheres* 130.8, e2024JD041258. DOI: 10.1029/2024JD041258.

Chen, T. and C. Guestrin (2016). "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, pp. 785–794. ISBN: 9781450342322. DOI: 10.1145/2939672.2939785.

Christopoulos, C. D. et al. (2018). "A machine learning approach to aerosol classification for single-particle mass spectrometry". In: *Atmospheric Measurement Techniques* 11.10, pp. 5687–5699. DOI: 10.5194/amt-11-5687-2018.

Connolly, P. J. et al. (2009). "Studies of heterogeneous freezing by three different desert dust samples". In: *Atmospheric Chemistry and Physics* 9.8, pp. 2805–2824. DOI: 10.5194/acp-9-2805-2009.

David, R. O. et al. (2019). "Development of the DRoplet Ice Nuclei Counter Zurich (DRINCZ): validation and application to field-collected snow samples". In: *Atmospheric Measurement Techniques* 12.12, pp. 6865–6888. DOI: 10.5194/amt-12-6865-2019.

DeMott, P. J. et al. (2010). "Predicting global atmospheric ice nuclei distributions and their impacts on climate". In: *Proceedings of the National Academy of Sciences* 107.25, pp. 11217–11222. DOI: 10.1073/pnas.0910818107.

Efron, B. (1979). "Bootstrap Methods: Another Look at the Jackknife". In: *The Annals of Statistics* 7.1, pp. 1–26. DOI: 10.1214/aos/1176344552.

Friedman, J. H. (2001). "Greedy function approximation: A gradient boosting machine". In: *The Annals of Statistics* 29. DOI: 10.1214/aos/1013203451.

Garimella, S. et al. (2017). "Uncertainty in counting ice nucleating particles with continuous flow diffusion chambers". In: *Atmospheric Chemistry and Physics* 17.17, pp. 10855–10864. DOI: 10.5194/acp-17-10855-2017.

Georgakaki, P. et al. (2022). "Secondary Ice Production Processes in Wintertime Alpine Mixed-Phase Clouds". In: *Atmospheric Chemistry and Physics* 22.3, pp. 1965–1988. DOI: 10.5194/acp-22-1965-2022.

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep Learning*. MIT Press.

Guyon, I. and A. Elisseeff (2003). "An Introduction to Variable and Feature Selection". In: *Journal of Machine Learning Research* 3, pp. 1157–1182. ISSN: 1532-4435.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer Series in Statistics. New York, NY: Springer, pp. XXII, 745. ISBN: 978-0-387-84857-0. DOI: 10.1007/978-0-387-84858-7.

Heintzenberg, J. (1994). "Properties of the Log-Normal Particle Size Distribution". In: *Aerosol Science and Technology* 21.1, pp. 46–48. DOI: 10.1080/02786829408959695.

Herbert, R. J. et al. (2014). "Representing time-dependent freezing behaviour in immersion mode ice nucleation". In: *Atmospheric Chemistry and Physics* 14.16, pp. 8501–8520. DOI: 10.5194/acp-14-8501-2014.

Herbert, R. J. et al. (2025). "Gaps in our understanding of ice-nucleating particle sources exposed by global simulation of the UK Earth System Model". In: *Atmospheric Chemistry and Physics* 25.1, pp. 291–325. DOI: 10.5194/acp-25-291-2025.

Hiranuma, N. et al. (2015). "A comprehensive laboratory study on the immersion freezing behavior of illite NX particles: a comparison of 17 ice nucleation measurement techniques". In: *Atmospheric Chemistry and Physics* 15.5, pp. 2489–2518. DOI: 10.5194/acp-15-2489-2015.

Hoose, C. and O. Möhler (2012). "Heterogeneous ice nucleation on atmospheric aerosols: a review of results from laboratory experiments". In: *Atmospheric Chemistry and Physics* 12.20, pp. 9817–9854. DOI: 10.5194/acp-12-9817-2012.

Hornik, K., M. Stinchcombe, and H. White (1989). "Multilayer feedforward networks are universal approximators". In: *Neural Networks* 2.5, pp. 359–366. ISSN: 0893-6080. DOI: 10.1016/0893-6080(89)90020-8.

Huang, S. et al. (2021). "Overview of biological ice nucleating particles in the atmosphere". In: *Environment International* 146, p. 106197. ISSN: 0160-4120. DOI: 10.1016/j.envint.2020.106197.

Kanji, Z. A. et al. (2017). "Overview of Ice Nucleating Particles". In: *Meteorological Monographs* 58, pp. 1.1–1.33. DOI: 10.1175/AMSMONOGRAPHS-D-16-0006.1.

Ke, G. et al. (2017). "LightGBM: A Highly Efficient Gradient Boosting Decision Tree". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc.

Knopf, D. A. and P. A. Alpert (2023). "Atmospheric ice nucleation". In: *Nature Reviews Physics* 5.4, pp. 203–217. ISSN: 2522-5820. DOI: 10.1038/s42254-023-00570-7.

Knopf, D. A. et al. (2020). "Stochastic nucleation processes and substrate abundance explain time-dependent freezing in supercooled droplets". In: *npj Climate and Atmospheric Science* 3.1, p. 2. ISSN: 2397-3722. DOI: 10.1038/s41612-020-0106-4.

Korolev, A. and G. Isaac (2003). "Phase Transformation of Mixed-Phase Clouds". In: *Quarterly Journal of the Royal Meteorological Society* 129.587, pp. 19–38. DOI: 10.1256/qj.01.203.

Korolev, A. et al. (2022). "Observation of Secondary Ice Production in Clouds at Low Temperatures". In: *Atmospheric Chemistry and Physics* 22.19, pp. 13103–13113. DOI: 10.5194/acp-22-13103-2022.

Lawson, C. L. and R. J. Hanson (1995). *Solving Least Squares Problems*. Vol. 15. Classics in Applied Mathematics. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM), pp. xii+337. ISBN: 0-89871-356-0.

Louppe, G. (2014). "Understanding Random Forests: From Theory to Practice". In: *arXiv: Machine Learning*.

Lundberg, S. M. and S.-I. Lee (2017). "A unified approach to interpreting model predictions". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., pp. 4768–4777. ISBN: 9781510860964.

Lundberg, S. M. et al. (2020). "From local explanations to global understanding with explainable AI for trees". In: *Nature Machine Intelligence* 2.1, pp. 56–67. DOI: 10.1038/s42256-019-0138-9.

Manning, C. D., P. Raghavan, and H. Schütze (2008). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

May, A. A. and H. Li (2022). "Application of machine learning approaches in the analysis of mass absorption cross-section of black carbon aerosols: Aerosol composition dependencies and sensitivity analyses". In: *Aerosol Science and Technology* 56.11, pp. 998–1008. DOI: 10.1080/02786826.2022.2114312.

McKay, M. D., R. J. Beckman, and W. J. Conover (1979). "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code". In: *Technometrics* 21.2, pp. 239–245. DOI: 10.2307/1268522.

Möhler, O. et al. (2006). "Efficiency of the deposition mode ice nucleation on mineral dust particles". In: *Atmospheric Chemistry and Physics* 6.10, pp. 3007–3021. DOI: 10.5194/acp-6-3007-2006.

Möhler, O. et al. (2008). "Heterogeneous ice nucleation activity of bacteria: new laboratory experiments at simulated cloud conditions". In: *Biogeosciences* 5.5, pp. 1425–1435. DOI: 10.5194/bg-5-1425-2008.

Murray, B. J. et al. (2012). "Ice nucleation by particles immersed in supercooled cloud droplets". In: *Chem. Soc. Rev.* 41.19, pp. 6519–6554. DOI: 10.1039/C2CS35200A.

Niemand, M. et al. (2012). "A Particle-Surface-Area-Based Parameterization of Immersion Freezing on Desert Dust Particles". In: *Journal of the Atmospheric Sciences* 69.10, pp. 3077–3092. DOI: 10.1175/JAS-D-11-0249.1.

Orser, C. et al. (1985). "Cloning and expression of bacterial ice nucleation genes in Escherichia coli". In: *Journal of Bacteriology* 164.1, pp. 359–366. DOI: 10.1128/jb.164.1.359-366.1985.

Palmer, T. N. et al. (Oct. 2009). *Stochastic parametrization and model uncertainty*. eng. DOI: 10.21957/ps8gbwbdv.

Phillips, V. T. J., P. J. DeMott, and C. Andronache (2008). "An Empirical Parameterization of Heterogeneous Ice Nucleation for Multiple Chemical Species of Aerosol". In: *Journal of the Atmospheric Sciences* 65.9, pp. 2757–2783. DOI: 10.1175/2007JAS2546.1.

Schrod, J. and H. G. Bingemer (2025). "A view on recent ice-nucleating particle intercomparison studies: why the uncertainty of the activation temperature matters". In: *Atmospheric Measurement Techniques* 18.12, pp. 2591–2605. DOI: 10.5194/amt-18-2591-2025.

Ullrich, R. et al. (2017). "A New Ice Nucleation Active Site Parameterization for Desert Dust and Soot". In: *Journal of the Atmospheric Sciences* 74.3, pp. 699–717. DOI: 10.1175/JAS-D-16-0074.1.

Vali, G. (1971). "Quantitative Evaluation of Experimental Results an the Heterogeneous Freezing Nucleation of Supercooled Liquids". In: *Journal of Atmospheric Sciences* 28.3, pp. 402–409. DOI: 10.1175/1520-0469(1971)028<0402:QEOERA>2.0.CO;2.

Vali, G. et al. (2015). "Technical Note: A proposal for ice nucleation terminology". In: *Atmospheric Chemistry and Physics* 15.18, pp. 10263–10270. DOI: 10.5194/acp-15-10263-2015.

Viering, T. and M. Loog (2023). "The Shape of Learning Curves: A Review". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.6, pp. 7799–7819. DOI: 10.1109/TPAMI.2022.3220744.

Wang, G. et al. (2024). "Machine learning approaches for automatic classification of single-particle mass spectrometry data". In: *Atmospheric Measurement Techniques* 17.1, pp. 299–313. DOI: 10.5194/amt-17-299-2024.

Wang, Y. et al. (Nov. 2025). "Advancing Aerosol Chemistry with Machine Learning: A Short Review". In: *ACS EST Air* 2.11, pp. 2323–2341. DOI: 10.1021/acsestair.5c00001.

Wex, H. et al. (2015). "Intercomparing different devices for the investigation of ice nucleating particles using Snomax as test substance". In: *Atmospheric Chemistry and Physics* 15.3, pp. 1463–1485. DOI: 10.5194/acp-15-1463-2015.

Whale, T. F. et al. (2015). "A technique for quantifying heterogeneous ice nucleation in microlitre supercooled water droplets". In: *Atmospheric Measurement Techniques* 8.6, pp. 2437–2447. DOI: 10.5194/amt-8-2437-2015.

Yakobi-Hancock, J. D., L. A. Ladino, and J. P. D. Abbatt (2013). "Feldspar Minerals as Efficient Deposition Ice Nuclei". In: *Atmospheric Chemistry and Physics* 13.22, pp. 11175–11185. DOI: 10.5194/acp-13-11175-2013.