TinyTrail: A Lightweight Transformer Model for Contrail Risk Nowcasting

by

Omar Hayat

A thesis
presented to the University of Waterloo
in fulfillment of the
research paper requirement for the degree of
Masters of Mathematics
in
Computational Mathematics

Waterloo, Ontario, Canada, 2025

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Data-driven paradigms for atmospheric forecasting now challenge the leading numerical weather prediction (NWP) techniques, offering competitive accuracy and superior inference speeds. This work draws on architectural decisions by leading data-driven models to present a lightweight transformer based approach for tackling the forecasting task of condensation-trail (contrail) risk prediction. Contrails are artificial cloud-like structures formed in the upper atmosphere when aircraft fly through ice-supersaturated regions, and are the main actor in global warming from air travel. This has motivated the development of reliable contrail avoidance systems, in which contrail risk prediction plays a key role. Our work highlights the contrail risk prediction capabilities of our lightweight transformer model and the architectural decisions that enable it to learn complex dynamical-system relationships.

Acknowledgements

I would like to thank my supervisors Dr. Giang Tran and Dr. Saeed Ghadimi for their patience and support throughout my masters degree. I've grown immensely in the skill of doing research in large part due to their investment in me.

Dedication

To my family.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The term contrails is used colloquially in reference to condensation trails. Condensation trails are artificial cloud-like particle structures observable as white streaks that trail aircraft. These trails are well studied in their formation conditions [10, 42, 58, 41] and warming effect on the global climate [2, 46, 48]. Conseqeuence of this warming effect and the efficacy of slight flight path deviations on contrail prevention [18], the study of contrail avoidance systems has become a focal challenge in the aviation community.

Addressing contrail avoidance is a multi-faceted problem. It demands sustained research efforts into alternative fuel sources for air travel, operational systems for live aircraft re-routing, and predictive tooling that can accurately forecast contrail prone regions. Of these objectives, adoption of effective re-routing systems is the more immediately tractable given maturity of forecasting capabilities and the effectiveness of small deviations to flight path on contrail prevention. Furthermore, commercial aircraft already perform re-routing to avoide areas of strong turbulance which is almost operationally identical to avoidance of contrail risk regions.

Currently, there is no commercially deployed operational system that guide aircraft away from contrail risk zones. At present, only small-scale operational intervention studies and historical re-routing analyses have been done [36, 13]. Therefore, there is pressing need for development of a preliminary system. These systems are primarily limited by accurate measures of contrail risk at fine enough spatial and temporal resolutions [7]. Though performant numerical weather prediction models currently exist [4] and competitive rivaling data-driven atmospheric models [5, 30], operationally deployable contrail risk models do

not exist.

In this study, we design and evaluate a small transformer-based model to assist in the contrail avoidance task. The architecture decisions are inspired by other leading data-driven models [5, 19] for large scale atmospheric prediction. We adapt their techniques to fit our smaller architecture setting. We call our small transformer model for contrail risk prediction **TinyTrail**.

The layout of our report is as follows. We first cover study of the environmental impact of contrails, the atmospheric conditions under which their formation is favourable, and recent literature on contrail avoidance systems. We also provide background on leading numerical and data-driven atmospheric forecasting models. Following coverage of the background and related work, we motivate the design decisions in our model architecture for **TinyTrail** and our approach to predicting contrail risk regions. We then discuss our experimental set up and results. Lastly, we end with a discussion contextualizing our results in the larger contrail avoidance space and explore avenues for further research.

Chapter 2

Background and Related Work

Contrails (condensation trails) are white cloud-like structures that form in wake of aircraft. Though seemingly harmless, these trails contribute to a large fraction of the warming effect due to aviation. For this reason, they have become of pressing concern and led to substantial efforts researching measures of avoiding contrails partly or entirely, particularly in commecial settings. In this section we cover relevant background on contrail formation, climate impact, and avoidance. Following, we introduce the leading datasets for atmospheric modelling and the data-driven methods that use them. Additionally, we discuss where these data-driven methods stand in comparison to the leading numerical weather prediction (NWP) techniques.

2.1 Contrails

Documented observation of contrails dates back to late 1910s [55], however, serious research interest only started when the detectability of aircraft started to become a military interest in the 1940s [28]. The idea that the water vapour emitted from the engines could cause supersaturation with respect to the liquid water was suggested early (1921 [52]), but was dismissed for a long time till Schmidt (1941) [39, 40] and Appleman (1953) [1] developed a theory showing that contrail formation conditions can be modelled as a function of ambient pressure, humidity, temperature, and specific properties of the fuel combustion and engine efficency. This developed into the Schmidt-Appleman Criterion (SAC) for contrail formation. After formation, contrails will persist when the relative humidity with respect to ice (RHi) is above 100% [43]. Extensive observational study has further supported the

SAC as a valid threshold for contrail occurence [21]. The regions of atmosphere that are referred to as RHi>100% are called Ice Supper Saturated Regions (ISSRs).

2.1.1 Formation of Condensation Trails

Contrails exist in two varieties, persistent and cirrus. Persistent contrails retain their linear shape and typicall last 2-4 hours but can endure for 18 hours or more depending on atmospheric conditions [7, 26]. Contrail cirrus is the term for aged persistent contrails that have lost their initial linear shape. These cirrus spread spatially due to atmospheric factors like wind sheer and turbulence and eventually become indistinguishable from naturally occuring cirrus clouds. Together these two types of contrails are collectively referred to as aircraft-induced clouds or cloudiness (AIC) [7]. Worth noting is that contrails may also be extremely short-lived lasting only a few minutes, and after 10 minutes are defined as "Cirrus Homogenitus" by The World Meteorological Organization (WMO). However, only those contrails that peresist for extended periods of time are contributors to significant wamming effects [7].

Figure 2.1: Contrail types. (a) Exhaust contrail (photo by Josef P. Williams; Unterstrasser et al. 2012). (b) Aerodynamic contrail (photo by Dieter Klatt; Gierens et al. 2011). (c) Aircraft-induced lines and holes in supercooled liquid clouds (cloud-top temperatures −35°C to −25°C); section of image with blue border lines, near northwest corner of Texas (29 Jan 2007, NASA, Jeff Schmaltz, MODIS Rapid Response Team). (d) Contrail visible shortly behind B747-400 engines, 38,000 ft, −61°C, 28 May 2004; photo by Robert Falk. (e) "Soot cirrus" observed at DLR, Oberpfaffenhofen, 0905 [51]

The formation of contrails is a result of water vapour exhausted from an aircraft engine mixing into sufficiently cold humid air. The vapour is a byproduct of the combution process, which in favourable surrounding atmospheric conditions condenses and contributes to ice crystal growth. This condensation binds to solid carbon particles in the exhaust of the jet engines as well as onto the atmospheric aerosol particles [22]. Though a thermoodynamic theory defined by Schmidt and Appleman provides a framekwork for understanding their formation as a function of fuel source and engine type, there are still microphysical and microchemical processes that are not captured by their criterion (SAC) [56].

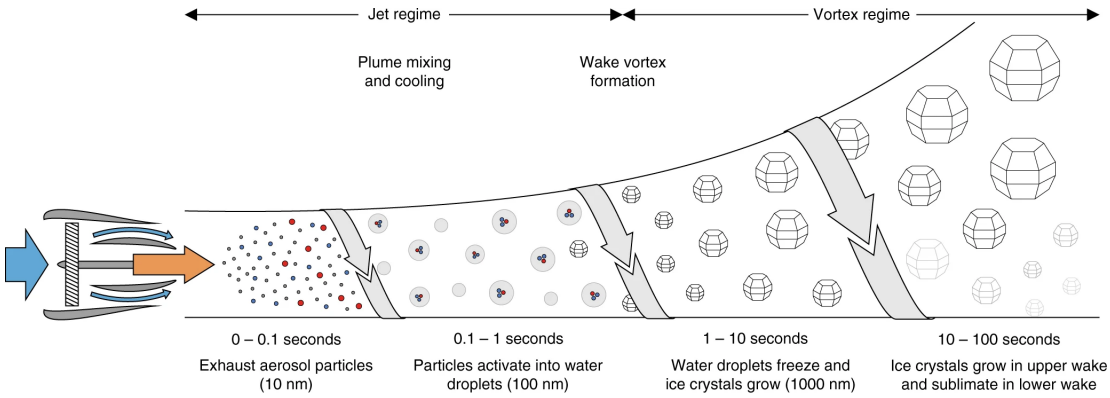

Figure 2.2: Early stage lifecycle of contrail formation. Full details available at [22].

A neccessary condition for contrail peristence is the presence of ice-supersaturated regions (ISSRs). These ISSRs are non-homogenous in their coverage with inconsistencies across both vertical and horizontal dimensions. Meaning, there is no general rule for what flight levels or geographic regions contrails may form in; pockets of ISSRs exist in irregular patterns [46]. For this reason, the primary operational avoidance measure being explored is contrail formation mitigation through trajectory adjustment [49, 50, 47, 58, 3, 18].

ISSRs are also relatively rare, in our study we had an average of approximate 8% of our region classified as ice-supersaturated (ISS). In another recent study (2023), it was observed that airspace over the UK was classified as ISSR approximate 10-15% of the time [25]. Furthemore, a recent modelling study of 40.2 million flights in 2019 found 5% of the total distance flown formed persistent contrails with the mean contrail segment lifetime of 2.4 hours [48].

Correlations have also been observed between contrail formation, the thermal tropopause, and maximal points along jet streams [57]. The thermal tropopause is the boundary between the troposphere and stratosphere where the lowest temperature in the atmosphere occurs. This layer starts around 9km over the poles to around 17-20km at the equator. In this layer, the combination of low temperatures, adiabatic cooling, and enchanced relative humidity is favourable for ice cloud formation [7].

After formation, a necessary condition for contrail persistence is ISS, identifiable by via an observed relative humidity with repsect to ice (RHi) above 100%. Other measures of contrail sensitive areas exist that take into account both ISS and the predicted net warming effect on the atmosphere. The leading metric is the Schmidt-Appleman criterion, which takes into account the thermal efficiency of the engine, the lower heating value of the fuel, and the relative humidity of the surrounding air [21]. Although the SAC condition considers more factors in modelling contrail risk, it introduces significant complexity when designing a machine learning model. In our setting we choose not to use the SAC for modelling contrail risk of a region.

For simplicity, we leverage equation 2.3 and model contrail risk when this quantity, RHi $> 100\%$.

$$e_{si}(T) = 6.112 \exp\left(\frac{22.46\,(T - 273.15)}{T - 0.55}\right) \tag{2.1}$$

$$e = \frac{q\,p}{0.622 + 0.378\,q} \cdot \frac{1}{100} \tag{2.2}$$

$$\text{RHi} = \frac{e}{e_{si}(T)} \tag{2.3}$$

For our purposes, we assign binary labels to our geographic and temporal datapoints as a function of the observed temperature (T), specific humidity (q), and pressure (p) at that point. More formally for a given dataset $\mathcal{X}$, of spatio-temporal tensors $X \in \mathbb{R}^{H \times W \times T \times C}$, we assign to each X label $Y \in \{0, 1\}$ where,

$$Y = \mathbf{I}(\text{RHi}(q_X, p_X, t_X) > 100).$$

Here, q_X, p_X, and t_X denote the relative hummidity, pressure, and temperature variables over the region X and $\mathbf{I}(\cdot)$ denotes the indicator function with value 1 when RHi $> 100\%$ is observed and 0 otherwise. In this setting our tensor X spans spatial region $H \times W$,

temporal history T, and features C. In our study we selected 10 atmospheric variables to make up C but only q, p, and t are required for identifying ice-supersaturation. We admit that as a proxy for contrail risk, strict RHi thresholding is crude, however for preliminary study setting we deem this sufficient.

2.1.2 Climate Impact

The standard measure of climate impact due to man-made (anthropogenic) factors is Radiative Force in watts per meter squared (W/m^2) [7]. The Intergovernmental Pannel on Climate Change (IPCC) estimates the total annual anthropogenic Net Radiative Forcing (NRF) at 2.38 W/m^2, of which aviation's total contribution is 0.090 W/m^2, approximately 3.7% [26]. Contrails alone are estimated to have a NRF of 0.050 W/m^2, which makes up 55% of aviation's total net warming effect.

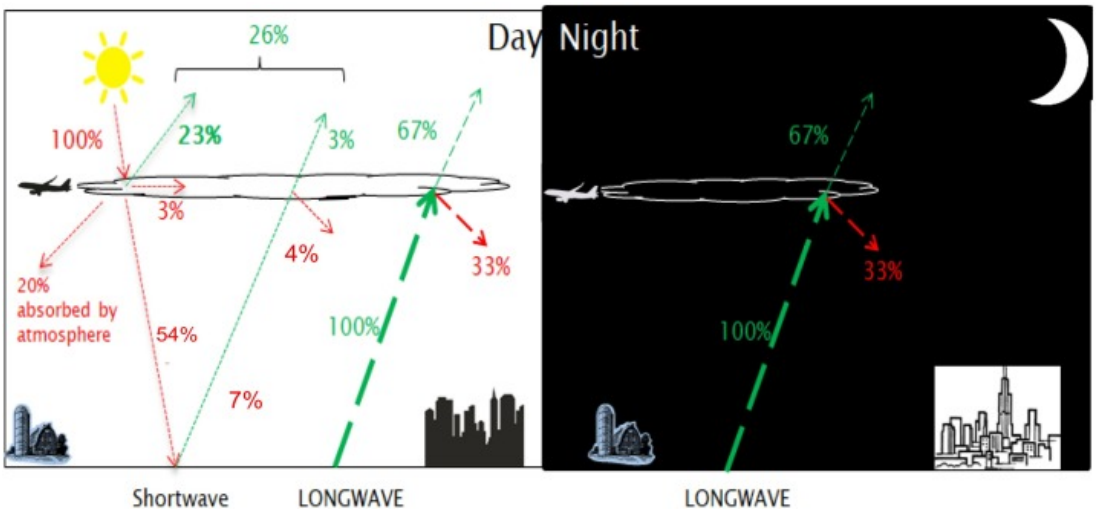

Figure 2.3: Effect of contrails on incoming shortwave radiation and longwave radiation. Taken from [8].

Contrails reflect away some incoming light from the sun in the form of shortwave radiation due to their albedo effect. However, they cause a blanket effect which keeps warmpth trapped in the lower atmosphere through the absoption of outgoing longwave radiation from the earth's surface [8]. Moreover, this effect prevails overnight resulting in a 24-hour warming effect. It's been shown that in general the warming effect is greater for flights during the night than during the day [47]. A recent study has found, based on real North

Atlantic flights in 2023 and 2024, that $\sim 2.8\%$ of flights sampled accounted for 80% of the total radiative force [16].

One key difference between warming due to contrails and warming due to release of gases (such as CO_2 and nitrogen oxides, NO_X) is that of immediacy. The warming effect due to contrails is experienced immediately in the atmosphere, as opposed to gas emissions which are projected to have an effect 20-40 years from now.

2.1.3 Current Work on Contrail Avoidance

There exist models that inform radiative forcing impact of a single contrail during its lifespan, that are utilized by studies that measure the effect of contrail formation. This plays a role in avoidance studies as well to derive measures of usefulness of interventions. The leading model is the Contrail Cirrus Prediction Tool (CoCiP) [41], which is computationally efficent but the nature of it makes it incompatible with flight planning. CoCiP is useful as a way to evaluate contrail formation and simulate evolution, however lacks instantaneous reactivity to be compatible with trajectory optimization methodologies.

As a result, work on contrail avoidance has largely been limited to simulation studies, with only two actual flight trials have been documented, both in low-density traffic situations [36, 13]. The avoidance strategy can be planned ahead of flight time, or implemented during the flight. Trajectory adjustments can be made vertically or horizontally, where vertical interventions are significantly more effective from a cost perspective due to the large width of ISSRs [45]. These regions can have large horizontal spread ($150 \pm 250\ km$) but shallow vertical spread on the order of 1000 to 2000 feet [7].

A study of the Japanese air-psace estimates that a deviation of 0.5-4.1% of flights can lead to a contrail radiative force reduction of 50 to up to 93%, depending on atmospheric variations [50]. Other recent studies have shown that only 2-16% of flight plans need to be adjusted to avoid 54-80% of contrail-induced warming, depending on the location, season, and mateorological conditions [7, 8, 16].

Efforts are beginning to inform customers of contrail risks to mitigate contrail formation through informed customer flight scheduling. Travellers that book flights through Google now are notified of the contrail risk of the flight, motivating the informed customer to take decisions to mitigate their contrail footprint.

2.2 Atmospheric Datasets in Climate Research

Modern climate research relies on large, coherent atmospheric reanalysis datasets that provide physically consistent estimates of the global state of the atmosphere. The leading product is the ERA5 dataset [20], produced by ECMWF, which offers hourly global fields at high spatial resolution and incorporates satellite, radiosonde, and surface observations through an advanced data assimilation system. ERA5 has become the de facto standard for climate modelling, validation, and long-term atmospheric studies.

Other reanalysis products complement ERA5 in coverage or methodological emphasis. NASA's MERRA–2 dataset [17] is optimized for aerosol, chemistry, and radiation applications; NOAA's CFSR and CFSv2 reanalyses [34, 35] are widely used in subseasonal-to-seasonal prediction; and the JRA–55 dataset from the Japan Meteorological Agency [23] provides a long, homogeneous record suitable for climate variability and trend analysis. Together, these datasets form the backbone of contemporary climate research, enabling reproducible modelling, evaluation of atmospheric processes, and construction of climatologies across multiple temporal and spatial scales.

2.3 Data-Driven and Numerical Prediction Models

Recent progress in atmospheric prediction reflects two parallel modelling paradigms: (i) numerical models derived from physical principles and discretized differential equations, and (ii) data-driven models that learn dynamical relationships directly from historical atmospheric states. Numerical weather prediction (NWP) remains the operational standard for medium–range forecasting [4], while data-driven models have demonstrated competitive skill at global and regional prediction horizons [38]. Understanding the distinctions and complementarities between these approaches is critical when designing specialized predictors, such as models for contrail-risk classification or upper–tropospheric moisture dynamics.

2.3.1 Numerical and Physical Models

Global Climate Models (GCMs) and numerical weather prediction systems solve approximations to the governing equations of atmospheric motion, namely, the Navier–Stokes equations under specific assumptions, coupled with radiation, thermodynamics, and cloud

microphysics parameterizations [54, 4]. Modern operational NWPs, such as ECMWF's Integrated Forecasting System (IFS), combine these physical equations with variational or ensemble-based data assimilation to produce globally consistent initial states [31]. GCMs extend this framework to multi-decadal time scales, resolving climate feedbacks such as water vapour, cloudiness, and circulation changes [15]. Although physically interpretable and highly constrained, these models require substantial computational resources and rely on parameterizations for unresolved processes (e.g., convection, cloud formation, sub-grid turbulence).

2.3.2 Data-Driven Models

Deep learning systems have emerged as flexible forecast models capable of emulating atmospheric dynamics without explicit physical parameterizations. Early approaches employed convolutional architectures such as ResNets and U-Nets for downscaling, post-processing, and limited-area prediction [33, 9]. More recent work focuses on transformer-based architectures, motivated by their ability to process high-dimensional spatial fields and capture long-range dependencies.

FourCastNet [30] demonstrated that a Fourier Neural Operator (FNO) combined with a lightweight transformer-style architecture can perform competitive global medium-range forecasts. GraphCast [24] extends this idea using graph neural networks to propagate information along a spherical mesh, achieving state-of-the-art skill among purely data-driven models.

ClimaX [29] adopts a modular transformer backbone based on a vanilla Vision Transformer (ViT), where variables are embedded via variable-tokenization and aggregated through learned attention maps. This design allows flexible conditioning on arbitrary variable subsets or predictive tasks, making it suitable for broad climate-modelling contexts.

Pangu-Weather [5] uses a Swin-transformer-based backbone with hierarchical spatial attention and introduces *Earth-Specific Positional Bias* (ESB), which encodes latitudinal anisotropy in atmospheric dynamics. ESB improves interpolation across latitude bands and supports long-range spatial dependencies critical for global prediction. Pangu's architecture also incorporates 3D Swin layers for vertical coupling and demonstrates strong

deterministic forecast skill.

Recent work explores faster transformer variants such as the Faster-Swin Transformer [59], which improves computational throughput via window-shift optimization and sparse attention, suggesting potential scalability for climate and weather prediction workloads.

Transformer Backbones in Climate Models

Transformers are used as the principal encoder in several models: (1) ClimaX (ViT), (2) Pangu-Weather (3D Swin), (3) FourCastNet (hybrid FNO + attention), (4) GraphCast (message–passing GN with transformer-like aggregation), (5) Swin-based regional models in downscaling and extreme-event prediction. Across these systems, the transformer serves as the foundational operator for representing spatial–temporal dependencies.

Chapter 3

Approach

3.1 Dataset

We leverage the ERA5 hourly data on pressure levels from 1940 to present [11] for our study. ERA5 is the fifth-generation global atmospheric reanalysis produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) under the Copernicus Climate Change Service. It provides physically consistent, hourly three-dimensional atmospheric fields on standard pressure levels from 1940 to the present. For contrail forecasting, ERA5 supplies the thermodynamic variables required to diagnose ice supersaturation, including temperature T, specific humidity q, and pressure p at upper-tropospheric flight levels (typically 200–300 hPa). These fields enable direct computation of relative humidity with respect to ice, $\mathrm{RHi}(T, p, q)$, which governs contrail persistence under ice-supersaturated conditions.

The geographic domain considered in this study covers the central United States, defined by the latitude band 30°N–50°N and longitude range 120°W–80°W. This region is chosen due to its high density of transcontinental air traffic and frequent occurrence of upper-tropospheric ice-supersaturated layers. All atmopsheric variables are extracted on the pressure levels 200, 225, 250, and 300 hPa, which span the typical cruise altitudes of commercial aviation and the primary altitude range of persistent contrail formation. These pressure levels correspond approximately to flight levels FL390 (200 hPa), FL370 (225 hPa), FL340 (250 hPa), and FL300 (300 hPa).

Figure 3.1: Geographic region corresponding to our dataset covering the spatial box 30-50°N, 120-80°W.

The atmospheric variables chosen for this study are highlighted in Table 3.1. These 10 variables, of the original 16, have the strongest observed correlation with features relevant for ISSR formation.

Variable	**Description**
t	Temperature (K)
q	Specific humidity (kg/kg)
r	Relative humidity (%)
u	Zonal wind component (m/s)
v	Meridional wind component (m/s)
w	Vertical velocity (Pa/s)
z	Geopotential (m^2/s^2)
clwc	Specific cloud liquid water content (kg/kg)
ciwc	Specific cloud ice water content (kg/kg)
cc	Cloud fraction (0–1)

Table 3.1: ERA5 pressure-level variables chosen for model training and short descriptions.

3.2 Input Representation

We define our full dataset to be the tensor $\mathcal{X}$ with dimension $(T_{\text{full}}, H, W) \times (C, P) := (17520, 81, 161) \times (10, 4)$. We differentiate the spatial and temporal element of the tensor with (T_{full}, H, W) which corresponds to the temporal size of the full dataset (in hourly timesteps), the longitudinal spread, and latitudinal spread respectively. The remaining (C, P) correponds to the number of channels per spatial point and the number of pressure levels respectively. In our case, we have 10 atmospheric variables to train on at 4 pressure levels. Our full dataset for training and validation covers the date range January 2022 to December 2023. We perform our model comparisons on a sample date and time from the month of November in 2024.

3.2.1 Label Generation

For a given sample $X \in \mathcal{X}$ with dimension $(T, H, W) \times (C, P)$ where our time history $T = 4$, number of channels $C = 10$, and number of pressure levels $P = 4$, we assign binary contrail risk labels $Y \in \{0, 1\}^{T \times H \times W \times C \times P}$. This risk label is identified via the formula for ISS to exist at that point given the temperature, specific humidity q, and pressure p. We use the formula specified in equation 2.3 which we restate here. Also note we refer to T_x as temperature at a point $x \in \mathbb{R}^{T \times H \times W}$,

$$e_{si}(T_x) = 6.112 \exp\left(\frac{22.46\,(T_x - 273.15)}{T_x - 0.55}\right)$$

$$e = \frac{q_x\, p_x}{0.622 + 0.378\, q_x} \cdot \frac{1}{100}$$

$$\text{RHi}(T_x, q_x, p_x) = \frac{e}{e_{si}(T)}$$

We reiterate here that an individual sample is made up of a pair (X, Y) where X is the tensor of atmospheric state over a history of four hours and our classification tensor Y where each spatial point is assigned a value governed by the following,

$$Y = \mathbf{I}(\text{RHi}(T_x, q_x, p_x) > 100),$$

where $\mathbf{I}(\cdot)$ denotes the indicator function with value 1 when RHi $> 100\%$ is observed and 0 otherwise.

3.2.2 Data Pre-processing

Before our raw samples, (X, Y), can be processed by the transformer it must be pre-processed. This includes data standardization and patchification. When preparing our data for model training, each variable in our tensor is normalized to zero mean and unit variance. Additionally, the region is patchified with each patch projected down to our embedding At the patchfication, we specify a patch size p which determines $N = \lceil H/p \rceil \lceil W/p \rceil$.

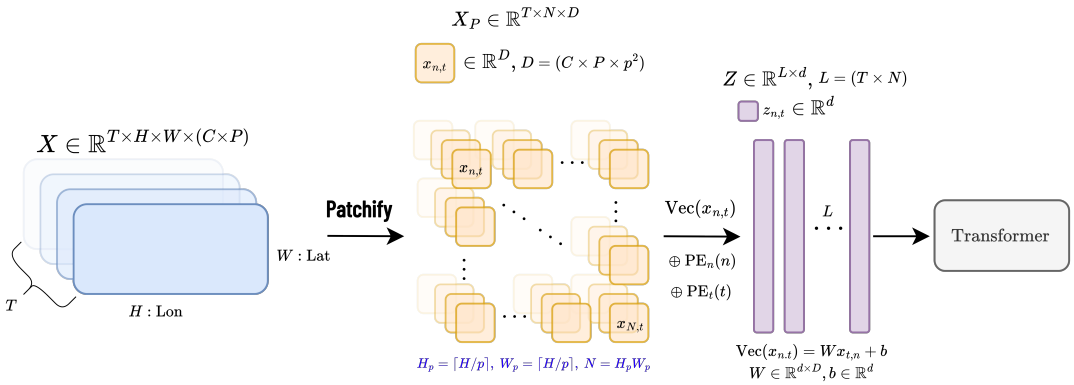

Figure 3.2: Starting from the left, we have our raw sample tensor. This gets processed via a patchification process which partitions the tensor into patches of size p. Patches at the edge of the region are padded by repeating the boundary values. These patches are then projected into a specified embedding dimension d via the Vec$(\cdot)$ function, which uses the learnable projection matrix $W \in \mathbb{R}^{d \times D}$, where $D = C \times P \times p^2$. Additionally spatial and temporal positional embedding terms are added at this stage to facilitate indexing for the attention mechanism in the transformer.

Each spatiotemporal patch $x_{t,n} \in \mathbb{R}^D$ is mapped to the model embedding dimension via a linear vectorization operator Vec$(\cdot)$, defined as

$$\text{Vec}(x_{t,n}) = W x_{t,n} + b,$$

where $W \in \mathbb{R}^{d \times D}$ is a learnable projection matrix and $b \in \mathbb{R}^d$ is a learnable bias. This operation converts each flattened patch into a token embedding in $\mathbb{R}^d$ forming the token sequence $X \in \mathbb{R}^{L \times d}$ prior to positional encoding where $L = T \times N$ and T denotes the number of temporal steps.

Let the tokenized spatiotemporal input be represented as $X \in \mathbb{R}^{L \times d}$ and $N = H_p \times W_p$ the number of spatial patches per time step obtained from an $H_p \times W_p$ patch grid.

We employ a separable spatiotemporal positional encoding composed of an independent temporal encoding and an independent spatial encoding.

The temporal encoding $\mathrm{PE}^{\mathrm{time}} \in \mathbb{R}^{T\times d}$ is defined using a 1D sinusoidal positional embedding. For time index $t \in \{0,\dots,T-1\}$ and channel index $i \in \{0,\dots,\lfloor d/2\rfloor - 1\}$,

$$\mathrm{PE}^{\mathrm{time}}_{t,\,2i} = \sin\!\left(t \cdot \exp\!\left(-\frac{2i}{d}\log(10000)\right)\right), \qquad \mathrm{PE}^{\mathrm{time}}_{t,\,2i+1} = \cos\!\left(t \cdot \exp\!\left(-\frac{2i}{d}\log(10000)\right)\right).$$

The spatial encoding $\mathrm{PE}^{\mathrm{space}} \in \mathbb{R}^{N\times d}$ is constructed using a 2D sinusoidal encoding over the patch grid by splitting the embedding dimension as $d = d_y + d_x$, with $d_y = \lfloor d/2\rfloor$ and $d_x = d - d_y$. For patch coordinates (y, x) with $y \in \{0,\dots,H_p-1\}$ and $x \in \{0,\dots,W_p-1\}$, the spatial encoding is defined as

$$\mathrm{PE}^{\mathrm{space}}_{y,x} = \left[\mathrm{PE}^{(y)}_{y} \;\middle|\; \mathrm{PE}^{(x)}_{x}\right] \in \mathbb{R}^{d},$$

where, for $i \in \{0,\dots,\lfloor d_y/2\rfloor - 1\}$ and $j \in \{0,\dots,\lfloor d_x/2\rfloor - 1\}$,

$$\mathrm{PE}^{(y)}_{y,\,2i} = \sin\!\left(y \cdot \exp\!\left(-\frac{2i}{d_y}\log(10000)\right)\right), \quad \mathrm{PE}^{(y)}_{y,\,2i+1} = \cos\!\left(y \cdot \exp\!\left(-\frac{2i}{d_y}\log(10000)\right)\right),$$

$$\mathrm{PE}^{(x)}_{x,\,2j} = \sin\!\left(x \cdot \exp\!\left(-\frac{2j}{d_x}\log(10000)\right)\right), \quad \mathrm{PE}^{(x)}_{x,\,2j+1} = \cos\!\left(x \cdot \exp\!\left(-\frac{2j}{d_x}\log(10000)\right)\right).$$

After flattening the spatial grid, this yields $\mathrm{PE}^{\mathrm{space}} \in \mathbb{R}^{N\times d}$.

Each token at time index t and spatial patch index n is then assigned the combined positional encoding

$$\mathrm{PE}_{t,n} = \mathrm{PE}^{\mathrm{time}}_{t} + \mathrm{PE}^{\mathrm{space}}_{n},$$

which is added to the corresponding token embedding prior to attention, yielding the final transformer input $Z = X_{t,n} + \mathrm{PE}_{t,n} \in \mathbb{R}^{L\times d}$. The positional encoding describe here is consistent with the 1D-positional encoding presented in the original Attention is All You Need paper [53] and its natural extension to 2D presented in the vision transformer setting [12].

3.3 Transformer Model Architecture

Our model, and many other leading data-driven models [6, 30, 24, 29], use the vision transformer (ViT) encoder model [12] as the backbone for learning the spatio-temporal features

of the atmosphere. The task of vision is similar to the task of atmospheric modelling, the longitude and latitude grid mirrors a screen's resolution and the RGB channels parralells our atmospheric variables under consideration. Our proposed model architecture for **TinyTrail** takes inspiration from two other models: Pangu-Weather [5] and FourCastNet [30]. Namely, we borrow the ideas of an earth specific positional bias term in the attention mechnasim and the usage of learning in the frequency domain respectively from each paper.

We perform ablations on the architectural decisions, using the vanilla vision transformer as our baseline. In total we compare the performance of the baseline, baseline with Fourier neural operator head layer, baseline with earth spatial positioning bias, and baseline with both.

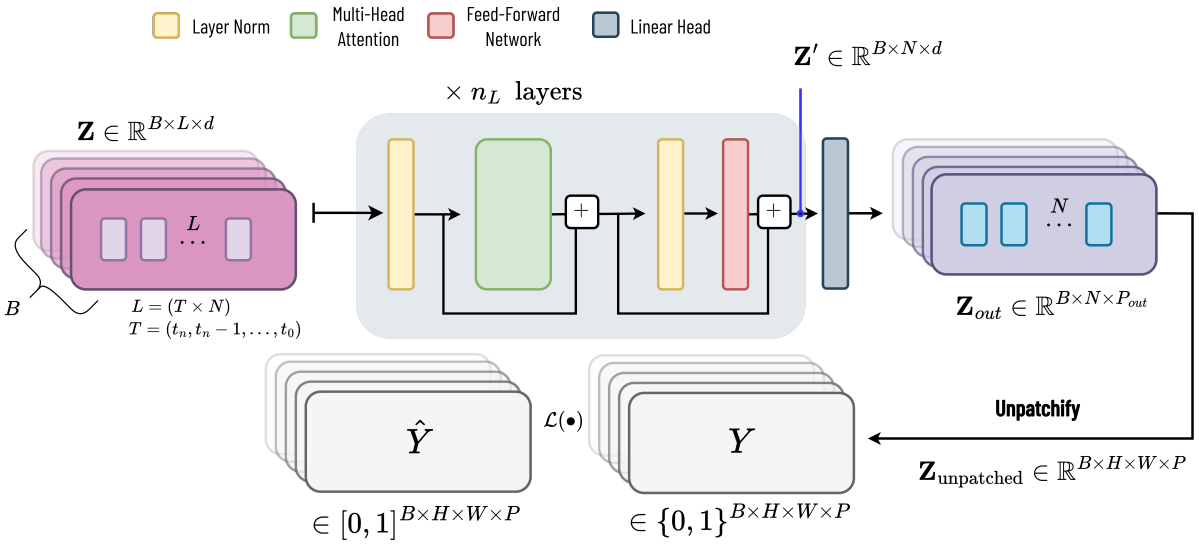

Figure 3.3: Baseline encoder-only Transformer architecture for next-hour contrail-risk prediction. Input atmospheric fields over T historical time steps are patchified into N spatial tokens per step, forming a sequence of length $L = TN$. Each patch token is linearly projected to the model dimension and enriched with separable temporal and spatial sinusoidal positional encodings before being processed by a stack of Transformer encoder layers. From the encoded sequence, only the tokens corresponding to the most recent time step are retained and passed through a linear projection head to produce per-patch logits. These logits are subsequently unpatchified to reconstruct the predicted contrail-risk field $\widehat{Y}$ on the physical grid, which is compared against the ground-truth field Y.

In Fig 3.3 we visualize the baseline architecture. The details of this encoder structure are well covered in original transformer and ViT works [53, 12].

To recover the full-resolution prediction grid, we view the coarse patchwise logits $\widehat{Y}_{\text{patch}} \in \mathbb{R}^{H_p \times W_p}$ as samples of a continuous field and define the high-resolution reconstruction $\widehat{Y} \in \mathbb{R}^{H_{\text{out}} \times W_{\text{out}}}$ by evaluating this field under a bilinear sampling operator. Let (x, y) denote output-grid coordinates and let $(u(x), v(y))$ be the corresponding continuous coordinates in the coarse grid. The reconstruction can be written compactly as

$$\widehat{Y}(x, y) = \sum_{i=0}^{H_p-1} \sum_{j=0}^{W_p-1} K_x\big(u(x) - i\big)\, K_y\big(v(y) - j\big)\, \widehat{Y}_{\text{patch}}(i, j),$$

where K_x and K_y are the 1D bilinear sampling kernels,

$$K_x(t) = \max(1 - |t|,\, 0), \qquad K_y(t) = \max(1 - |t|,\, 0),$$

and the 2D kernel is given by their separable product $K(x, y) = K_x(x)\, K_y(y)$. This operator corresponds exactly to continuous bilinear interpolation, in which each output pixel (x, y) is obtained as a weighted combination of the four nearest coarse-grid samples. In practice, this operation is implemented in Python using `torch.nn.functional.interpolate` with `mode="bilinear"` and `align_corners=False`, which applies the same separable bilinear sampling kernel over the coarse prediction grid.

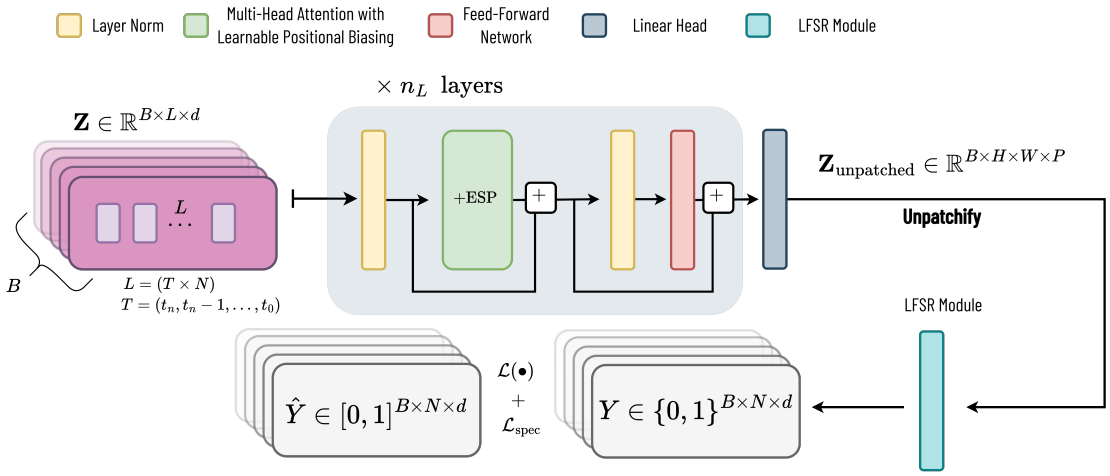

Figure 3.4: Overview of the proposed Transformer architecture, highlighting its differences from the baseline model. As in the baseline, atmospheric fields over T historical time steps are patchified into N spatial tokens per step, linearly projected to the model dimension, and enriched with separable temporal and spatial sinusoidal positional encodings before entering a stack of Transformer encoder layers. In contrast to the baseline, each self-attention block here incorporates an Earth-Specific Positional (ESP) bias, which injects geophysical structure directly into the attention logits and allows the model to better capture latitude–longitude dependencies relevant for upper-tropospheric contrail formation. After encoding, only the final-time tokens are retained and passed through a linear projection head, as in the baseline; however, the resulting per-patch logits are subsequently refined by a low-frequency spectral redisual (LFSR) module head and spectral regularized loss term $\mathcal{L}_{\text{spec}}$ that penalizes spurious high-frequency artifacts. Together, the ESP-enhanced attention and the LFSR modules extend the baseline architecture with physically informed inductive biases tailored for contrail-risk prediction.

3.3.1 Training Loss

To train the model for binary contrail–risk prediction, we employ the binary cross–entropy (BCE) loss applied independently at every spatial location and pressure level. Let

$$\widehat{Y} \in \mathbb{R}^{B\times P\times H\times W}$$

denote the model logits and

$$Y \in \{0,1\}^{B\times P\times H\times W}$$

the corresponding ground–truth labels. The sigmoid function

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

maps logits to probabilities. The per-element BCE loss is defined by

$$\ell(y, \hat{y}) = -\Big[y \log \sigma(\hat{y}) + (1 - y) \log\big(1 - \sigma(\hat{y})\big) \Big], \qquad y \in \{0, 1\},\ \hat{y} \in \mathbb{R}.$$

Aggregating over batch index b, pressure level p, and spatial coordinates (i, j) yields the full training objective

$$\mathcal{L}(\theta) = -\frac{1}{BPHW} \sum_{b=1}^{B} \sum_{p=1}^{P} \sum_{i=1}^{H} \sum_{j=1}^{W} \Big[Y_{b,p,i,j}\ \log \sigma(\widehat{Y}_{b,p,i,j}) + (1 - Y_{b,p,i,j})\ \log\big(1 - \sigma(\widehat{Y}_{b,p,i,j})\big) \Big].$$

This loss penalizes low predicted probability at locations where $Y_{b,p,i,j} = 1$ (regions exhibiting contrail–favorable conditions) and high predicted probability where $Y_{b,p,i,j} = 0$. The normalization by $BPHW$ ensures that $\mathcal{L}(\theta)$ represents the mean cross–entropy across all pixels and channels in the batch. This loss is utilized in both the baseline, ablations, and full model architecture.

3.3.2 ESP Attention

Let $Z \in \mathbb{R}^{L \times d}$ denote the sequence of token embeddings, and define

$$Q = ZW_Q, \qquad K = ZW_K,$$

with projection matrices $W_Q, W_K \in \mathbb{R}^{d \times d_k}$. Scaled dot-product attention augmented with an Earth-Specific Positional (ESP) bias is given by

$$A = \mathrm{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + B_{\mathrm{esp}} \right), \qquad A \in \mathbb{R}^{L \times L},$$

where Softmax is applied row-wise. Elementwise,

$$A_{i,j} = \frac{\exp\Big(\frac{Q_i K_j^\top}{\sqrt{d_k}} + B_{\mathrm{esp}}(i, j) \Big)}{\sum_{\ell=1}^{L} \exp\Big(\frac{Q_i K_\ell^\top}{\sqrt{d_k}} + B_{\mathrm{esp}}(i, \ell) \Big)},$$

where $Q_i \in \mathbb{R}^{d_k}$ and $K_j \in \mathbb{R}^{d_k}$ denote the ith and jth rows of Q and K, respectively.

ESP Bias Construction

Fix a spatial patch grid of size $H_p \times W_p$ and let $N := H_p W_p$ denote the number of spatial tokens per time slice. Let $T \in \mathbb{N}$ denote the number of time slices, so that the total sequence length is $L := TN$. For each spatial patch index $i \in \{1, \dots, N\}$, define normalized coordinates

$$p_i := [x_i, y_i]^\top \in [0, 1) \times [-1, 1],$$

where x_i is longitude-like (periodic) and y_i is latitude-like.

Learned Fourier embedding

Let $M \in \mathbb{N}$ denote the number of learned Fourier bands and let

$$B \in \mathbb{R}^{M \times 2}, \qquad \beta \in \mathbb{R}^M$$

be learned parameters. Define the learned Fourier positional embedding $\phi_i^{(\mathrm{LFE})} \in \mathbb{R}^{2M}$ by

$$\phi_i^{(\mathrm{LFE})} = \begin{bmatrix} \sin(2\pi(Bp_i + \beta)) \\ \cos(2\pi(Bp_i + \beta)) \end{bmatrix},$$

where $Bp_i + \beta \in \mathbb{R}^M$ and the sine/cosine are applied elementwise.

Low-rank bilinear bias

Let $r \in \mathbb{N}$ denote the ESP rank and let

$$W_q^{\mathrm{pos}}, W_k^{\mathrm{pos}} \in \mathbb{R}^{2M \times r}$$

be learned projection matrices. Define the projected positional features

$$u_i := (\phi_i^{(\mathrm{LFE})})^\top W_q^{\mathrm{pos}} \in \mathbb{R}^r, \qquad v_i := (\phi_i^{(\mathrm{LFE})})^\top W_k^{\mathrm{pos}} \in \mathbb{R}^r.$$

Let $\alpha \in \mathbb{R}$ be a learned global scale. The spatial ESP bias for a single time slice is the bilinear form

$$B_{\mathrm{sp}}(i, j) := \alpha\, u_i^\top v_j = \alpha \left((\phi_i^{(\mathrm{LFE})})^\top W_q^{\mathrm{pos}} \right) \left((\phi_j^{(\mathrm{LFE})})^\top W_k^{\mathrm{pos}} \right)^\top, \qquad B_{\mathrm{sp}} \in \mathbb{R}^{N \times N},$$

which satisfies $\mathrm{rank}(B_{\mathrm{sp}}) \le r$.

Block-diagonal extension in time

Define the full ESP bias over the length-L token sequence by repeating the same spatial bias independently within each time slice and assigning zero bias across distinct time slices:

$$B_{\text{esp}} := \text{blkdiag}(B_{\text{sp}}, \ldots, B_{\text{sp}}) \in \mathbb{R}^{L \times L},$$

i.e., for $t, s \in \{1, \ldots, T\}$ and $i, j \in \{1, \ldots, N\}$,

$$B_{\text{esp}}\big((t,i),(s,j)\big) = \begin{cases} B_{\text{sp}}(i,j), & t = s, \\ 0, & t \neq s. \end{cases}$$

Resulting Attention Kernel

With the ESP bias defined above, the attention weights take the form

$$A_{a,b} = \frac{\exp\Big(\frac{Q_a K_b^\top}{\sqrt{d_k}} + B_{\text{esp}}(a,b)\Big)}{\sum_{\ell=1}^{L} \exp\Big(\frac{Q_a K_\ell^\top}{\sqrt{d_k}} + B_{\text{esp}}(a,\ell)\Big)}, \qquad a, b \in \{1, \ldots, L\}.$$

This construction yields an attention mechanism in which the additive term B_{esp} acts as a learned, low-rank, low-frequency positional prior over the spatial grid (with longitude periodicity encoded through $x_i \in [0, 1)$), while leaving cross-time coupling to be learned through the content-dependent term $QK^\top$.

3.3.3 Low-Frequency Spectral Residual (LFSR) Module with Spectral Regularization

Given the patchwise predictions produced by the Transformer, let

$$\mathbf{Z} \in \mathbb{R}^{B \times N \times P}$$

denote the output tensor corresponding to the final-time tokens, where B is the batch size, N is the number of spatial patches, and P is the number of output channels (e.g., pressure levels). These patchwise predictions are unpatchified and rearranged to form a spatial grid

$$h_{\text{map}} \in \mathbb{R}^{B \times H \times W \times P},$$

where $H \times W$ denotes the spatial resolution of the reconstructed field.

The LFSR module is designed as a residual post-processing step: it does not replace the Transformer's prediction, but instead provides a mechanism for learning global, large-scale corrections when beneficial.

Forward Fourier Transform

Each channel of h_{map} is transformed into the frequency domain using a two-dimensional real-valued Fourier transform. For each batch element b and channel p, we compute

$$\mathcal{H}_f(b, u, v, p) = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} h_{\text{map}}(b, h, w, p) \, \exp\!\left(-2\pi i \left(\tfrac{uh}{H} + \tfrac{vw}{W}\right)\right).$$

Because h_{map} is real-valued, its Fourier coefficients satisfy the Hermitian symmetry property. Namely, the coefficients at negative frequencies are the complex conjugates of the corresponding positive-frequency coefficients. Consequently, the discrete spectrum contains redundant information and need only be stored on a nonredundant half-plane.

In our implementation, we retain all vertical frequency indices $u \in \{0, \ldots, H-1\}$ and only the nonnegative horizontal indices $v \in \{0, \ldots, \lfloor W/2 \rfloor\}$, yielding the compact representation

$$\mathcal{H}_f \in \mathbb{C}^{B \times H \times (W/2+1) \times P}.$$

The restriction $v \leq \lfloor W/2 \rfloor$ reflects the fact that indices $v > W/2$ correspond to wrapped negative frequencies whose coefficients are fully determined by Hermitian symmetry. Thus, although only $W/2+1$ horizontal frequencies are explicitly stored, the full $H \times W$ spectrum is implicitly represented and h_{map} can be reconstructed exactly via the inverse real FFT meaning no spectral information is lost.

Low-Frequency Spectral Residual Update

Fix integers $K_x \in \{1, \ldots, H\}$ and $K_y \in \{1, \ldots, \lfloor W/2 \rfloor + 1\}$, and define the retained low-frequency index set

$$\mathcal{K}_{\text{low}} := \{(u, v) \in \mathbb{Z}^2 : 0 \leq u < K_x, \;\; 0 \leq v < K_y\}.$$

For each $(u, v) \in \mathcal{K}_{\text{low}}$ and batch index $b \in \{1, \ldots, B\}$, we form a real-valued feature vector by concatenating real and imaginary parts across channels:

$$\mathbf{r}_b(u, v) := \big(\mathbf{Re}\ (\mathcal{H}_f(b, u, v, 1{:}P))\,,\ \mathbf{Im}\ (\mathcal{H}_f(b, u, v, 1{:}P))\big) \in \mathbb{R}^{2P}.$$

All coefficients with $(u, v) \notin \mathcal{K}_{\text{low}}$ are left unchanged. Next, a two-layer multi-layer perceptron (MLP) $g_\theta : \mathbb{R}^{2P} \to \mathbb{R}^P$ and a learned scalar $\alpha \in \mathbb{R}$ produce a channel-wise correction shared across frequency pairs (u, v) and batches b,

$$\Delta_b(u, v) := \alpha\, g_\theta\big(\mathbf{r}_b(u, v)\big) \in \mathbb{R}^P.$$

We implement g_θ with a hidden width $d_{\text{mid}} = \max(8, \lfloor r_{\text{hidden}} P \rfloor)$, where $r_{\text{hidden}} \in (0, 1]$ controls the capacity of the spectral correction. Concretely, the MLP g_θ is a two-layer network of the form

$$g_\theta(\mathbf{r}) = W_2\, \sigma(W_1 \mathbf{r} + b_1) + b_2, \qquad g_\theta : \mathbb{R}^{2P} \to \mathbb{R}^P,$$

where σ denotes the GELU activation applied coordinate-wise. For a vector $\mathbf{x} \in \mathbb{R}^{d_{\text{mid}}}$, it is defined componentwise by

$$\sigma(\mathbf{x})_i = \mathbf{x}_i\, \Phi(\mathbf{x}_i), \qquad i = 1, \ldots, d_{\text{mid}},$$

where $\Phi : \mathbb{R} \to [0, 1]$ is the standard normal CDF applied coordinate-wise. We also have the learnable parameters $W_1 \in \mathbb{R}^{d_{\text{mid}} \times 2P}$, $b_1 \in \mathbb{R}^{d_{\text{mid}}}$, $W_2 \in \mathbb{R}^{P \times d_{\text{mid}}}$, and $b_2 \in \mathbb{R}^P$. The hidden width is set to $d_{\text{mid}} = \max(8, \lfloor r_{\text{hidden}} P \rfloor)$. Larger d_{mid} increases the number of learnable parameters and the expressivity of the spectral correction Δ, while smaller d_{mid} provides a computationally efficient constraint on the complexity of the learned update.

The modified spectrum $\mathcal{H}'_f \in \mathbb{C}^{B \times H \times (W/2+1) \times P}$ is then defined by updating only the real part on $\mathcal{K}_{\text{low}}$: for $(u, v) \in \mathcal{K}_{\text{low}}$ and $p \in \{1, \ldots, P\}$,

$$\begin{aligned} \mathbf{Re}\, \mathcal{H}'_f(b, u, v, p) &= \mathbf{Re}\, \mathcal{H}_f(b, u, v, p) + \Delta_b(u, v)_p \\ \mathbf{Im}\, \mathcal{H}'_f(b, u, v, p) &= \mathbf{Im}\, \mathcal{H}_f(b, u, v, p), \end{aligned}$$

and $\mathcal{H}'_f(b, u, v, p) = \mathcal{H}_f(b, u, v, p)$ for $(u, v) \notin \mathcal{K}_{\text{low}}$.

At the spectral-update stage, the shared MLP produces an input-adaptive, channel-wise correction in the low-frequency Fourier domain. For each batch element b and retained frequency $(u, v) \in \mathcal{K}_{\text{low}}$, the module computes a correction vector $\Delta_b(u, v, :) \in \mathbb{R}^P$. Although

the correction is emitted per channel, it is cross-channel coupled, meaning each entry $\Delta_b(u, v, p)$ may depend on all channels because the MLP input concatenates the real and imaginary components across the P channels. The correction is then applied as a residual update to the low-frequency Fourier coefficients (in our implementation, to the real part only), while all coefficients outside $\mathcal{K}_{\text{low}}$ are left unchanged.

Inverse Fourier Transform

Let $\mathcal{H}'_f \in \mathbb{C}^{B\times H\times(W/2+1)\times P}$ denote the modified Fourier spectrum after the low-frequency residual update. We map this spectrum back to physical space by applying the two-dimensional inverse real Fourier transform channel-wise. For each batch element $b \in \{1, \dots, B\}$, spatial index $(h, w) \in \{0, \dots, H-1\} \times \{0, \dots, W-1\}$, and channel $p \in \{1, \dots, P\}$, we define

$$\delta(b, h, w, p) := \sum_{u=0}^{H-1} \sum_{v=0}^{\lfloor W/2 \rfloor} \mathcal{H}'_f(b, u, v, p)\, \exp\!\big(2\pi i \left(\tfrac{uh}{H} + \tfrac{vw}{W}\right)\big) \;\in\; \mathbb{R}.$$

Equivalently, in operator form,

$$\delta \;:=\; \text{irFFT2}(\mathcal{H}'_f) \;\in\; \mathbb{R}^{B\times H\times W\times P},$$

where $\text{irFFT2} : \mathbb{C}^{B\times H\times(W/2+1)\times P} \to \mathbb{R}^{B\times H\times W\times P}$ denotes the inverse operator associated with the forward real FFT on an $H \times W$ grid. By construction, δ is real-valued and has the same spatial resolution and channel dimension as the original prediction h_{map}.

Residual Blending

Let $\delta \in \mathbb{R}^{B\times H\times W\times P}$ denote the spatial correction field obtained after the inverse real Fourier transform. We apply a learnable pointwise channel-mixing operator to δ before adding it residually to the original logits. Concretely, define

$$\text{Blend} : \mathbb{R}^{B\times H\times W\times P} \to \mathbb{R}^{B\times H\times W\times P}$$

as a 1×1 convolution across channels. For each batch index $b \in \{1, \dots, B\}$, spatial location $(h, w) \in \{0, \dots, H-1\} \times \{0, \dots, W-1\}$, and output channel $p \in \{1, \dots, P\}$, the operator is given by

$$\big[\text{Blend}(\delta)\big](b, h, w, p) = \sum_{q=1}^{P} A_{pq}\, \delta(b, h, w, q) + c_p \in \mathbb{R},$$

where $A \in \mathbb{R}^{P\times P}$ and $c \in \mathbb{R}^P$ are learnable parameters shared across all spatial locations, applied pointwise in (h, w). This operation mixes information across channels while preserving the spatial resolution.

The refined logits are then obtained via a residual update

$$h_{\text{map}} \leftarrow h_{\text{map}} + \text{Blend}(\delta), \qquad h_{\text{map}} \in \mathbb{R}^{B\times H\times W\times P}.$$

In our implementation, the blending parameters (A, c) are initialized to zero, so $\text{Blend}(\delta) \approx 0$ at initialization and the refinement head acts as a near-identity mapping; the contribution of the residual correction is therefore learned only when supported by the training objective.

Spectral Regularization

In addition to the LFSR head, we impose a spectral penalty on the predicted logits during training to discourage excess high-frequency energy. We define the high-frequency set using normalized spatial frequencies in Nyquist units. For a discrete field on an $H \times W$ grid, the Nyquist wavenumber corresponds to the highest resolvable oscillation (one cycle every two grid points), hence normalized frequencies lie in $[0, 0.5]$ (nonnegative half-spectrum) or $[-0.5, 0.5]$ (signed spectrum).

Concretely, for the real FFT index $v \in \{0, \dots, \lfloor W/2\rfloor\}$ we define the normalized horizontal frequency by

$$\kappa_x(v) := \frac{v}{W} \in [0, 0.5],$$

and for the vertical FFT index $u \in \{0, \dots, H-1\}$ we define the signed, normalized vertical frequency by

$$\kappa_y(u) := \begin{cases} \frac{u}{H}, & 0 \le u \le \lfloor H/2\rfloor, \\ \frac{u-H}{H}, & \lfloor H/2\rfloor < u \le H-1, \end{cases} \qquad \in [-0.5, 0.5].$$

Thus $\kappa_y(u)$ is positive for $u \le H/2$ and represents wrapped negative frequencies for $u > H/2$. Given cutoffs $k_x^{\text{cut}}, k_y^{\text{cut}} \in [0, 0.5]$, we define

$$\mathcal{K}_{\text{high}} := \Big\{(u, v) \in \{0, \dots, H-1\} \times \{0, \dots, \lfloor W/2\rfloor\} : \ \kappa_x(v) > k_x^{\text{cut}} \ \text{ or } \ |\kappa_y(u)| > k_y^{\text{cut}}\Big\}.$$

The spectral regularization loss is then

$$\mathcal{L}_{\text{spec}}(h_{\text{map}}) := \lambda \, \frac{1}{B\,P\,H\,(\lfloor W/2\rfloor + 1)} \sum_{b=1}^{B}\sum_{p=1}^{P}\sum_{u=0}^{H-1}\sum_{v=0}^{\lfloor W/2\rfloor} \mathbf{1}[(u, v) \in \mathcal{K}_{\text{high}}] \, \big|\mathcal{H}_f(b, u, v, p)\big|^2, \ \ \lambda \in \mathbb{R}_{\ge 0}$$

where $\mathbf{1}[\cdot]$ is the indicator function. This term is added to the training objective (weighted by λ) and does not alter the inference-time architecture. This spectral regularization term penalizes excessive energy in high-frequency Fourier modes of the predicted logits. By assigning a cost to the power spectrum outside a prescribed low-frequency band, this penalty discourages spurious fine-scale oscillations while preserving the model's ability to represent large-scale structure.

Motivation

The LFSR module and spectral regularization serve complementary roles. The LFSR module provides the model with an explicit mechanism for learning global, large-scale corrections in Fourier space, compensating for the inherently local and patchwise inductive biases of Transformer-based architectures. Importantly, it does not impose any preference toward smooth or low-frequency predictions; it simply enables such corrections when they improve task performance.

Spectral regularization, by contrast, introduces an explicit inductive bias that discourages unconstrained high-frequency energy in the predicted fields. For coarse-resolution geophysical prediction tasks, fine-scale spectral components are often poorly supervised, under-resolved, or dominated by noise, making such penalties a principled form of output-space regularization.

Both mechanisms are motivated by prior work on Fourier-domain operator learning and spectral token mixing. Fourier Neural Operators (FNO) parameterize global convolutional kernels in Fourier space and have shown to provide an efficient way to model long-range spatial interactions in PDE solution operators [27]. Adaptive Fourier Neural Operators (AFNO) extend this idea as an efficient Fourier-domain token mixer for high-resolution fields [19], and have been deployed at scale for global data-driven weather forecasting in FourCastNet [30]. Related global spectral filtering architectures, such as Global Filter Networks, also demonstrate that Fourier-domain filtering can capture long-range spatial dependencies efficiently via learned frequency-domain transformations [32]. Our approach outlined here is inspired by these works.

Chapter 4

Experimental Setup and Results

4.1 Model Specification and Ablations

Baseline Model Configuration

For the baseline Transformer architecture, we adopt a lightweight encoder–only design tailored to short–horizon contrail–risk prediction. The model is trained using a sliding temporal window of length $T_{\text{in}} = 4$ hours, from which the network predicts the contrail–risk field at a lead time of $\Delta t = 1$ hour. At each training step, the input therefore consists of four consecutive hourly atmospheric states, and the target corresponds to the binary contrail–risk label one hour into the future.

Spatial inputs are patchified with a patch size of $p = 8$, and each patch is embedded into a latent representation of dimension $d = 256$. The Transformer encoder comprises 4 layers with 4 attention heads per layer, using standard pre–norm residual blocks and GELU activations. Training is performed with a batch size of 8 for 40 epochs using the Adam optimizer with learning rate 10^{-4} and weight decay 5×10^{-3}. These hyperparameters were selected as a standard starting point for the model; hyperparameter optimization search may lead to better performance.

Baseline + ESP Ablation

To assess the contribution of Earth–Specific Positional (ESP) bias, we augment the baseline Transformer with the learnable additive attention bias $B_{\text{esp}}(i, j)$. For the ablation experiments, the ESP module is configured with the following hyperparameters: `esp_rank = 8`, `esp_kx = 2`, and `esp_ky = 2`. Each parameter controls a distinct component of the ESP mechanism:

- **Rank $r = 8$.** The ESP bias is parameterized as a low–rank factorization $B_{\text{esp}}(i, j) = (\phi_i^{(\text{LFE})} W_q^{\text{pos}}) \cdot (\phi_j^{(\text{LFE})} W_k^{\text{pos}})^\top$, where r determines the dimensionality of the learned projection matrices $W_q^{\text{pos}}, W_k^{\text{pos}} \in \mathbb{R}^{2M \times r}$. Increasing the rank increases the expressiveness of the spatial bias field, while keeping $r = 8$ yields a compact, computationally efficient prior.

- The quantity M denotes the number of learned Fourier modes used in the Least–Frequency Encoding (LFE) of the true latitude–longitude coordinates. It is determined by the product of the chosen kernel sizes, $M = k_x \times k_y$, so that larger (k_x, k_y) allow the positional embedding to represent finer spatial variation while remaining dominated by smooth, low–frequency structure.

- **Fourier kernel size in latitude ($k_x = 2$).** ESP uses a set of learned Fourier features $\phi^{(\text{LFE})}(x, y)$ to encode geospatial coordinates. The parameter k_x specifies the number of low–frequency Fourier bands used along the meridional (north–south) direction, controlling how much large–scale latitudinal structure the model can represent.

- **Fourier kernel size in longitude ($k_y = 2$).** Analogously, k_y determines the number of azimuthal (east–west) low–frequency components. Together, (k_x, k_y) set the spatial resolution of the learnable geophysical prior, ensuring that ESP biases emphasize planetary–scale, slowly varying patterns rather than high–frequency noise.

This configuration introduces a mild but meaningful inductive bias that reflects the natural structure of atmospheric fields, while keeping the additional parameter cost negligible. The ablation isolates the effect of this positional prior when added on top of the baseline Transformer without the LFSR head component.

The ESP mechanism adds only a small number of learnable parameters relative to the baseline Transformer. The LFE embedding uses $M = k_x k_y = 4$ learned Fourier modes, producing a $2M$–dimensional positional feature vector for each token. The projection

matrices $W_q^{\text{pos}}, W_k^{\text{pos}} \in \mathbb{R}^{2M \times r}$ therefore contain $2Mr = 64$ parameters each, and the learned frequency matrix $B \in \mathbb{R}^{M \times 2}$ together with the phase vector $\beta \in \mathbb{R}^M$ contributes an additional $3M = 12$ parameters. In total, the ESP module introduces fewer than 150 parameters, several orders of magnitude smaller than the baseline model's Transformer layers.

From a computational standpoint, the cost of forming the bias matrix $B_{\text{esp}} \in \mathbb{R}^{L \times L}$ is dominated by the low–rank bilinear form $(\phi_i W_q^{\text{pos}})(\phi_j W_k^{\text{pos}})^\top$, which scales as $\mathcal{O}(Lr)$ rather than $\mathcal{O}(L^2)$ due to the rank–r factorization. Since $L = TN$ is relatively small in our setting (with $T_{\text{in}} = 4$ and moderate patch count), the additional computation is negligible compared to the self–attention layers themselves. The ESP augmentation yields a geophysical inductive bias at effectively no additional parameter or runtime cost.

Baseline + LFSR Head Ablation

In this ablation, the baseline Transformer is extended with an LFSR (Low-Frequency Spectral Residual) head that refines the predicted contrail–risk field in the frequency domain. After unpatchifying the model output into $h_{\text{map}} \in \mathbb{R}^{B \times H \times W \times P}$, the LFSR head computes its 2D Fourier transform $\mathcal{H}_f = \mathcal{F} h_{\text{map}}$ and applies residual correction term to selected low-frequency modes. The hyperparameters `lfsr_kx = 0.25` and `lfsr_ky = 0.25` determine which portion of the spectrum this correction is applied to. The hyperparameters `spec_kx = 0.35`, `spec_ky = 0.35`, and `spec_lambda = 0.2` determine the loss penalty for select high-frequency modes.

Full Architecture (TinyTrail)

The full architecture integrates both components examined in the ablation studies. Specifically, it augments the baseline Transformer simultaneously with the Earth–Specific Positional (ESP) bias—configured with `esp_rank = 8`, `esp_kx = 2`, `esp_ky = 2`—and the LFSR spectral refinement head using the parameters `lfsr_kx = 0.25` and `lfsr_ky = 0.25`. In combination, ESP contributes a geophysically structured, low-rank additive attention prior, while LFSR performs a low-frequency spectral refinement of the output field and high-frequency loss penalty. This joint configuration yields the complete model used in our experiments, incorporating both a physically informed attention bias and a large-scale spectral smoothing mechanism on top of the baseline architecture.

4.2 Dataset

We construct our training and evaluation dataset from two full years (2022-2023) of ERA5 hourly reanalysis on pressure levels, restricted to ten atmospheric state variables (listed in Table 1). The data are cropped to the geographic domain (50 N, 30 N, 120 W, 80 W), corresponding to a mid–latitude region over the central United States, as visualized in Figure 3.1. For each pressure level, the ERA5 fields are stored at a spatial resolution of $0.25° \times 0.25°$ and hourly temporal frequency, yielding a high–resolution, multi–year dataset for short–term contrail–risk prediction. Across all variables, pressure levels, and timesteps, the resulting dataset occupies approximately 30 GB.

4.3 Hardware

All models were trained on a single NVIDIA H100 GPU. Across the full set of experiments, including the baseline, ESP-augmented, LFSR-augmented, and full architectures, the average end-to-end training time per run was approximately 15 minutes.

4.4 Qualitative Evaluation

To qualitatively assess model behaviour, we examine predictions for a fixed example on **26 November 2024 at 16:00 UTC**. For this timestep, each model variant produces a spatial heatmap of predicted contrail–risk probabilities, which we compare directly against the ground–truth labels derived from ERA5 RHi. Visual differences in the intensity and extent of high–risk regions highlight the characteristic biases of each architecture: sharper or more fragmented patterns indicate over–sensitivity to local features, while smoother, broader structures reflect a stronger large–scale prior.

4.5 Quantitative Evaluation

A key aspect of probabilistic prediction quality is *calibration*. A model is considered well calibrated if its predicted probabilities reflect true event frequencies. Formally, for a predicted probability $\hat{p}$, a perfectly calibrated model should satisfy

$$\Pr(Y = 1 \mid \hat{y} = \hat{p}) \approx \hat{p}.$$

In other words, among all pixels for which the model assigns probability $\hat{p}$ to contrail formation, approximately a fraction $\hat{p}$ of them should indeed be labeled positive. In our quantitative evaluation we similarly compare performance on **26 November 2024 at 16:00 UTC** contrail risk prediction.

To quantify calibration quality, we use the *Expected Calibration Error* (ECE), which partitions predictions into M probability bins $B_1, \ldots, B_M$ and measures the discrepancy between accuracy and average predicted confidence within each bin. The ECE is defined as

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|,$$

where

$$\mathrm{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}\{y_i = 1\}, \qquad \mathrm{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i.$$

A lower ECE indicates better calibration, with $\mathrm{ECE} = 0$ achieved only for a perfectly calibrated classifier. In our evaluation, we compute ECE across all spatial pixels and pressure levels to provide a robust measure of probabilistic fidelity for contrail–risk prediction.

4.6 Results

4.6.1 Qualitative Results

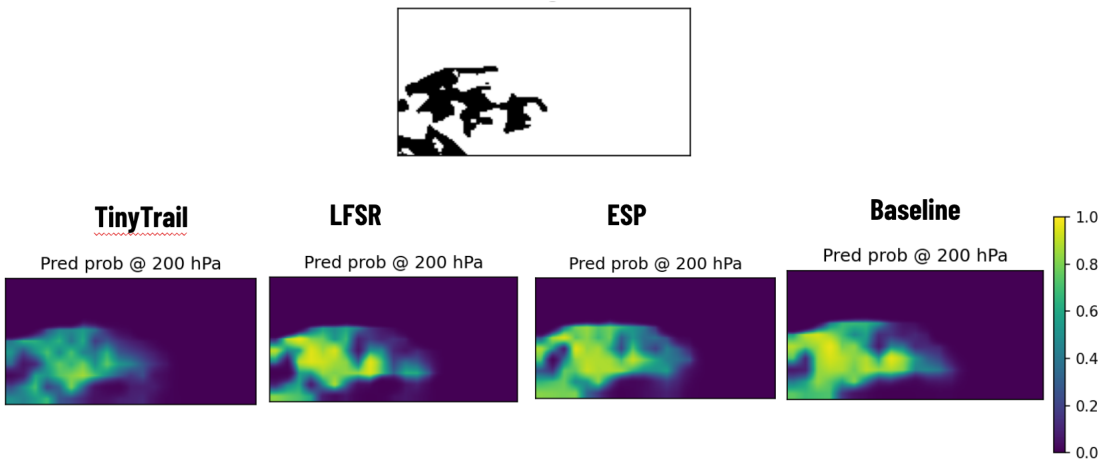

Figure 4.1: Qualitative comparison of contrail–risk predictions for **26 November 2024 at 16:00 UTC**. The top panel shows the ground-truth binary contrail mask (RHi $> 100\%$) over the evaluation region. The bottom row displays predicted probability heatmaps at 200 hPa for the full TinyTrail model, and for each ablation: LFSR-only, ESP-only, and the baseline Transformer.

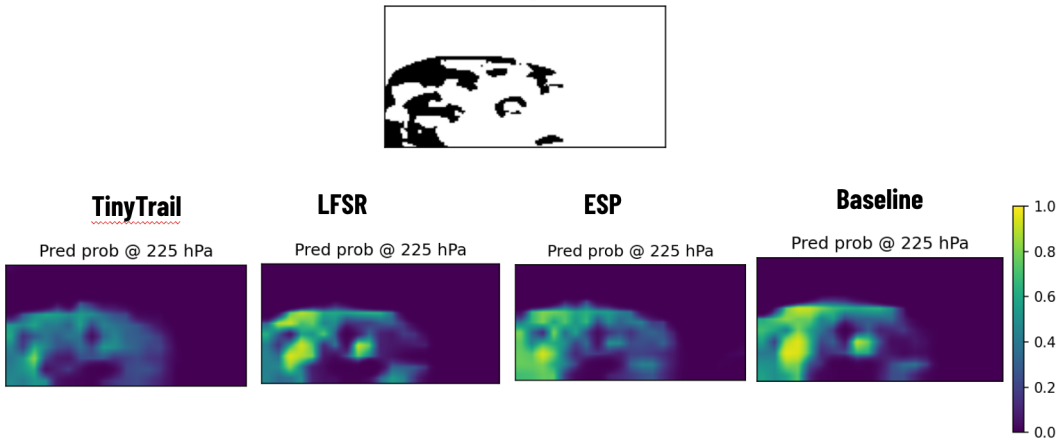

Figure 4.2: Qualitative comparison of contrail–risk predictions for **26 November 2024 at 16:00 UTC** and **225** hPa.

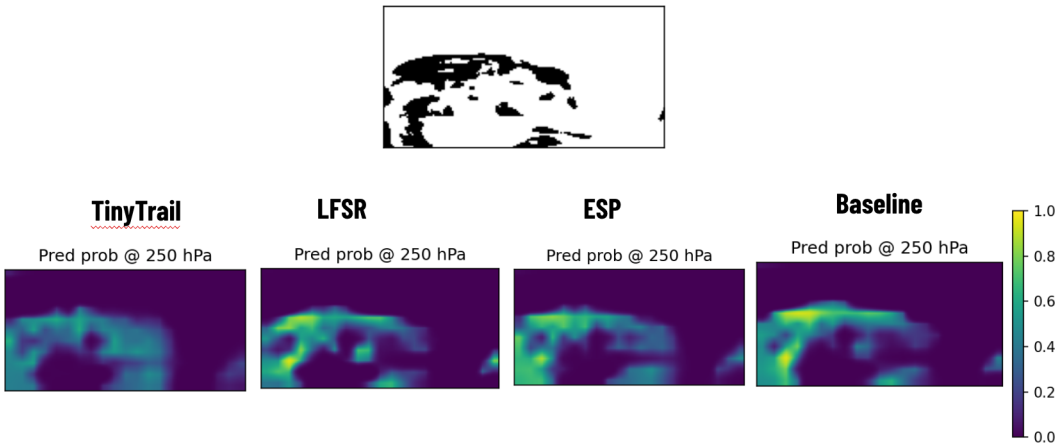

Figure 4.3: Qualitative comparison of contrail–risk predictions for **26 November 2024 at 16:00 UTC** and **250** hPa.

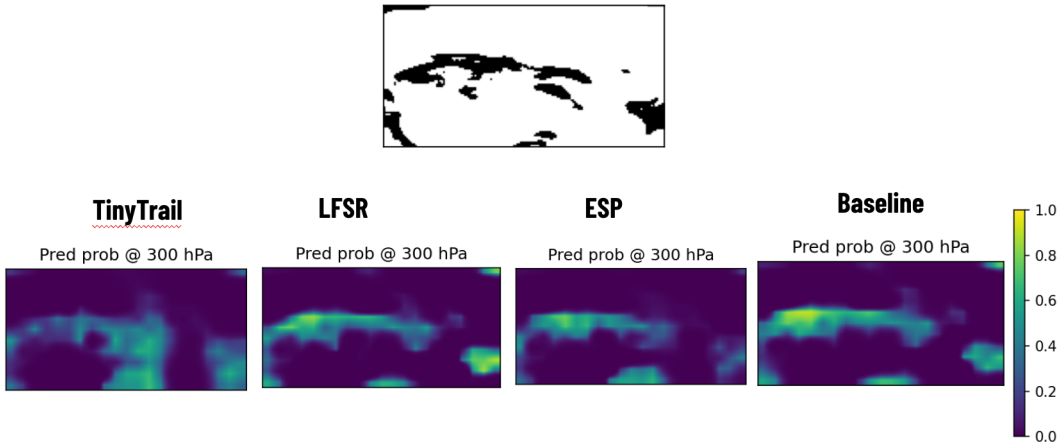

Figure 4.4: Qualitative comparison of contrail–risk predictions for **26 November 2024 at 16:00 UTC** and **300** hPa.

4.6.2 Quantitative Results

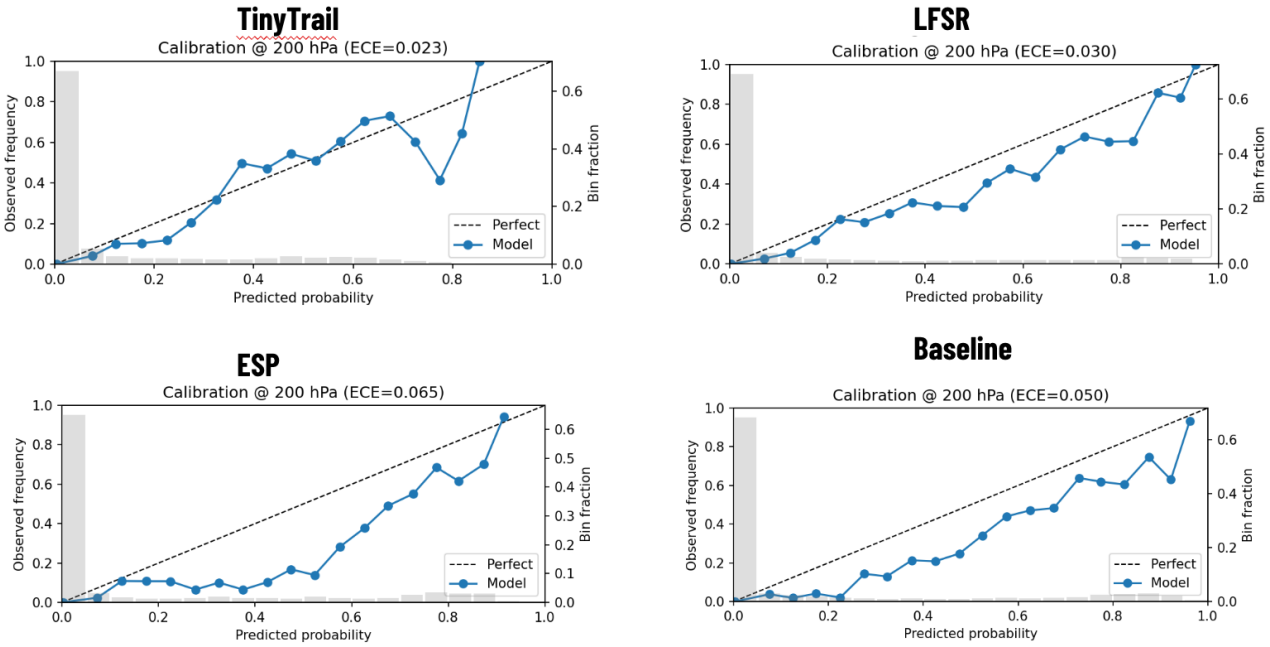

Figure 4.5: Calibration curves for contrail–risk prediction at **200 hPa** for the validation example on **26 November 2024 at 16:00 UTC**. Each panel shows, for a given model variant (TinyTrail, LFSR-only, ESP-only, Baseline), the relationship between the predicted contrail probability (horizontal axis) and the observed event frequency within each probability bin (vertical axis). The dashed diagonal represents perfect calibration: points lying on this line indicate that a predicted probability p corresponds to an empirical event frequency of p. The blue curve shows the model's actual calibration behaviour, with deviations from the diagonal reflecting over- or under-confidence. The light grey bars along the lower axis depict the distribution of predicted probabilities (bin fractions), indicating where most predictions occur. Models with lower ECE values (reported in each title) achieve closer alignment between predicted risk and true contrail occurrence.

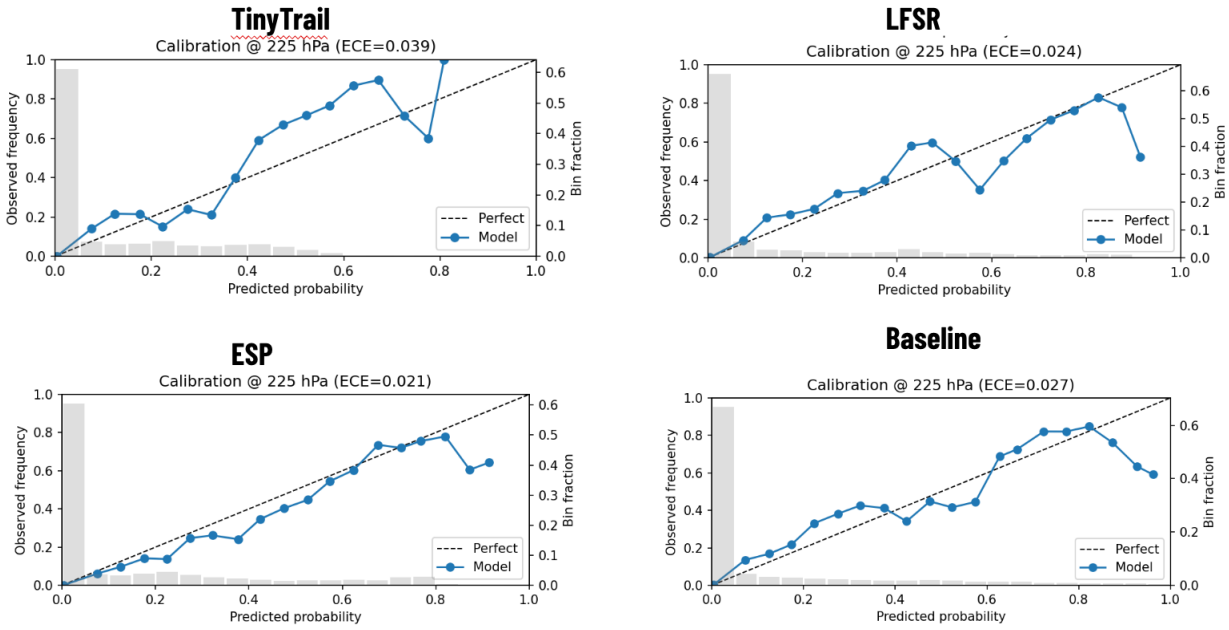

Figure 4.6: Calibration curves for contrail–risk prediction at **225 hPa** for the validation example on **26 November 2024 at 16:00 UTC**.

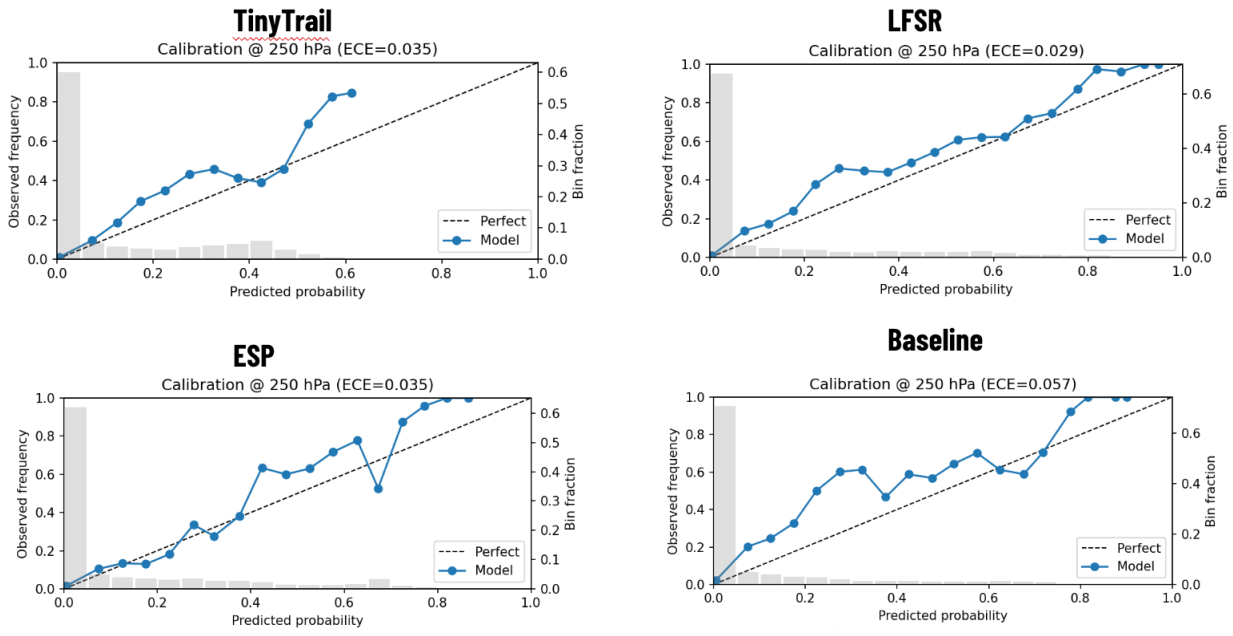

Figure 4.7: Calibration curves for contrail–risk prediction at **250 hPa** for the validation example on **26 November 2024 at 16:00 UTC**.

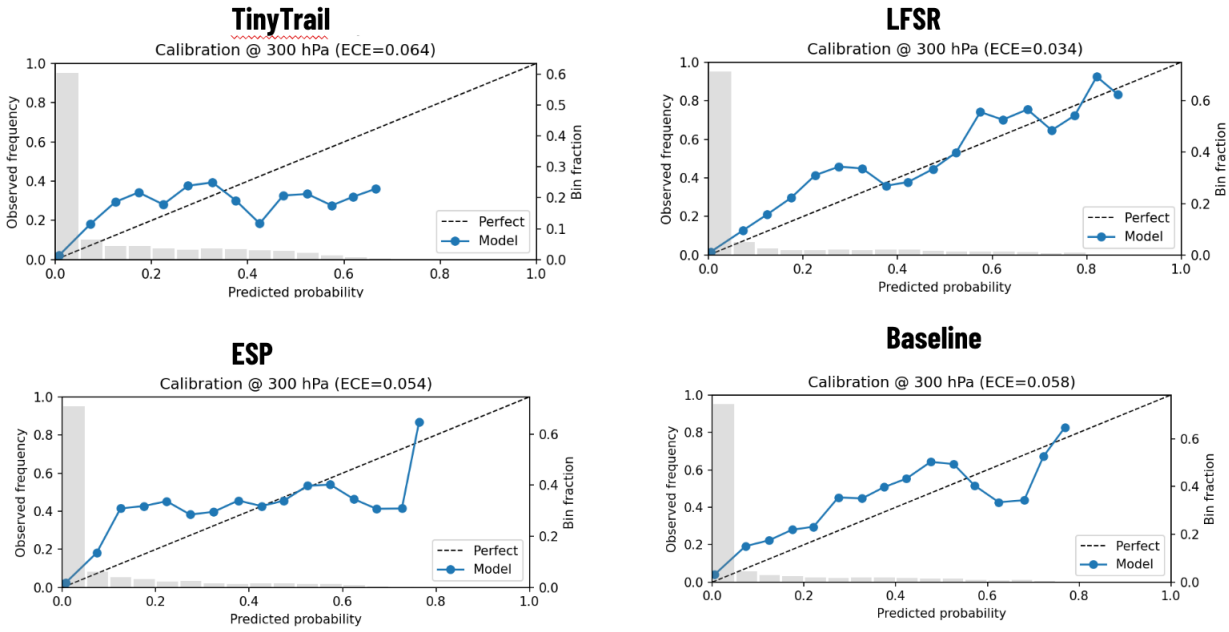

Figure 4.8: Calibration curves for contrail–risk prediction at **300 hPa** for the validation example on **26 November 2024 at 16:00 UTC**.

Chapter 5

Discussion and Conclusion

5.1 Qualitative Evaluation Across Pressure Levels

Figures 4.1–4.4 show qualitative predictions of contrail–risk probabilities for the case study on 2024–11–26 at 16:00 UTC across pressure levels 200, 225, 250, and 300 hPa. Each figure pairs the binary RHi $> 100\%$ label field (top) with a set of predicted probability maps from TinyTrail, the LFSR head ablation, the ESP ablation, and the baseline model (bottom row).

While these heat maps do not by themselves allow strong conclusions about relative model performance, they do help reveal characteristic spatial biases introduced by each architectural modification. The LFSR head tends to produce probability fields with more pronounced spatial gradients and sharper transitions, consistent with its spectral refinement mechanism. The ESP ablation, by contrast, often yields smoother, more latitudinally organized structures, reflecting the influence of its learned geophysical positional bias. The baseline model generally exhibits softer spatial variation, with broader regions of moderate confidence. TinyTrail combines both inductive biases and therefore displays a mixture of localized structure and large–scale organization.

Across pressure levels, all models show some sensitivity to altitude, with changes in the apparent coherence, smoothness, or contrast of their predicted fields. These qualitative differences highlight how architectural components shape the spatial character of predictions; however, a more precise assessment of model accuracy and calibration is provided in the subsequent quantitative analysis.

5.2 Quantitative Evaluation Across Pressure Levels

To complement the qualitative inspection of spatial prediction patterns, we evaluate model behaviour quantitatively through calibration analysis at four pressure levels (200, 225, 250, and 300 hPa). Overall, all architectures—Baseline, ESP-augmented, LFSR-refined, and the full TinyTrail model—exhibit reasonably well-calibrated probability estimates, as indicated by their low Expected Calibration Error (ECE) values across levels. This is noteworthy given the difficulty of predicting contrail-permitting conditions from high-dimensional atmospheric fields, and it underscores a key trend in the literature: even pure ViT-style architectures can perform strongly on geophysical tasks. In particular, ClimaX [29] achieves state-of-the-art results using a vanilla ViT backbone with no explicit physical priors, illustrating the surprising effectiveness of the transformer's general-purpose representation capacity.

While TinyTrail does not achieve the lowest ECE at every pressure level, its performance remains competitive relative to the baselines. Small variations in calibration across models are expected, as the probability distribution of RHi$>$ 100% is highly skewed and sensitive to subtle errors in humidity and temperature dynamics. Differences among models likely reflect several factors: (i) slightly different inductive biases (spectral filtering, geophysical positional structure), (ii) small variances in optimization dynamics given the relatively short training schedule, and (iii) limited calibration sample size for computing empirical reliability curves.

It is important to emphasize that the calibration results do not suggest any model is failing; rather, they indicate that all variants—including the raw ViT baseline—are effective at producing meaningful probabilistic estimates. TinyTrail's performance, even when not strictly the best, should not be interpreted as dismissing its architectural design. Instead, the results suggest that further improvements could plausibly be achieved via hyperparameter search, extended training, temperature scaling or post-hoc calibration, or wider evaluation samples. Given the sensitivity of ECE to binning choices and dataset size, additional calibration diagnostics (such as adaptive binning or Bayesian reliability estimation) may yield even clearer distinctions among model families.

In summary, all models demonstrate strong quantitative behaviour across pressure levels, and TinyTrail remains a viable, compact architecture whose performance could be further enhanced with more extensive experimentation. These early results invite natural follow up experiments of small architecture designs for this task.

5.3 Model Size Comparison

To contextualize the compactness of the proposed TinyTrail architecture, we compare its parameter count to several leading data-driven weather forecasting models. Despite incorporating both spectral refinement (LFSR) and geophysically informed positional structure (ESP), TinyTrail remains orders of magnitude smaller than state-of-the-art systems, underscoring its suitability for lightweight research and rapid experimentation.

Model	**Parameter Count**	**Description**
TinyTrail	$\sim$ **3.82M**	A lightweight Transformer for short-term contrail-risk prediction, optionally equipped with ESP positional bias and LFSR spectral refinement. Designed for efficiency and rapid experimentation.
FourCastNet	$\sim$ 433M	A high-capacity global forecasting model built on Adaptive Fourier Neural Operators, emphasizing spectral mixing and large-scale parallel training.
Pangu-Weather	$\sim$ 256M	A leading 3D Earth-system Transformer incorporating Earth-specific positional biases, achieving operational-scale deterministic weather forecasting performance.
ClimaX	$\sim$ 115M	A general-purpose climate foundation model using a pure ViT encoder with no explicit physical priors, demonstrating strong results across diverse Earth-system tasks.

Table 5.1: Parameter counts and descriptions of TinyTrail and leading weather–climate forecasting architectures. The compact TinyTrail model remains 30–100$\times$ smaller than large operational systems while achieving competitive contrail-risk predictive performance.

This comparison highlights an important point: although TinyTrail is approximately *30–100 times smaller* than prominent operational-scale systems, its performance on contrail-risk prediction remains competitive. The results suggest that carefully chosen inductive biases, efficient token representations, and task-specific transformer design can yield strong performance even at modest parameter scales.

5.4 Future Work

The preliminary results presented in this study motivate a number of natural extensions aimed at improving the fidelity, robustness, and practical utility of data-driven contrail–risk prediction. Future experiments should explore scaling the temporal conditioning window beyond the four-hour history used here. Longer input sequences (e.g., 8–12 hours) may enable the model to capture slower mesoscale moisture dynamics that influence the onset and persistence of ice-supersaturated regions. Observational studies indicate that ice-supersaturated regions—a necessary precursor for persistent contrail formation—can exhibit variability on time scales ranging from minutes up to several hours, with mean lifetimes on the order of hours and maxima extending to roughly 16 hours under certain atmospheric conditions [44]. Such evidence motivates extending the temporal context beyond the current four-hour window to allow learning models to better integrate temporal dependencies in moisture and thermodynamic evolution.

Similarly, hyperparameter tuning, increased model capacity, and training on a larger temporal span of ERA5 (beyond the two years considered in this work) may reveal additional gains. Additionally, the use of higher-vertical-resolution datasets such as ERA5's *137-level* [14] product can allow interpolation to a finer set of aviation-relevant flight levels and offer contrail–risk estimation with improved re-routing capabilities. This was not used in our study due to the more copmlex nature of the dataset and stronger access restrictions. Beyond model scaling, there are several promising avenues for operational evaluation. With a denser set of flight-level predictions, one can perform historical case studies to estimate how much contrail formation might have been reduced had certain altitude-adjustment interventions been applied. Such analyses would help quantify the potential climate impact of contrail avoidance strategies and assess the practical value of real-time contrail–risk maps.

Towards better calibration evaluations, we can measure calibration against true *observed* contrail occurrence. One method for testing model calibration would be to take a large sample of flight data (e.g., from the OpenSky Network [37]) and compare predicted contrail formation against observational contrail detections, such as those visualized on the Contrails.org contrail map [10]. Comparing predicted contrail probabilities with real historical contrail formation events would enable more rigorous calibration diagnostics and support the development of correction methods tailored to aviation operations. This backtesting against live flight data also lends itself to a post-training finetuning. Furthermore, in a live operational setting, using observed contrail risk in a reinforcement learning con-

text may also be worth exploring.

Another direction is the exploration of alternative proxy labels for contrail formation beyond ice supersaturation. While $\mathrm{RHi} > 100\%$ is widely used and physically justified, additional variables—such as Schmidt–Appleman threshold diagnostics, relative temperature and pressure gradients, or multi-variable contrail formation models—may capture dynamics missing from a single-threshold label. Jointly modeling these factors could yield more robust contrail–risk estimators and expand the applicability of transformer-based architectures to aviation–climate interaction studies.

Taken together, these directions highlight a rich landscape for future research: larger models, longer temporal context, higher-vertical-resolution inputs, historical intervention studies, and calibration against real flight and contrail-formation records. The preliminary effectiveness of TinyTrail and its ablations suggests that transformer-based systems offer a strong foundation for the next generation of operational contrail–mitigation tools.

5.5 Conclusion

This work presented TinyTrail, a compact Transformer architecture for short-term contrail–risk prediction built from ERA5 pressure-level data. Through a series of ablations, we explored the contributions of two lightweight inductive biases—Earth–Specific Positional (ESP) structure and LFSR-based spectral refinement—and demonstrated that each produces characteristic spatial effects in the prediction field while maintaining strong overall calibration performance. Despite its small parameter count, TinyTrail its ablated variants, aligning with broader evidence in the literature that even vanilla ViT backbones can serve as effective forecasting models in atmospheric settings.

Our qualitative and quantitative analyses show that all examined architectures capture essential patterns of contrail-permitting conditions, highlighting both the strength of transformer-based approaches and the promise of further task-specific refinement. At the same time, the variability across pressure levels and the modest differences in calibration suggest that there remains considerable room for advancement. The results motivate deeper investigations into model scaling, richer temporal context, higher-vertical-resolution atmospheric inputs, and calibration against observed contrail formation events.

Overall, these findings support the feasibility of lightweight, data-driven models for operational contrail-avoidance applications and establish a foundation upon which more

comprehensive forecasting and decision-making systems can be developed.

References

[1] H. Appleman. The formation of exhaust condensation trails by jet aircraft. *Bulletin of the American Meteorological Society*, 34:14–20, 1953.

[2] Denis Avila, Lance Sherry, and Terry Thompson. Reducing global warming by airline contrail avoidance: A case study of annual benefits for the contiguous united states. *Transportation Research Interdisciplinary Perspectives*, 2:100033, 2019.

[3] K. P. A. M. Barten. Contrail mitigation through flight planning. Master's thesis, Delft University of Technology (TU Delft), 2017.

[4] Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525:47–55, 2015.

[5] Kaifeng Bi, Rui Yan, Tian Zhang, and et al. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619:720–726, 2023.

[6] Kailai Bi, Shilong Li, Yifan Liu, Guang Song, Yu Wang, Jun Wu, and Guoqiang Yang. Pangu-weather: A 3D high-resolution model for fast and accurate global weather forecast. *Nature*, 619:317–324, 2023.

[7] CAEP Working Group 2, ICAO. Report on operational opportunities to reduce contrails and non-CO2 effects. Technical report, International Civil Aviation Organization (ICAO), 2024. Accessed: 2025-12-08.

[8] Zach Cathcart, R. Teoh, U. Schumann, R. Contreras, et al. Understanding contrail management. Google Research and Breakthrough Energy White Paper, 2024.

[9] Robert Chen, Eng-Shien Ong, and Michael Pritchard. U-net architectures for weather and climate applications. *Geoscientific Model Development*, 2020.

[10] Contrails.org. Contrails map and observational contrail data. https://contrails.org/, 2025. Accessed: 2025-12-XX; contains live and historic contrail formation maps derived from satellite and other observations.

[11] Copernicus Climate Data Store. ERA5 reanalysis: Hourly data on pressure levels. https://pernicus.eu/datasets/reanalysis-era5-pressure-levels?tab=overview, 2024. Accessed: 2024-12-10.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[13] C. Elkin and D. Sanekommu. How AI is helping airlines mitigate the climate impact of contrails. https://blog.google/technology/ai/ai-airlines-contrails-climate-change/, 2023. Google AI Blog.

[14] European Centre for Medium-Range Weather Forecasts (ECMWF). ECMWF reanalysis v5 (era5). https://www.ecmwf.int/en/forecasts/dataset/ecmwf-reanalysis-v5, 2025. ERA5 resolves the atmosphere using 137 vertical levels from the surface up to about 80 km.

[15] Gregory Flato, Jochem Marotzke, Babatunde Abiodun, and et al. Evaluation of climate models. *IPCC AR5 WG1 Chapter 9*, 2013.

[16] A. Frías, S. Gómez, O. Boucher, R. Teoh, U. Schumann, M. E. J. Stettler, E. Gryspeerdt, and et al. Feasibility of contrail avoidance in a commercial flight planning system: an operational analysis. *Environmental Research: Infrastructure and Sustainability*, 4(1), 2024.

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

[18] V. Grewe, K. Dahlmann, S. Matthes, and T. Greßhöner. Climate-optimized air traffic: Climate cost functions, routing strategies, and climate impact. *Atmospheric Environment*, 167:394–407, 2017.

[19] John Guibas, Hossein Azizpour, Vishal Mahadevan, and Jaideep Pathak. Adaptive fourier neural operators: Efficient token mixers for transformers. In *International Conference on Learning Representations (ICLR)*, 2022.

[20] H. Hersbach, B. Bell, P. Berrisford, and et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

[21] Eric J. Jensen, Owen B. Toon, Stephanie A. Vay, Jean Ovarlez, Robin May, Cameron Becle, Margaret LeMone, Cynthia Twohy, Bruce W. Gandrud, Ralf F. Pueschel, and Ulrich Schumann. Environmental conditions required for contrail formation and persistence. *Journal of Geophysical Research: Atmospheres*, 103(D4):3929–3936, 1998.

[22] Bernd Kärcher. Formation and radiative forcing of contrail cirrus. In Jacques Lenoble and Michael I. Mishchenko, editors, *Radiation in the Atmosphere and Ocean*, volume 4 of *Springer Series in Light Scattering*, pages 397–441. Springer, 2018.

[23] S. Kobayashi, Y. Ota, Y. Harada, and et al. The jra-55 reanalysis: General specifications and basic characteristics. *Journal of the Meteorological Society of Japan*, 93(1):5–48, 2015.

[24] Robin Lam, Yujia Chen, Suman Ravuri, and et al. Graphcast: Learning skillful medium-range global weather forecasting. *Science*, 2023.

[25] D. S. Lee, L. L. Lim, D. Holzer, and et al. Uncertainties in mitigating aviation non-co_2 emissions for climate and air quality using hydrocarbon fuels. *Environmental Sciences: Atmospheres*, 23(12), 2023.

[26] David S. Lee, David W. Fahey, Ashley Skowron, Matthew R. Allen, Ulrike Burkhardt, Qi Chen, Sara J. Doherty, Sasha Freeman, Piers M. Forster, Jan S. Fuglestvedt, Andrew Gettelman, Raúl R. De León, Ling L. Lim, Marianne T. Lund, Richard J. Millar, Bethan Owen, Joyce E. Penner, Giovanni Pitari, Michael J. Prather, Robert Sausen, and Laura J. Wilcox. The contribution of global aviation to anthropogenic climate forcing for 2000 to 2018. *Nature Communications*, 9(1):1–10, 2018.

[27] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

[28] Hermann Mannstein and Helmut Ziereis. A case study on the spreading of aircraft induced cirrus cloud. DLR Internal Report DLR-IB 511-96/03, Deutsches Zentrum für Luft- und Raumfahrt (DLR), 1996.

[29] An Nguyen, Albert Soret, Peter Dueben, and et al. Climax: A foundation model for weather and climate. *ICML*, 2023.

[30] Jaideep Pathak, Shashank Subramanian, Peter Harrington, and et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *Nature Machine Intelligence*, 2022.

[31] Florence Rabier, Heikki Järvinen, Erland Klinker, and et al. The ECMWF operational implementation of four-dimensional variational assimilation: I. experimental results. *Quarterly Journal of the Royal Meteorological Society*, 126:1143–1170, 2000.

[32] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. In *Advances in Neural Information Processing Systems*, 2021. arXiv:2107.00645.

[33] Stephan Rasp, Michael Pritchard, and Pierre Gentine. Deep learning to represent subgrid processes in climate models. *PNAS*, 115(39):9684–9689, 2018.

[34] S. Saha, S. Moorthi, H.-L. Pan, and et al. The NCEP climate forecast system reanalysis. *Bulletin of the American Meteorological Society*, 91:1015–1058, 2010.

[35] S. Saha, S. Moorthi, X. Wu, and et al. The NCEP climate forecast system version 2 (CFSv2). *Journal of Climate*, 27(6):2185–2208, 2014.

[36] R. Sausen, K. Gierens, M. Ponater, C. Frömming, V. Grewe, S. Matthes, and et al. Can we successfully avoid persistent contrails by small altitude adjustments of flights in the real world? *Meteorologische Zeitschrift*, 33(1), 2024.

[37] Matthias Schafer, Martin Strohmeier, Vincent Lenders, Ivan Martinovic, and David Basin. The opensky network: A large-scale ADS-B sensor network for research. In *Proceedings of the 2014 ACM Conference on Embedded Networked Sensor Systems*, pages 1–14. ACM, 2014.

[38] David Schmidt, Peter Dueben, Matthew Chantry, and et al. Machine learning for weather and climate modelling. *Quarterly Journal of the Royal Meteorological Society*, 149:1–47, 2023.

[39] E. Schmidt. Die entstehung von eisnebel aus den auspuffgasen von flugmotoren. Technical Report 44, Deutsche Akademie der Luftfahrtforschung, München und Berlin, 1941. Schriften der Deutschen Akademie der Luftfahrtforschung; also reprinted in: Jahrbuch der Deutschen Akademie der Luftfahrtforschung, 1940/1941, Berlin, pp. 126–135.

[40] E. Schmidt. *Einführung in die Technische Thermodynamik*. Springer, Berlin, 10 edition, 1963.

[41] U. Schumann. A contrail cirrus prediction model. *Geoscientific Model Development*, 5:543–580, 2012.

[42] Ulrich Schumann. On conditions for contrail formation from aircraft exhausts. *Meteorologische Zeitschrift*, 5(4):4–23, 1996.

[43] Ulrich Schumann. Influence of propulsion efficiency on contrail formation. *Aerospace Science and Technology*, 4(6):391–401, 2000.

[44] P. Spichtinger et al. Horizontal scales of ice-supersaturated regions. *Tellus A: Dynamic Meteorology and Oceanography*, 68(1):29020, 2016.

[45] B. Sridhar, H. K. Ng, F. Linke, and N. Y. Chen. Benefits analysis of wind-optimal operations for trans-atlantic flights. In *14th AIAA Aviation Technology, Integration, and Operations Conference*, page 2014, Atlanta, GA, 2014. American Institute of Aeronautics and Astronautics.

[46] X. Tan and et al. An assessment of the radiative effects of ice supersaturation based on in situ observations. *Geophysical Research Letters*, 43(11), 2016.

[47] R. Teoh, U. Schumann, E. Gryspeerdt, M. Shapiro, M. E. J. Stettler, and et al. Climate-optimized north atlantic flight planning reduces contrail radiative forcing. *Environmental Research Letters*, 17(6):064044, 2022.

[48] R. Teoh, U. Schumann, E. Gryspeerdt, M. E. J. Stettler, M. Shapiro, G. Boselli, and et al. Global aviation contrail climate effects from 2019 to 2021. *Atmospheric Chemistry and Physics*, 24:6071–6097, 2024.

[49] R. Teoh, U. Schumann, L. Johansson, and N. Metz. Aviation contrail avoidance through optimal small-scale flight path adjustments. *Environmental Research Letters*, 15(5):054001, 2020.

[50] R. Teoh, U. Schumann, and M. E. J. Stettler. Beyond contrail avoidance: Reducing the climate impact of aviation using a multi-disciplinary approach. *Meteorological Applications*, 27(5):e1958, 2020.

[51] Simon Unterstrasser, Klaus Gierens, and Peter Spichtinger. Contrail microphysics. *Atmospheric Chemistry and Physics*, 16:2059–2083, 2016.

[52] B. M. Varney. The argonne battle cloud. *Monthly Weather Review*, 49:348–349, 1921.

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[54] Warren M. Washington and Claire L. Parkinson. *An Introduction to Three-Dimensional Climate Modeling*. University Science Books, 2005.

[55] L. Weickmann. Wolkenbildung durch ein flugzeug. *Die Naturwissenschaften*, 7(34):625, 1919.

[56] K. Wolf, N. Bellouin, and O. Boucher. Long-term upper-troposphere climatology of potential contrail occurrence over the paris area derived from radiosonde observations. *Atmospheric Chemistry and Physics*, 23:287–303, 2023.

[57] K. Wolf, N. Bellouin, and O. Boucher. Long-term upper-troposphere climatology of potential contrail occurrence over the paris area derived from radiosonde observations. *Atmospheric Chemistry and Physics*, 23:287–303, 2023.

[58] F. Yin, V. Grewe, and S. Matthes. Impact on flight trajectory characteristics when avoiding the formation of persistent contrails. *Transportation Research Part D: Transport and Environment*, 65:466–480, 2018.

[59] Mingbao Zheng, Han Hu, Rui Xie, and et al. Fasterswin transformer. *International Journal of Computer Vision*, 2024.