Article

# The ozone radiative forcing of nitrogen oxide emissions from aviation can be estimated using a probabilistic approach

Check for updates

Pratik Rao [1,5] ✉, Richard Dwight [2], Deepali Singh [2], Jin Maruhashi [1], Irene Dedoussi [1,3], Volker Grewe [1,4] & Christine Frömming [4]

Reliable prediction of aviation's environmental impact, including the effect of nitrogen oxides on ozone, is vital for effective mitigation against its contribution to global warming. Estimating this climate impact however, in terms of the short-term ozone instantaneous radiative forcing, requires computationally-expensive chemistry-climate model simulations that limit practical applications such as climate-optimised planning. Existing surrogates neglect the large uncertainties in their predictions due to unknown environmental conditions and missing features. Relative to these surrogates, we propose a high-accuracy probabilistic surrogate that not only provides mean predictions but also quantifies heteroscedastic uncertainties in climate impact estimates. Our model is trained on one of the most comprehensive chemistry-climate model datasets for aviation-induced nitrogen oxide impacts on ozone. Leveraging feature selection techniques, we identify essential predictors that are readily available from weather forecasts to facilitate the implementation therein. We show that our surrogate model is more accurate than homoscedastic models and easily outperforms existing linear surrogates. We then predict the climate impact of a frequently-flown flight in the European Union, and discuss limitations of our approach.

The air transport sector has been growing rapidly in most regions of the world, transporting around 4.5 billion passengers in 2019[1]. Airbus[2] and Boeing[3] predict annual growth of about 5% a year in terms of revenue passenger kilometres (RPK, a measure of revenue collected from passengers for the distance travelled) for the next two decades, with the COVID-19 pandemic inducing only temporary disruption to this growth[4]. Aviation contributed to 3.5% of anthropogenic climate change in terms of effective radiative forcing in 2018 from aviation emissions since 1940[5], which is expected to grow continually unless effective political, technical and operational measures are undertaken[6].

The growth rate of fuel usage and consequently $CO_2$ emissions, has been lower than that of RPK due to various engineering novelties and higher passenger load factors[5]. Aviation transport efficiency has experienced an approximately eight-fold improvement since 1960, reaching 125 g($CO_2$) $RPK^{-1}$[5]. From 1960 to 2014, a 1.3% average reduction in fuel consumption per passenger-km at the global fleet level has been reported[7], which

addresses $CO_2$. On the other hand, non-$CO_2$ effects from jet fuel combustion from 1940 to 2018, contribute to about 2/3 of aviation's net warming in terms of effective radiative forcing with moderate to large uncertainties[5] and are currently not included in the international climate agreements such as the Paris Agreement[8]. The Kyoto Protocol designated the International Civil Aviation Organization (ICAO) as responsible for addressing international aviation emissions. While these emissions are estimated and reported, they are not included in countries' total emissions or targets, which poses an additional challenge[9]. The most important non-$CO_2$ effects include persistent line-shaped contrails, contrail-induced cirrus clouds[10,11], and nitrogen oxide ($NO_x = NO + NO_2$) emissions that alter the ozone ($O_3$), methane ($CH_4$) and stratospheric water vapour ($H_2O$) concentrations[12–17], all of which are greenhouse gases, and the emission of water vapour ($H_2O$)[18,19]. Studies[20,21] have found that rapid increases in aircraft emissions significantly contribute to global tropospheric ozone trends, radiative impacts and degradation of air quality.

[1]Faculty of Aerospace Engineering, Department of Control & Operations, Section Aircraft Noise & Climate Effects, Delft University of Technology, 2629 HS Delft, The Netherlands. [2]Faculty of Aerospace Engineering, Department of Flow Physics & Technology, Delft University of Technology, 2629 HS Delft, The Netherlands. [3]Department of Engineering, University of Cambridge, CB3 0DY Cambridge, United Kingdom. [4]Deutsches Zentrum für Luft- und Raumfahrt, Institut für Physik der Atmosphäre, 82234 Oberpfaffenhofen, Germany. [5]Present address: SRON Netherlands Institute for Space Research, 2333 CA Leiden, The Netherlands. ✉e-mail: P.V.Rao@tudelft.nl

Unlike $CO_2$, the climate impact of non-$CO_2$ effects from aviation is strongly dependent on location, such that climate impact can potentially be reduced by strategic routing of flights. This operational measure has been addressed with two different approaches: weather-dependent and weather-independent operation changes. Various studies[22–24] discuss weather-dependent options for mitigating aviation's climate impact whereby climate-sensitive areas are detected and avoided by aircraft. On the other hand, the weather-independent approach[25,26] involves quantifying the climate impact mitigation potential and related costs resulting from changes in aircraft operations using a multi-disciplinary model workflow. After analysing different flight altitudes and Mach numbers for more than 1000 flight routes, it was concluded[26,27] that a generally lower flight altitude and lower flight speed reduces the climate impact. This operational mitigation option can be combined with a redesign of the aircraft, as the original aircraft would be operated in off-design conditions. While the redesign further allows a climate impact reduction at no additional cash-operating cost and contributes thereby to increased eco-efficiency, the requirement of new aircraft is a downside.

Recent advancements in numerical weather prediction[28,29] and its compatibility with existing aircraft technology have prompted our focus on the weather-dependent feasibility of climate-optimised planning. Studies in this area are closely related to the development and use of climate change functions (CCFs[24]) for pre-defined locations on the globe. These functions quantify the climate effect of a locally confined aviation emission in terms of appropriate climate metrics and associated emission scenarios. These include the absolute global warming potential with pulse emissions (P-AGWP) and the global average temperature response with future increasing emissions (F-ATR) computed for contrails and the total $NO_x$ effect. A time horizon of 20 years and 100 years are taken into account for short-term and long-term climate effects, respectively. These metrics are derived from the computation of global annual mean instantaneous radiative forcing (iRF), which measures the change in the atmosphere's energy balance due to greenhouse gas emissions. If F-ATR is chosen, then the CCFs are measured in terms of temperature change per unit emission (K kg$^{-1}$) or per unit distance of contrail coverage (K km$^{-1}$). To test the concept of climate-optimised flight planning, CCFs were used as an objective function in an air-traffic optimisation routine. This approach aimed to avoid regions that were particularly sensitive to climate impact, and results indicated a large reduction potential. For example, a 25% reduction in P-AGWP over a time horizon of 100 years was shown to be possible for Westbound North-Atlantic flights for a slight increase in cash operating costs (0.5%) compared to conventional flights optimised for maximising airline profits. Larger reductions in climate impact were found to be the case for a time horizon of 100 years because of the reduced warming by $CH_4$, causing the total $NO_x$ effect to be negative in terms of P-AGWP. While the promising mitigation potential is evident in this study, the large computational expense of simulating CCFs in real-time hinders their operational application. As a solution, simple linear regression surrogate models called algorithmic climate change functions (aCCFs[30]) were obtained by regressing CCF data against local atmospheric variables at the time of emission. However, aCCFs exhibit limitations, such as low coefficient of determination ($R^2$) values and large variance in training data, making the potential improvements in climate impact unclear. A recent study[31] explored the effectiveness of using $O_3$ aCCFs to optimise aircraft trajectories for reducing aviation-induced climate impact resulting from $NO_x$ emissions. The study found that while $O_3$ aCCFs provide reasonable mean estimates, they are limited to certain geographical areas (parts of the Northern Hemisphere), are deterministic, and lack uncertainty estimates in their predictions. Thus, past studies have developed CCFs based on chemistry-climate model (CCM) simulations in the North Atlantic, built a simple surrogate model (aCCFs), and analysed them in the context of short-term $O_3$ iRF. Deterministic surrogates such as the aCCFs only provide point estimates (mean) with no confidence intervals (uncertainty estimates). If, for instance, the climate impact mitigation potential of a flight is high, but the level of confidence is very low, it is beneficial to avoid this trajectory altogether. A flight optimisation tool can

take these into account to suggest robust and effective flight paths. Both homoscedastic and heteroscedastic surrogate models yield uncertainty estimates, but the latter is more suitable because variability in predictions is not expected to be uniform subject to changing environmental conditions. Thus, we build and present an accurate surrogate model that is both global and heteroscedastic. This provides reliable uncertainty estimates for short-term $O_3$ iRF, to enable global climate-optimised flight planning in the future.

Here, we develop probabilistic algorithmic climate change functions (paCCFs) as a replacement for aCCFs via a two-step approach: (i) firstly, by using the first comprehensive global dataset based on the CCF approach[24,32–34], to recalculate iRF of $O_3$ induced by $NO_x$ (i.e., $O_3$ iRF) in more regions (North America (N. America), South America (S. America), Eurasia, Africa and Australasia) and days for a range of cruise level altitudes (Fig. 1) thereby encouraging the possibility of global flight planning, and (ii) secondly, by formulating a corresponding high-accuracy probabilistic surrogate model using a chained Gaussian process (GP) regression model that is heteroscedastic to predict $O_3$ iRF with reliable uncertainty estimates using the most influential spatial and meteorological features locally (Fig. 2). GP regression is a Bayesian nonparametric technique that exhibits great flexibility and captures more information about the data by using more parameters as the dataset grows. Moreover, predictions are made with varying confidence levels, using the most influential features obtained using feature selection techniques, which is especially desirable for the non-$CO_2$ effects of aviation. We show that both, the chained (heteroscedastic) and the standard GP model (homoscedastic) perform well and significantly better than the deterministic aCCFs. The chained GP model reproduces the data distribution more accurately than the standard GP model and has the added advantage of providing varying confidence levels for their predictions on test data. Some statistical outliers are underestimated by the mean predictions from both GP models but exhibit a large enough variance to capture many of them. We demonstrate the improvements by applying the method exemplarily to actual flight routes in the European airspace, enabling climate impact predictions, including confidence levels. The predominant source of uncertainty in our GP models arises from relying on feature information solely at the release location, thus neglecting the broader weather patterns' influence on emissions transport and the corresponding climate impact; we explore potential improvements in our "Discussion" section.

## Results
### Data on the relation of local $NO_x$ emissions to $O_3$ iRF

The development of our paCCFs is based on the work of the studies[33,34], which investigated the short-term $O_3$ iRF from local aviation $NO_x$ emissions at pre-defined points across the globe on 2 chosen days using a CCM. The emission regions are representative of aircraft flying at typical subsonic cruise levels ($\approx$10–12 km) and consist of 28 release locations per region (N. America, S. America, Eurasia, Africa and Australasia) at atmospheric pressures of 200, 250, and 300 hPa (Fig. 1a). The CCM data corresponding to the iRF is visualised for each pressure level and all emission scenarios in Fig. 1b. Since there is no clear pattern and there is large variation in the data, it is informative to detect statistical outliers. The outliers are those data points that are larger than the Interquartile Range (IQR) ($>Q_3 + 1.5 \times$ IQR) or smaller than the IQR ($<Q_1 - 1.5 \times$ IQR), where $Q_1$ and $Q_3$ represent the first and third quartiles, respectively. Figure 1b also depicts a boxplot with the 15 outliers, all of which deviate from the first and third quartile by more than 1.5 times the IQR. These outliers (highly climate sensitive regions) are associated with the pressure levels of 200 and 300 hPa, and are mostly associated with Australasia. These elevated iRF values, especially in Australasia, are explained by the increased efficiency of $O_3$ production for a given quantity of $NO_x$. This is due to the heightened sensitivity of a $NO_x$-deficient atmosphere[13,14,33,35]. On the other hand, the lowest mean climate impact from $NO_x$ corresponds to 250 hPa, and is consistent with another study[32]. Figure 1c depicts the evolution of $NO_x$ and $O_3$ over the simulation period starting from 1st January 2014 for two release locations. The solid line represents the outlier associated with grid index '2' of Australasia at 200 hPa
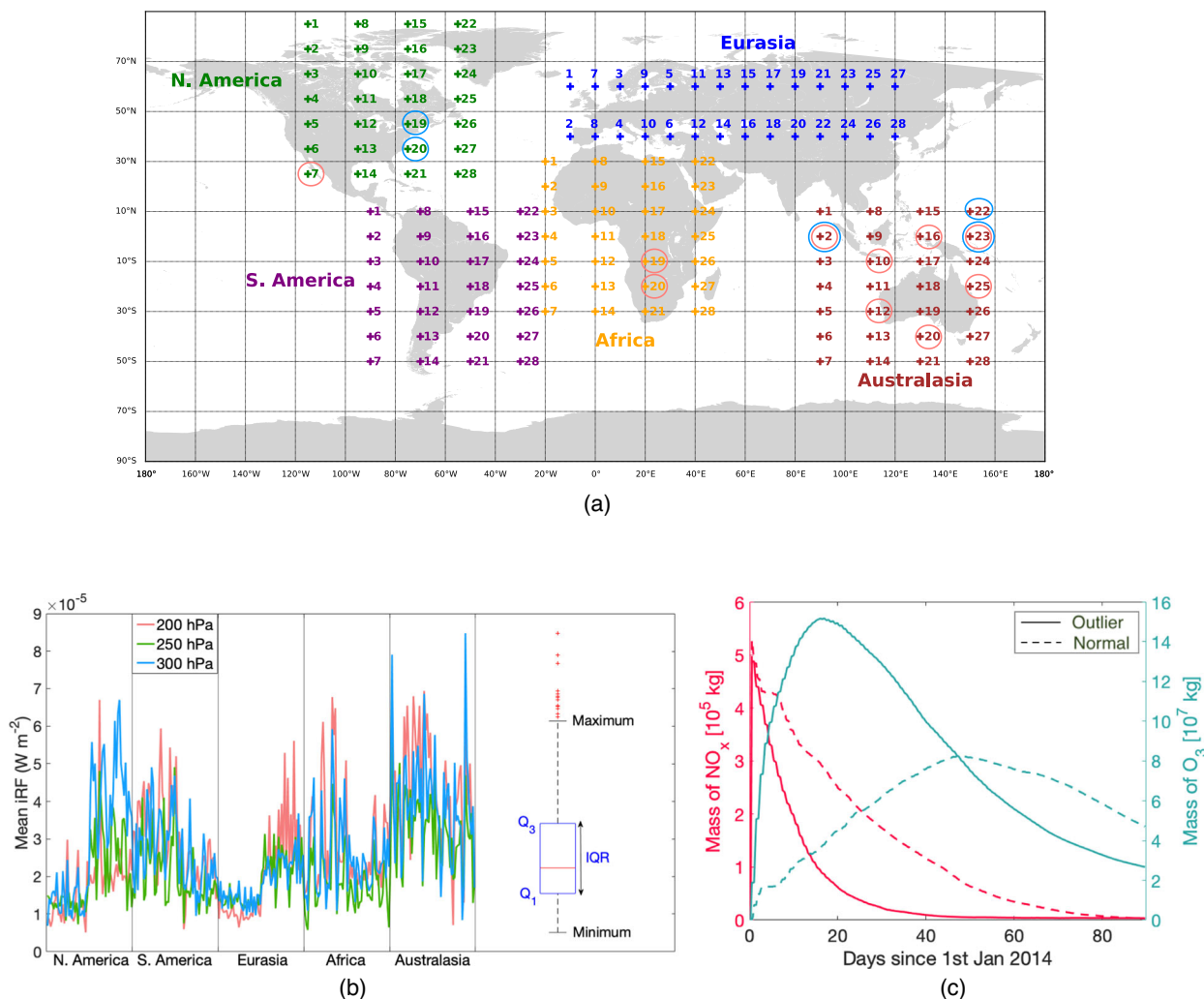
(a)



(b)



(c)

**Fig. 1 | The emission locations and corresponding iRF [W m$^{-2}$] dataset from the CCM simulations for the three cruise pressure levels. a** The time region grid (trg) comprising 28 emission locations for each of N. America (green), S. America (purple), Eurasia (blue), Africa (mustard) and Australasia (crimson red), for which the iRF is calculated. This figure is adapted from ref. 33 and falls under the Creative Commons Attribution 4.0 license. The salmon pink and sky blue circles represent regions associated with statistical outliers in the iRF dataset corresponding to 200 and 300 hPa, respectively. No outliers are present for 250 hPa. **b** The iRF values, where the horizontal axis represents the corresponding geographical regions per pressure level (200 hPa in salmon pink, 250 hPa in bright green, 300 hPa in sky blue) where emissions were released. The boxplot of the same data is shown, where the red

horizontal line represents the median and the 15 outliers are marked with red crosses. The minimum, maximum, first and third quantiles ($Q_1$, $Q_3$), and Inter Quantile Range (IQR) are labelled in the plot. For an alternative visualisation of this data in terms of a scatter plot, see Supplementary Fig. S1). **c** The evolution of global mean tracer mass of $NO_x$ (red pink) and $O_3$ (teal green) associated with two trgs over the 3-month simulation period since 1st January 2014. The solid lines represent the outlier associated with trg index '2' of Australasia at 200 hPa and the dashed lines represent trg index '4' of N.America at 250 hPa. The peak $O_3$ mass and the area under the curve are significantly larger for the outlier, ultimately resulting in a larger iRF.

and the dashed line represents grid index '4' of N.America at 250 hPa. Although the $NO_x$ is consumed faster for the point associated with the outlier, the peak $O_3$ mass and the area under the curve are significantly larger, indicating a greater $O_3$ production, and a larger iRF as a result. Apart from the magnitude of the peak $O_3$ mass, its position shows that it also occurs much sooner. The $NO_x$ and $O_3$ lifetimes at the outlier location (high climate sensitivity) are approximately 10 and 41 days, respectively. In contrast, at the other emission point (relatively low climate sensitivity), the lifetimes are around 26 and 82 days, respectively. Lifetime is defined here as the time it takes for the peak mass of the species to decrease by a factor of e (thus, the e-folding time) due to chemical reactions, deposition, or other removal processes. In the high climate sensitivity region, $NO_x$ is more reactive, leading to a higher $O_3$ production efficiency. Consequently, the peak $O_3$ concentration occurs much earlier and declines steeply afterwards. The temporal evolution of $NO_x$ and $O_3$ is characterised by high variability. For example, lifetimes of 20 ± 11 days for $NO_x$ and 72 ± 26 days for $O_3$ have

been reported in the North-Atlantic region[24], which aligns roughly with another findings[36]. Here, the $O_3$ lifetime associated with the outlier location in Australasia falls outside this range. Thus, we see from the data that the relation between $NO_x$ emissions and $O_3$ forcings is not straightforward, making a reliable surrogate model for predicting $NO_x$–$O_3$ impact even more essential.

**Towards probabilistic algorithmic climate change functions**

The fundamental issue lies in the computational burden borne by chemistry-climate models (CCMs), which can only be executed for a limited set of emission scenarios and cannot be run effectively in real-time. We start with the premise that these CCM outputs serve as our potentially noisy ground-truth data, and consequently, we do not account for epistemic uncertainties within the CCM. Our aim is to reproduce these climate impact forecasts by using information from influential variables (features) at the emission source, i.e., locally. However, the variability in climate impact is not
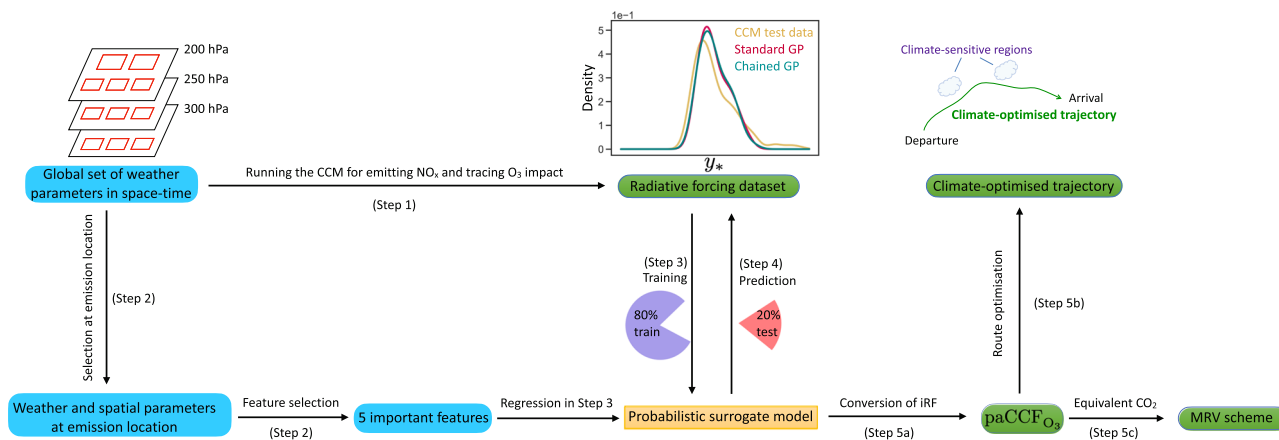
**Fig. 2 | A schematic representation to predict climate impact and its potential application in climate-optimised flight planning and use in an MRV scheme.** A high-cost CCM requires a wide range of input parameters (such as boundary conditions, emission inventories, weather data, initial conditions, chemical reaction rates of various atmospheric species, etc.) to accurately simulate complex phenomena in the atmosphere, and is used to generate the radiative forcing dataset (iRF) from aviation $NO_x$ emissions released at pre-defined locations for three pressure altitudes (200, 250, and 300 hPa). We use 80% of this data ($y$) to train probabilistic surrogate models based on chained and standard Gaussian processes (GP). Five local features are determined using objective feature selection techniques to use as predictors for the GP regression models. Predictions are made using the probabilistic surrogate models on the remaining (20%) test data ($y_*</Subscript>$) and their corresponding probability density functions (pdf) are visualised in the panel. For an arbitrary day, the feature information is readily available from local forecasts, and the corresponding predictive distribution of iRF from the chained GP can be converted to paCCF$_{O_3}$, and used as an objective function in an air traffic optimisation tool to generate climate-optimised trajectories that avoid the most climate-sensitive regions. Additionally, the prediction can be converted into equivalent $CO_2$ to enable its potential use in the MRV scheme.

constant; it fluctuates across different geographic areas, over time, and under varying input conditions (Fig. 1b). In statistical terms, this phenomenon is heteroscedastic. Conventional deterministic and most probabilistic surrogates disregard this phenomenon, providing an incomplete view of the climate response. In this section, we introduce a probabilistic framework designed to train a surrogate that takes heteroscedasticity into account, utilising available data to make accurate climate impact predictions along with uncertainty estimates. Furthermore, we explore the potential of employing this model for climate-optimised flight planning on a global scale in the context of climate-conscious decision-making. Our framework consists of five main steps, of which the final one is a future suggestion, as shown in Fig. 2:

Step 1: Run the CCM to emit $NO_x$ at specific points across the globe and trace the $O_3$ impact on a supercomputer which outputs a wide range of data, including weather-related variables, chemical concentrations, and $O_3$ iRF. The latter is our quantity of interest, and therefore other aviation $NO_x$ effects are not taken into consideration.

Step 2: Select a large number of weather and spatial parameters at the emission location from the global set. That is, the data is local. Next, perform objective feature selection firstly by calculating the mutual information (MI) between the iRF and these local parameters and consider the 10 highest scoring features (shown in bold text in Supplementary Table S1). Secondly, we narrow the selection down to five features with the help of Automatic Relevance Determination (ARD[37]), due to the curse of dimensionality. We find that the five most important features are dry air temperature ($T$), geopotential ($\phi$), solar irradiance ($S_r$), zonal wind velocity ($u_w$) and release location (latitude) of emissions ($R_{lat}$). We know that solar irradiance provides the energy needed for photochemical reactions for $NO_x$ in the atmosphere, leading to the formation of $O_3$. The corresponding increase in $T$ due to increase in solar irradiance also influences these reactions[38], while $R_{lat}$, $\phi$, and $u_w$, govern the location and subsequent pathways for $NO_x$ emissions. While background $NO_x$ is known to have an influence on $O_3$ production, a low correlation was found between its value at the emission location and the $O_3$ iRF. In this form, background $NO_x$ is not a good predictor for photochemistry about ten days since the release of emissions[39] and is thus excluded as a feature. The set described above is used as a proxy of the most influential features for step 3.

Step 3: Split the iRF data ($d$), into a training set (80%), $y$ and a test set (20%), $y_*$. The former dataset is used to train probabilistic surrogate models based on a chained GP model[40] that is heteroscedastic, and a standard GP model[41] that is homoscedastic, using the features selected in step 2. The choice of an 80–20 data split is standard practice in statistical learning.

Step 4: Estimate the probability density function (pdf) of the test set from the CCM and the predictions made on it by the chained and standard GP model. Visually, the estimated pdfs of the two GP models, which are always Gaussian, have a close resemblance and exhibit a considerable overlap with the data distribution. However, the position and magnitude of the peak of the chained GP pdf aligns more closely with that of the CCM pdf compared to the standard GP model (panel above "Radiative forcing dataset" in Fig. 2). A number of metrics are used to assess the performance of the prediction (see "Methods" for more details).

Step 5: This is a future suggestion, which entails the use of the chained GP model in predicting an improved estimate of iRF on an arbitrary day, and converting them to probabilistic algorithmic climate change functions (step 5a) for $O_3$ (paCCF$_{O_3}$), which represent the Average Temperature Response (ATR) of $O_3$ over a selected time horizon (e.g. 20 years) caused by a local aviation $NO_x$ emission. This can then be used as an objective function in an air traffic optimisation tool (e.g., AirTraf[42]) to generate climate-optimised trajectories, that avoid the most climate-sensitive regions (step 5b), or converted into an equivalent $CO_2$ effect[43] to enable its potential use in e.g., the forthcoming EU-wide Monitoring, Reporting, and Verifying (MRV) scheme (step 5c).

**Estimating climate impact**

After randomly selecting 80% of the dataset in training ($y$) the GP models, we visualised their climate impact estimates on the remaining 20% test data ($y_*$) in the form of pdfs (Fig. 2). The splitting of the data is done to prevent overfitting, and thus provide models that can be generalised. The two characteristic aspects of GP models are the mean and variance estimates. Figure 3a shows the mean predictions from the chained GP model against the test data. It can be seen that the six largest values of iRF test data are underestimated (enclosed by the magenta box) by the mean of the GP model, and are part of the set of outliers shown in Fig. 1b.
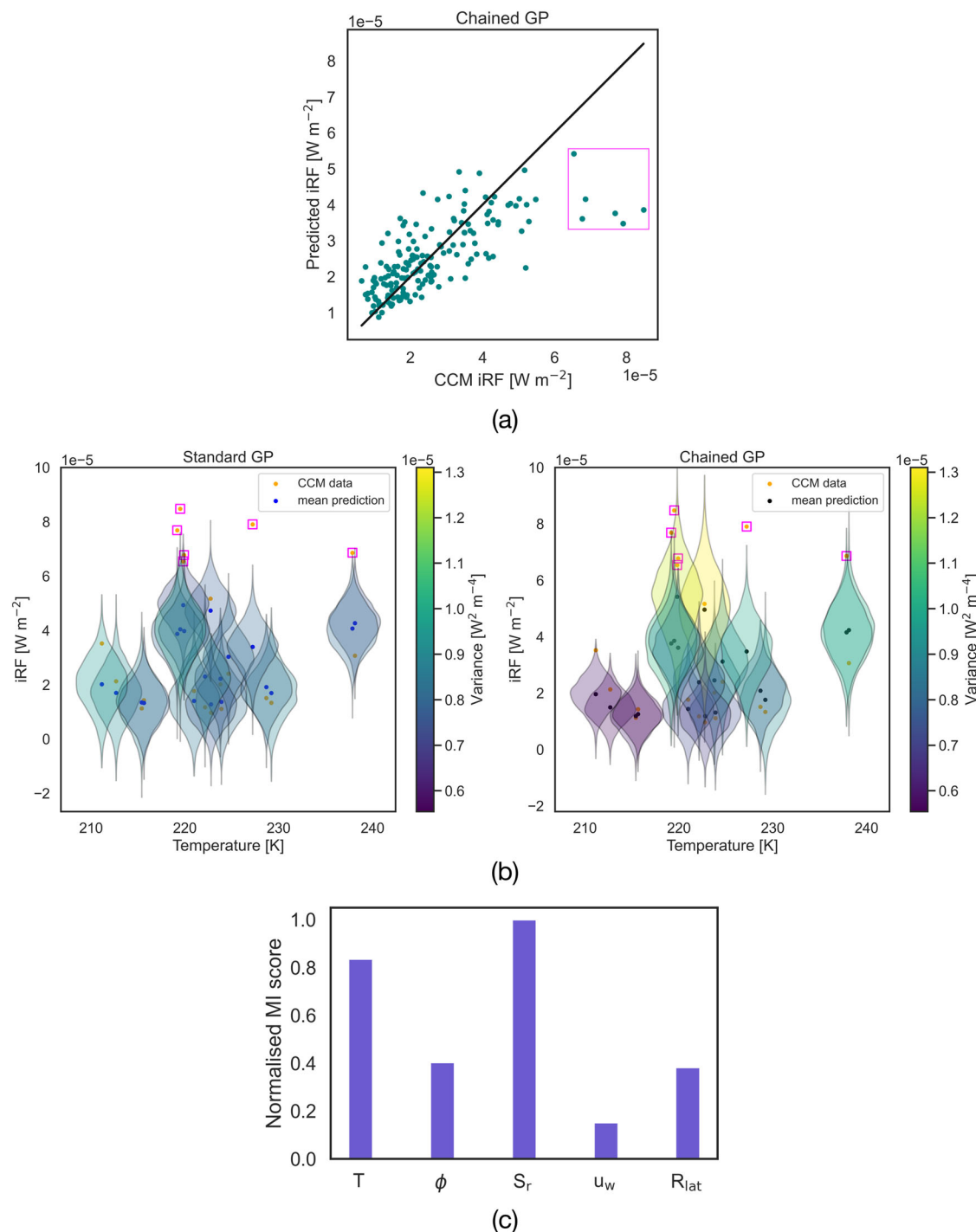
(a)



(b)



(c)

**Fig. 3 | Comparing predictions from the GP models and the test data. a** Mean predictions from the chained GP model (in teal) plotted against the CCM test data on the horizontal axis. The box (in magenta) shows the outliers from the test set. **b** Violin plot of the predictions from the two GP models, plotted against one of the features (temperature), for 20 chosen test indices. The variance predicted for each test index by the standard GP model (left) and chained GP model (right) is colour-coded with respect to its magnitude shown in the colour bars. For the standard GP (left), the variances are almost the same for changes in temperature and do not capture most outliers, but for the chained GP (right), the variance changes, as it is a function of the feature space. For example, the prediction at 210 K is associated with low uncertainty, while 220 K is associated with a relatively large uncertainty. The mean predictions of the standard and chained GP models are depicted as blue and black points, respectively and the test data points are shown in orange. The magenta boxes indicate the outliers, and it is seen that the chained GP captures most of them. **c** Bar chart (in slate blue) containing normalised MI scores between each feature and iRF of the test dataset.

Since our feature space is high-dimensional ($\mathbb{R}^5$) and thus difficult to visualise, we plot test data and predictions against one of the features (i.e., temperature) in Fig. 3b. Here, only 20 predictions are shown from the test set so as to avoid clutter. The left and right panels of each figure represent the mean and variance predictions from the standard GP and chained GP models, respectively. The standard GP model has a nearly constant variance (uncertainty estimate) and captures most data points but misses some of the outliers. On the other hand, the chained GP provides varying uncertainty estimates and captures almost every single test data point. A smaller (larger) variance corresponds to a higher (lower) confidence in predicting the target.

**Fig. 4 | Convergence study of GP models and comparison with linear regression models.** The complete dataset ($d$) is split into training data ($y$), which never exceeds 80% $d$, and test data ($y_*$) is fixed at 20% $d$. We analyse if the standard GP (in crimson red) and chained GP (in teal) models perform better by gradually increasing the amount of training data for four scenarios: 40%, 60%, 80%, and 100% of $y$ and testing them on $y_*$. **a** Convergence of $R^2$ of the mean predictions on the test data increases as more training data is used. **b** The statistical distance between the predictive distribution and the test data distribution shrinks in terms of KL divergence as more training data is used. **c** We compare the mean predictions of the linear regression model (in bright blue) against test data with ($T, \phi, T \cdot \phi$) as the feature basis. The mean predictions from the standard GP model (in crimson red) are overlaid as a reference, and the $R^2$ values are included in the legend.



(a)



(b)



(c)
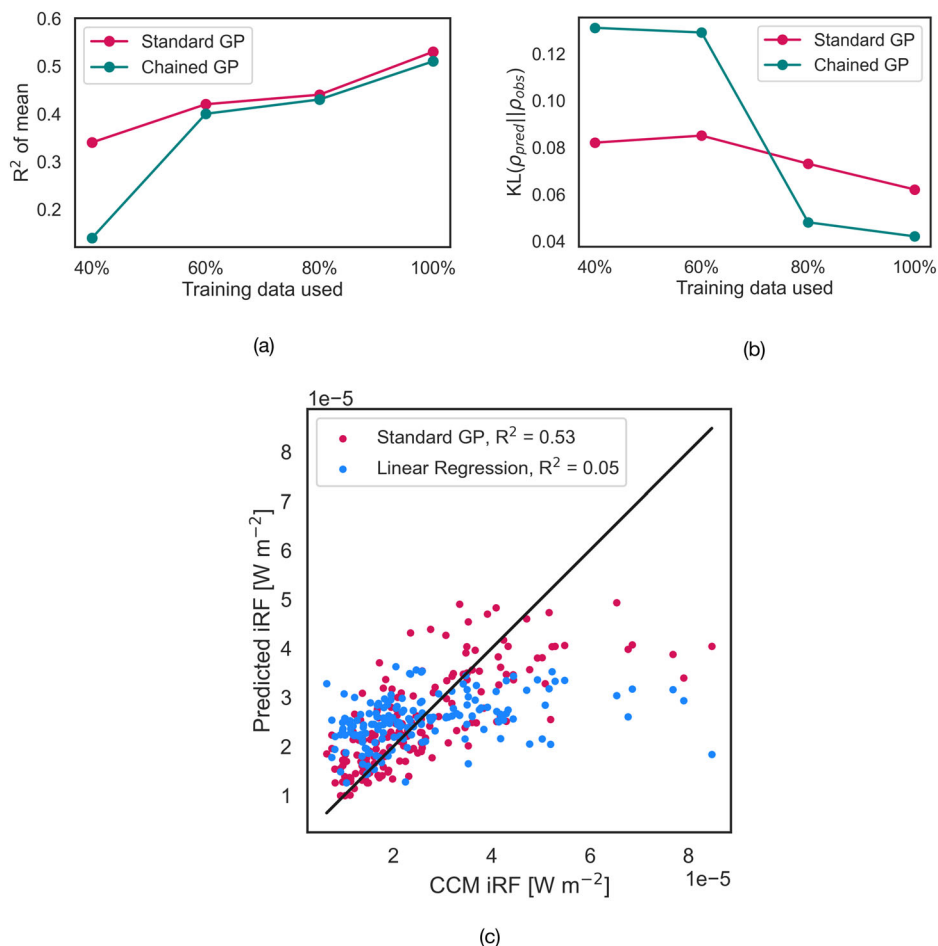
It is seen though, that while the mean prediction is similar to the standard GP model, the variance in some cases is smaller (larger) and it captures these test data points with relatively higher (lower) probability. The predictions corresponding to outliers are characterised by large variances. Thus, the climate impact is estimated with varying confidence levels, making the chained GP model more realistic. We also see that there is a nonlinear trend between iRF and temperature, which cannot be captured by simple parametric surrogates that have been used so far[30]. Violin plots against the other features are available in the Supplementary Figs. S2–S5). Going further, we score each of the five features based on the test dataset, by computing their corresponding MI with iRF. We then normalise these scores by dividing them by the highest score and deduce that solar irradiance ($S_r$) and zonal wind velocity ($u_w$) are the most and least important predictors of aviation $NO_x$–$O_3$ effects, respectively (Fig. 3c).

**Performance of the model and comparisons**
We evaluate the performance of the GP models and compare them to an existing linear regression model by following the $R^2$ test for the mean predictions on the test data. Additionally, we measure the statistical distance between the predictive distribution of the GP models and the test data distribution using the Kullbak–Leibler (KL) divergence[44].

Keeping the test data fixed, we vary the amount of data used for training the models: 100%, 80%, 60%, and 40% of $y$, which serves as a simple convergence study (Fig. 4). The $R^2$ increases as the amount of training data is increased, as shown in Fig. 4a. The $R^2$ is only marginally higher for the standard GP model (=0.53) than the chained GP model (=0.51) when all the training data is used (scenario in Fig. 3) and this can be attributed to the latter requiring more data to learn both, the mean and variance functions (see the "Methods" section). Hence, both models show basically the same quality in this respect. Since KL divergence measures the statistical distance

between the distributions, the value should fall as more training data is used by the GP models. For both GP models, there is a reduction in KL divergence as the amount of training data increases beyond 60% (Fig. 4b). For the chained model, there are sharper changes while moving from 60% to 80% $y$, which can be attributed to the requirement of learning two latent functions (and hence more hyperparameters) at the same time, which requires more data. The global minimum in KL-divergence is lower for the chained GP model, which indicates a higher accuracy compared to the standard GP model.

So how do these surrogate models perform compared to a linear regression model? We use the best available model derived before[30], which involves using $T$, $\phi$, and $T \cdot \phi$ as features. The third feature, which is the product of $T$ and $\phi$ was used to model the non-linearity of $NO_x$–$O_3$ effects. After training a linear regression model with these features for our data, we visualise the mean predictions and also overlay the predictions from the standard GP model as a reference in Fig. 4c. The $R^2$ value of 0.05 is too low, and Fig. 4c depicts that the linear regression model does not fit the data at all. The $R^2$ value of ≥0.51 for the GP models is a reflection of relatively high accuracy, considering the relative score of linear models (=0.05), the limited test dataset which contains outliers, and the complexity of the phenomenon that is being modelled. The higher $R^2$ of 0.41 that was obtained in the earlier study[30] for $O_3$ aCCFs can be attributed to two reasons. Firstly, the data was based only on the North Atlantic region and secondly, all the data was used for regression, unlike in this work.

**Climate impact estimation for a frequently flown flight**
We showcase the application of the chained GP model to forecast the short-term $NO_x$–$O_3$ climate effects of a commonly operated flight within the European Union, considering various departure times on a selected day. Our analysis is based on actual flight path data retrieved from the

Eurocontrol database[45], and the details of these flights are provided in Table 1. In addition to this, we include a great circle flight path, which is the shortest path, for the same origin-destination pair at a typical cruise altitude of 10.7 km to highlight differences in predictions. The weather-related feature data ($T$, $\phi$, $S_r$, $u_w$) are obtained from the CCM corresponding to the departure time of the flights to enable the evaluation of the chained GP model. Figure 5b shows the three actual flight paths and the great circle path from take-off from the departure airport in the Netherlands (Amsterdam Schiphol (EHAM)) to the landing at the arrival airport in Spain (Madrid–Barajas airport (LEMD)), while Fig. 5a shows the climate impact prediction from the chained GP model during the cruise phase of the flights. The predictions are offered in terms of $W\,m^{-2}\,kg(NO_2)^{-1}$. This is because the CCM-generated $O_3$ iRF data is associated with a release of $5 \times 10^5$ kg of NO per emission point, as shown in Fig. 1b. However, actual flights do not emit this amount per location and emission models report $NO_x$ emissions in $kg(NO_2)$. We thus convert the emission from $kg(NO)$ to $kg(NO_2)$ and normalise the climate impact prediction from the GP model with respect to this mass. To obtain the $O_3$ iRF estimate of these flight paths (in $W\,m^{-2}$), the $NO_2$ emitted at each time step can be multiplied with the scaled predictions (in $W\,m^{-2}\,kg(NO_2)^{-1}$). Since this data is not publicly available from Eurocontrol, we do not perform this step here.

In Fig. 5a, it can be seen that the three actual flight paths have their maximum mean impact at 45°N. Flight 3 predicts a high $O_3$ iRF in this region, with lower uncertainty, which translates to a high probability of reducing $O_3$ iRF impact if this region is avoided. Figure 5b also shows a colour map of the wind velocity field embedded with geopotential height contour lines at a cruise level of 216 hPa at an arbitrary time of the day on which the flights take place. These fields show the weather patterns, including the jet stream and transport pathway affecting the region of interest. The orange diamond in Fig. 5b represents the region where the maximum mean climate impact was predicted for the three flights. Emitted

species from these flights are expected to be transported to lower latitudes by the jet stream and this region has been identified earlier[39] with a large and early $O_3$ maximum compared to higher latitudes characterised by a small and late $O_3$ maximum. When we look at the variation of uncertainty estimates throughout the cruise flight, it is lowest for flight 1, and highest for flight 3. The total mean climate impact associated with the short-term $NO_x$–$O_3$ effect for the cruise level, however, is largest for flight 1 and smallest for flight 2. The difference in prediction between the great circle path and the three actual flight paths is clear, indicating that the GP model is influenced by the chosen features, and partly by the noise. The release location and flight timing affect the mean predicted $O_3$ iRF, however this prediction is associated with a large uncertainty range. The varying confidence intervals, subject to varying input conditions, indicate that optimising flights based on climate impact using a heteroscedastic model (Fig. 2) allows us to estimate the probability of achieving a reduction in short-term $O_3$ iRF. This means that the differences in uncertainty estimates translate to a probability of a certain level of climate impact and, therefore, if used in a predictive manner, a probability of climate impact avoidance. It is important to note that this probability reflects the uncertainty inherent in the surrogate model only since we have assumed the source data to be ground truth.

## Discussion

The application of climate-optimised routing requires us to be able to predict the climate impact of various forcing agents (here, $NO_x$ emissions) as a function of readily available forecasts of relevant features. We show that probabilistic surrogate modelling is useful in providing uncertainty estimates to the modelled climate impact from aviation $NO_x$ in terms of iRF. Both, the chained (heteroscedastic) and the standard GP model (homoscedastic) perform well and significantly better than the deterministic aCCFs[30]. The chained GP model reproduces the data distribution more accurately than the standard GP model and has the added advantage of providing varying confidence levels for their predictions on test data. Additionally, we found that temperature, geopotential, solar irradiance, wind velocity and the release location are all important predictors of $O_3$ iRF as a result of aviation $NO_x$. Thus, with this independent approach applied to a global dataset, features in addition to the earlier study[30] were found to be influential. Furthermore, it was observed that the uncertainty estimates due to varying input conditions are considerable (Fig. 3b and Supplementary Figs. S2–S5). While the chained GP model has a large enough variance to capture most outliers, the mean underestimates them, which poses a challenge in identifying high-risk areas

## Table 1 | Characteristics of three flights from EHAM to LEMD on 1st June 2016 at different departure times

| Flight | Departure time (UTC) | Arrival time (UTC) | Flight duration (h) |
|--------|----------------------|--------------------|---------------------|
| 1      | 09:25                | 11:49              | 2.4                 |
| 2      | 10:20                | 12:49              | 2.5                 |
| 3      | 18:49                | 20:56              | 2.1                 |



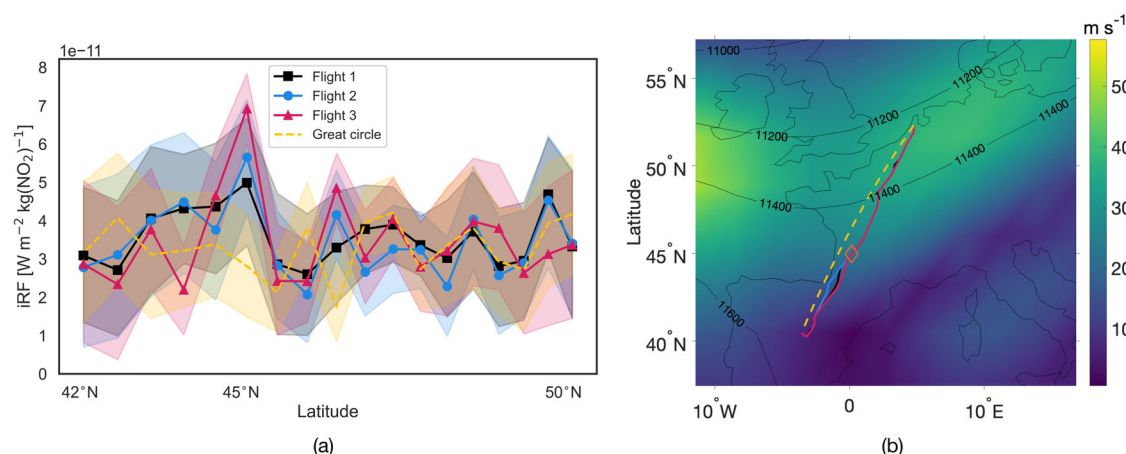(a)                                                                                       (b)

**Fig. 5 | Characteristics of the three flights described in Table 1 and the great circle path. a** The climate impact prediction from the chained GP model for the three actual flights (solid lines) and the great circle path (dashed line) at cruise level in terms of scaled $O_3$ iRF [$W\,m^{-2}\,kg(NO_2)^{-1}$]. The horizontal axis represents the latitude. The lines and the shaded areas represent the mean prediction and 95% confidence interval, respectively, for all four paths. The markers for the three actual flights denote specific points along the trajectory at cruise level, at which the model is

evaluated. At 45°N, these three flights show a peak mean climate impact. **b** The actual reported flight paths (solid lines) shown in 2D for the three flights, including take-off and landing retrieved from the EUROCONTROL database[45] and the 2D great circle path (dashed lines). The colour map and contour lines represent the wind velocity [$m\,s^{-1}$] and geopotential height [m], respectively, at a cruise atmospheric pressure of 216 hPa, while the orange diamond represents the region where maximum mean climate impact is predicted for the three flights.

for climate-optimised planning. This limitation must be overcome with additional investigations to better understand these outliers and the driving forces. Since the Gaussian processes represent a nonparametric approach, incorporating more iRF data into the model's training process could potentially reveal a reduction in uncertainties, ultimately leading to a decreased risk in climate-optimised flight planning. The large improvements relative to previous surrogates are demonstrated by applying the method exemplarily to actual flight routes in the European airspace, enabling climate impact predictions, including confidence levels. While ERA re-analysis data has been used as input in our GP models as a first step, using real-time ECMWF forecasts for future predictions and planning is not a problem. This is because, for $NO_x$–$O_3$ effects, the transport pathways seem to be most important, and these are well captured by both forecasts. We additionally note that if flights are optimised only with respect to short-term $NO_x$–$O_3$ effects, they need not effectively mitigate the effect of other non-$CO_2$ effects such as water vapour and contrails. Thus, once paCCFs are available for other forcing agents, the net climate impact must be minimised for an acceptable increase in operating costs.

We acknowledge that using a single CCM (here, EMAC), is a limitation of this study. While it is challenging to replicate this study using other CCMs, we acknowledge that different models will yield different responses due to variations in their underlying processes and parameterisations. While the approach and results used in this study are certainly influenced by the specific characteristics of the EMAC model, we also note that EMAC has undergone extensive validation[46,47] to assess its performance and accuracy in simulating various aspects of the atmosphere. For example, a study[48] evaluated the background chemistry setup against the observations of tropospheric $O_3$ and its precursors made by another study[49] for the years between 1983 and 2001 and found compelling agreement. Additionally, another study[50] performed a comprehensive evaluation across multiple models, including EMAC, to assess the impact of aircraft $NO_x$ emissions on the atmosphere by varying cruise altitudes. Apart from EMAC, these included four other models that were employed to include detailed representations of tropospheric and stratospheric chemistry to cover the upper troposphere-lower stratosphere region. The models were found to be in good agreement with other studies (e.g., refs. 12,25,51) with respect to chemical perturbations in $O_3$ and the sensitivity of aircraft $NO_x$ emissions to altitudinal changes. If we consider the submodel AIRTRAC, used by EMAC to compute the short-term $O_3$ iRF, a direct validation is not possible since most of the simulated effects cannot be measured directly. However, the recent study[34] performs a direct inter-comparison of three methods which include Lagrangian tagging (using AIRTRAC), Eulerian tagging, and the perturbation approach using the same regions and days considered in our manuscript. The study found differences between the tagging and perturbation approaches that align well with the literature (e.g., refs. 52–54). Additionally, various sensitivity studies were performed[24] on the basis of temporal and horizontal resolutions, and the number of air parcel trajectories used. As far as the chemistry is concerned, comparisons of the temporal evolution and lifetimes of $NO_x$ and $O_3$ were found to be roughly in agreement with another study[36]. The resulting RF of $O_3$ and $CH_4$ (from $NO_x$) and $H_2O$ were also found to be well within the range of other studies (e.g. refs. 55,56). Finally, even results for contrail properties and radiative impacts were found to agree reasonably well[57]. However, as noted earlier, different CCMs might produce varying results, which can be attributed to their unique configurations, parameterisations, tagging schemes for chemistry, and radiative transfer codes. This variability underscores the importance of multi-model evaluations to capture the range of possible outcomes and enhance our understanding of model-specific influences on climate impact prediction and mitigation strategies.

A surrogate model can be only as good as the quantity and quality of the data used to train it. To start with, data are available for aviation $NO_x$ emissions released on 2 days (and for three pressure levels), which constitutes $n = 840$ data points. Standard GP models, without a sparse approximation of the covariance matrix, has $\mathcal{O}(n^3)$ complexity limited by $n < 10,000$, or $\approx 24$ days. Thus, while there are computational limitations to

calculate the data, running the full order model for a few more days and regions (e.g., remaining parts of Asia), characterised by other seasons and cyclical events such as El Niño and La Niña could help further advance our understanding of local aviation $NO_x$ emissions on global warming. In addition, there are certain physical limitations in the $NO_x$–$O_3$ chemistry calculated by the EMAC submodel AIRTRAC[24], which is part of the CCM data generation process (see the "Methods" section). While these limitations include the use of linearised reaction rates and fewer tagged processes, simple and accurate correction factors were derived[34] considering various emission scenarios. These values can be used to appropriately scale the iRF data used by the GP models, but we also note that plume-scale processes are neglected, potentially leading to an overestimation of the short-term $O_3$ iRF by about 20–30% (e.g., refs. 58,59). In order to convert iRF to CCFs for the regions considered in this paper, the parameterisation has to be revised with the aid of additional simulations, which will further improve this work. Moreover, this is both, an atmospheric transport and chemistry problem at the same time[32,39] that requires to be addressed while relying on a limited amount of CCM data. Since atmospheric transport is dominated by advection and convection (rather than diffusion) due to the dynamic movement of air masses driven by various factors such as temperature gradients, pressure differences, and wind patterns, using local weather data as features may not capture all the relevant physics of the problem. This is the largest source of uncertainty in the aCCFs and GP models as the iRF is predicted solely based on weather data at the release site, overlooking the broader weather pattern's influence, as also indicated[39]. This could partly explain why some statistical outliers are underestimated with respect to the mean of the surrogate models. The use of nonlocal data, in terms of trajectory forecasts, for a better understanding of $NO_x$–$O_3$ chemistry has also been suggested[39], which can be leveraged in the future. Since the paCCFs do not take uncertainties in standard weather forecasts into account, they can be used along with, e.g., ref. 60 for robust climate-optimised flight planning to overcome this limitation.

While nonlocal data is readily available from the underlying chemistry-climate model, there are two challenges: choosing how to define a nonlocal region in space and the computational issue of dealing with its high-dimensional nature. To make this tractable, we would need to represent the nonlocal data in a lower-dimensional space. A linear dimensional reduction method such as principal component analysis (PCA) or multidimensional scaling could be attempted, or nonlinear reduction if it proves necessary, e.g. kernel-PCA, Isomap[61], or auto-encoder approaches[62]. Assuming the model can be successfully calibrated, it will provide specific information about the importance of local versus nonlocal effects, as well as the underlying noise level. It may be able to incorporate nonlocal effects to improve prediction, but given the convective nature of the underlying process, this is perhaps unlikely. If these dimension-reduction methods do not show a low-effective dimension, then we would require a method capable of native high dimensions, such as neural networks. They can be used in regression as replacements for any other surrogate. In the context of physical predictions however, experience shows it is necessary to incorporate physical constraints[63]. The use of nonlocal data as features can also help reduce the variance of our predictions (Fig. 3b and Supplementary Figs. S2–S5), which could be beneficial in increasing the climate-impact mitigation potential for flight planning. Nevertheless, uncertainties will remain. We believe that while our model is a big step forward in advancing our predictive ability of the short-term $NO_x$–$O_3$ effect, which is the main warming component, the long-term decrease of $CH_4$ and a $CH_4$-induced decrease of $O_3$ (primary mode ozone[64]) and stratospheric water vapour can also be estimated to get the total effect of aviation $NO_x$, which can even negate the $O_3$ iRF. The prediction of the total $NO_x$ effect can then be compared to other studies that used high-fidelity simulations that are local (e.g., see Fig. 2 of ref. 24). Going further, the methods mentioned in this section can be extended to other aviation climate-forcing agents such as $H_2O$, contrails, and aerosols.

Our chained GP model provides a means of assessing individual flights with respect to the $NO_x$- induced short-term $O_3$ iRF, which is an important

step for global climate-optimised flight planning. Note that model-to-model variation in the $O_3$ RF estimates shows some variation. For example, the contribution of global anthropogenic $NO_x$ emissions to the 1850–2000 $O_3$ RF has a variation of $221 \pm 33$ mW m$^{-2}$ (mean $\pm$ standard deviation) based on the evaluation of six different climate models[65]. However, the sensitivity of $O_3$ RF due to aircraft $NO_x$ emissions with respect to altitude is fairly similar[50] and also, earlier studies on the responses of chemical systems are well agreeing (see ref. 55, their section 5.3.2.1 and Fig. 10 and discussions in refs. 36,51,66). In the future, this model could be used, along with other methods, to predict the total climate impact of local $NO_x$ released from a given flight. This a key step in making climate optimised flight planning more viable and thereby enable its potential use in e.g. a non-$CO_2$ scheme.

## Methods
### Data generation
The simulations performed to obtain the complete dataset are described before[34]. The approach used is briefly summarised here. The ECHAM/MESSy Atmospheric Chemistry (EMAC) model is a global numerical CCM that contains submodels describing tropospheric, stratospheric and mesospheric processes and interactions with oceans, land, and human influences (e.g. anthropogenic emissions) are also implemented[46,47]. The core atmospheric model is based on the 5th generation European Centre Hamburg general circulation model (ECHAM5)[67]. The atmospheric model ECHAM5 and the Modular Earth Submodel System Model (MESSy), are the fundamental blocks of EMAC which are used along with various submodels that govern, for instance, atmospheric and chemical processes.

For more realistic atmospheric conditions, the CCM simulations were run using nudging (Newtonian relaxation) of the vorticity, the wind divergence, the logarithm of the pressure field at the surface, and the temperature towards the 2014 ERA-interim reanalysis data. Four submodels played a vital role in the investigation of short-term $O_3$ iRF from global aviation $NO_x$: TREXP (Tracer Release EXperiments from Point Sources[46]), ATTILA (Atmospheric Tracer Transport In a LAgrangian model[68]), AIRTRAC[24], and RAD[69]. TREXP was employed to define the positions and duration of release of $NO_x$ pulse emissions in terms of latitude, longitude, and pressure altitude for each of the five regions, as shown in Fig. 1a. An emission amount of $5 \times 10^5$ kg of nitric oxide (NO) was injected into the atmosphere at 06:00 UTC within a 15-min time step and equally split into 50 randomly generated Lagrangian air parcel trajectories within the grid cell of each emission point using ATTILA. The procedure followed is the same as another study[24] on the dates 1 January 2014 and 1 July 2014, each with a 3-month simulation period. Within each air parcel trajectory, the AIRTRAC submodel calculated the net contribution of $NO_x$ emissions to the atmospheric composition of $O_3$ by taking into account various chemical reactions, and phenomena such as scavenging and dry deposition. The RAD submodel computes the iRF every 2 h, as opposed to 15 min, due to computational burdens, and the averaged value is extracted with a 6 h output frequency over the 3-month simulation period. It is important to note that in this study, iRF refers to the mean global radiative impact resulting from a pulse emission rather than the cumulative RF contributions since pre-industrial times and is measured relative to the climatological tropopause since the stratospheric-adjusted RF and the effective radiative forcing are not feasible in this case[31,70]. A total of 30 simulations were conducted for five regions, two seasons and three pressure levels, totalling ~105,000 CPU hours using parallel computing on the Dutch supercomputer Snellius. Thus, CCM data generation is a computationally expensive process, and the goal is to adequately reproduce these predictions using the methodology described in the following sections.

### Supervised learning
Supervised learning aims to find a relation between a target variable, $y \in \mathbb{R}$ (climate impact in terms of iRF from aviation $NO_x$ emissions) and selected input variables that are influential, $\mathbf{x} \in \mathbb{R}^m$ (e.g., meteorological and spatial parameters), based on data generated from a high-cost CCM subject to

several input climate model parameters. More concretely, we have the dataset, $\mathcal{D} = \{(\mathbf{x}_i, y_i) | i = 1, \ldots, n\} \equiv (X, \mathbf{y})$ where $X \in \mathbb{R}^{n \times m}$ is called the design matrix and $\mathbf{y} \in \mathbb{R}^n$ is the target vector, created by aggregating the $n$ cases. Given this dataset, we would like to build a low-cost probabilistic surrogate model that can make predictions for new, unobserved cases $y_*$ using Gaussian process regression (GPR). This is a Bayesian nonparametric approach that does not yield a single best-fit point estimate, but provides a probability distribution for each estimate; they essentially provide a useful way of quantifying uncertainties in the model estimates for new test inputs. This runs contrary to deterministic surrogates such as the existing aCCFs[30], which provide a point estimate for a given input, that does not capture the uncertainty associated with it.

### Feature selection
There are several potentially important features $\mathcal{V} \in \mathbb{R}^p$ that inform the climate impact of aviation $NO_x$ on $O_3$, but statistical models are constrained by the curse of dimensionality for large $p$. However, we could still pick a subset of important features $\mathcal{S} \subseteq \mathcal{V}$ by: (i) calculating the mutual information (MI) between the target and each feature $\mathcal{I}(y; v_i)$ with $v_i \in \mathcal{V}$, $\forall i = 1, \ldots, p$ and selecting variables which are characterised by relatively large scores, and, (ii) using automatic relevance determination (ARD)[37], which automatically determines the relevance of different features for GPR. Thus, we obtain, $\mathcal{S} \in \mathbb{R}^m$, with $m < p$. Additionally it is informative to look at $\mathcal{I}(v_i; v_j) \forall i \neq j$ to detect multi-collinearity, which negatively impacts regression models. MI measures relationships between any two random variables, for which a moderate number of samples are available. Intuitively, it tells us how much we learn about one random variable by knowing the value of the other random variable. The MI between two univariate random variables $X$ and $Y$ is given by

$$\mathcal{I}(X; Y) = \text{KL}(\rho(x, y) || \rho(x)\rho(y)) = \int_{\mathcal{X}} \int_{\mathcal{Y}} \rho(x, y) \log \frac{\rho(x, y)}{\rho(x)\rho(y)} \, dx \, dy,$$

where $\text{KL}(\rho(a) || \rho(b))$ represents the KL divergence between two probability density functions (pdfs), is $\geq 0$ and it is a unit of statistical distance; $\rho(x, y)$ represents the joint pdf of $X$ and $Y$ over the space $\mathcal{X}$, $\mathcal{Y}$; and $\rho(x)$, $\rho(y)$ represent the marginal pdf of $X$ and $Y$, respectively.

### Using GPR to fit the data
A regression model with independently and identically distributed noise is a common assumption used in regression modelling, which treats observations as independent and providing the same kind of information, thereby simplifying parameter estimation and statistical inference. The regression model with Gaussian noise $\varepsilon$ can then be defined as

$$y = f(\mathbf{x}) + \varepsilon, \ \varepsilon \sim \mathcal{N}(0, \sigma_n^2), \tag{1}$$

where $\sigma_n^2$ represents data noise variance. Here, data is assumed to be homoscedastic, that is, the noise variance is constant. In linear regression, we find an optimal linear model $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w}$, where $\mathbf{w}$ represents the vector of unknown parameters to be determined. This is a parametric approach[30], whereby the complexity of the model is bounded by a fixed number of parameters. On the other hand, no assumptions are made about the characteristics of the underlying function in the case of GPR, allowing it to freely adapt to the complexity of the data and capture patterns that might be missed by parametric models. This way, the amount of information that can be captured about $\mathcal{D}$ grows as the amount of data available grows, making such models very flexible. In GPR, the key assumption is that the underlying function $f$ that generates the data is a Gaussian process (GP). A GP is a collection of random variables, any finite number of which have a joint Gaussian distribution and is used to model the relationship between the chosen features and the target variable. The standard GPR model is well documented[41].

While homoscedasticity is a common assumption used in regression modelling for computational and technical convenience, many real-world

phenomena are characterised by heteroscedasticity; that is, the variance of the data is some (simple or complex) function of certain variables. Thus we choose a heteroscedastic or chained GPR model with Gaussian noise $\varepsilon \sim \mathcal{N}(0, \sigma_n^2(\mathbf{x}))$,

$$y = f(\mathbf{x}) + \varepsilon(\mathbf{x}). \qquad (2)$$

To solve this, the noise variance is modelled using a log-GP prior to ensure that only positive values are possible. That is

$$\rho(y|f(\mathbf{x}), g(\mathbf{x})) = \mathcal{N}\big(f(\mathbf{x}), e^{g(\mathbf{x})}\big), \quad \text{where,}$$

$$\rho(\mathbf{f}) = \mathcal{N}\big(\boldsymbol{\mu}_f, k_f(\mathbf{x}, \mathbf{x}')\big), \quad \text{and}$$

$$\rho(\mathbf{g}) = \mathcal{N}\big(\boldsymbol{\mu}_g, k_g(\mathbf{x}, \mathbf{x}')\big).$$

Thus two GPs, $\mathbf{f}$ and $\mathbf{g}$, are used to learn both the mean and the variance of the Gaussian likelihood, respectively. These GPs have their individual means ($\boldsymbol{\mu}_f, \boldsymbol{\mu}_g$) and covariances ($k_f, k_g$), and are assumed to be independent. The covariances, which dictate the smoothness of the functions, are computed using the square exponential kernel[41]. The posterior distribution, however, is not analytically tractable, and we use chained GPs[40], which uses sparse variational inference to approximate the posterior with the aid of so-called inducing point methods. These methods introduce $q < n$ pseudo-points at locations $Z = \{\mathbf{z}_i\}_{i=1}^q$ to approximate the full covariance matrix. The locations are inferred by applying $K$-means clustering[71] to input data, $\mathbf{x}$ and using the cluster centres. The corresponding function values are $\mathbf{u}_f = f(Z)$ and $\mathbf{u}_g = g(Z)$. The posterior distribution obtained using Bayes' theorem is approximated as

$$\begin{aligned}\rho(\mathbf{f}, \mathbf{g}, \mathbf{u}_f, \mathbf{u}_g|\mathbf{y}) &\approx \rho(\mathbf{f}|\mathbf{u}_f)\rho(\mathbf{g}|\mathbf{u}_g)\rho(\mathbf{u}_f|\mathbf{y})\rho(\mathbf{u}_g|\mathbf{y}) \\ &\approx \rho(\mathbf{f}|\mathbf{u}_f)\rho(\mathbf{g}|\mathbf{u}_g)\pi(\mathbf{u}_f)\pi(\mathbf{u}_g),\end{aligned} \qquad (3)$$

where it has been assumed that the latent functions factorise for the tractability of the problem. In variational inference, we obtain a proxy for $\rho(\mathbf{u}_f|\mathbf{y})$ and $\rho(\mathbf{u}_g|\mathbf{y})$ by seeking an appropriate distribution $\pi(\mathbf{u}_f)$ and $\pi(\mathbf{u}_g)$, respectively, from a family of distributions $\mathcal{F}$, such that $\mathrm{KL}\big(\pi(\mathbf{f}, \mathbf{g}, \mathbf{u}_f, \mathbf{u}_g)||\rho(\mathbf{f}, \mathbf{g}, \mathbf{u}_f, \mathbf{u}_g|\mathbf{y})\big)$ is minimised. This requires the calculation of the likelihood $\rho(\mathbf{y})$, which is generally not tractable, leading us to find a lower bound for $\log \rho(\mathbf{y})$.

$$\begin{aligned}\log \rho(\mathbf{y}) &= \log \int \rho(\mathbf{y}|\mathbf{f}, \mathbf{g})\rho(\mathbf{f}|\mathbf{u}_f)\rho(\mathbf{g}|\mathbf{u}_g)\rho(\mathbf{u}_f)\rho(\mathbf{u}_g)d\mathbf{f}\,d\mathbf{g}\,d\mathbf{u}_f\,d\mathbf{u}_g \\ &\geq \int \pi(\mathbf{f})\pi(\mathbf{g})\log \rho(\mathbf{y}|\mathbf{f}, \mathbf{g})\,d\mathbf{f}\,d\mathbf{g} - \mathrm{KL}(\pi(\mathbf{u}_f)||\rho(\mathbf{u}_f)) \\ &\quad - \mathrm{KL}(\pi(\mathbf{u}_g)||\rho(\mathbf{u}_g)).\end{aligned} \qquad (4)$$

Following the study[40] and modelling $\pi(\mathbf{u}_f) = \mathcal{N}(\mathbf{u}_f|\mathbf{m}_f, \Sigma_f)$ and $\pi(\mathbf{u}_g) = \mathcal{N}(\mathbf{u}_g|\mathbf{m}_g, \Sigma_g)$ allows us to compute $\pi(\mathbf{f}) = \int \rho(\mathbf{f}|\mathbf{u}_f)\pi(\mathbf{u}_f)d\mathbf{u}_f$ and $\pi(\mathbf{g}) = \int \rho(\mathbf{g}|\mathbf{u}_g)\pi(\mathbf{u}_g)d\mathbf{u}_g$. The likelihood also factorises, i.e., $\rho(\mathbf{y}|\mathbf{f}, \mathbf{g}) = \prod_{i=1}^n \rho(y_i|f_i, g_i)$, allowing us to apply stochastic variational inference to the integral in Eq. (4),

$$\begin{aligned}\log \rho(\mathbf{y}) &\geq \sum_{i=1}^n \int \pi(f_i)\pi(g_i)\log \rho(y_i|f_i, g_i)\,df_i\,dg_i - \mathrm{KL}(\pi(\mathbf{u}_f)||\rho(\mathbf{u}_f)) \\ &\quad - \mathrm{KL}(\pi(\mathbf{u}_g)||\rho(\mathbf{u}_g)).\end{aligned}$$

The above integral can be solved analytically since the likelihood is Gaussian. Note that the variational parameters $\mathbf{m}_f, \Sigma_f, \mathbf{m}_g, \Sigma_g$ are learnt through maximising the bound. After optimisation, the posterior predictive distribution is calculated as

$$\rho(y_{i*}|y_i, \mathbf{x}_i, \mathbf{x}_{i*}) = \int \rho(y_{i*}|f_{i*}, g_{i*})\pi(f_{i*})\pi(g_{i*})\,df_{i*}\,dg_{i*}, \qquad (5)$$

where each prediction point is treated independently for the data pair $\{(\mathbf{x}_{i*}, y_{i*})\}_{i=1}^{n_*}$. This is the equation used by the chained GP model to predict climate impact for test data and arbitrary setups, as we demonstrated in our flight calculations. While inferring GPs (without a sparse approximation) have a computational cost of $\mathcal{O}(n^3)$, chained GPs with a sparse approximation have a computational cost of $\mathcal{O}(nq^2)$ for $q$ inducing points that are used to parameterise the covariance matrix. The inducing points are picked as a subset of the training data $X$, thus $q < n$, which makes it more computationally efficient than standard GPs, especially while dealing with large datasets. Additionally, chained GPs permit a nonlinear combination of any number of GPs, even with models with a non-Gaussian likelihood. The chained GP methodology is implemented using a machine-learning library called GPFlow (version 2.6.3[72]).

## Implementation

The methodology discussed herein has been implemented in Matlab and Python 3 using NumPy (version 1.22.3[73]), Pandas (version 1.4.2[74]), Matplotlib (version 3.5.2[75]), Scikit learn (version 1.2.1[76]), and GPFlow (version 2.6.3[72]) libraries. An academic license was used for Matlab, but Python 3 and the associated libraries are freely available.

## Data availability

The base data used to train the Gaussian process models is published and available in netcdf format[77]. The data plotted in Figs. 1a, b, and 5b are available on Zenodo[78].

## Code availability

The codes used to train the Gaussian process models, and visualise their results are available from the corresponding author upon request. The associated versions of the code related to the methodology are mentioned under "Implementation".

## References

1. ICAO (International Civil Aviation Organization). *Annual Report 2019: the World of Air Transport in 2019* (International Civil Aviation Organization, accessed 13 November 2023); https://www.icao.int/annual-report-2019/Pages/the-world-of-air-transport-in-2019.aspx.
2. Airbus. *Global Market Forecast 2018–2037* (Airbus, 2018).
3. Boeing. *Commercial Market Outlook 2019–2038* (Boeing, 2019).
4. Dube, K. Emerging from the COVID-19 pandemic: aviation recovery, challenges and opportunities. *Aerospace* **10**, 19 (2022).
5. Lee, D. S. et al. The contribution of global aviation to anthropogenic climate forcing for 2000 to 2018. *Atmos. Environ.* **244**, 117834 (2021).
6. Grewe, V. et al. Evaluating the climate impact of aviation emission scenarios towards the Paris Agreement including COVID-19 effects. *Nat. Commun.* **12**, 3841 (2021).
7. Kharina, A. & Rutherford, D. *Fuel Efficiency Trends for New Commercial Jet Aircraft: 1960 to 2014*. White paper of the International Council on Clean Transportation (2015).
8. Brazzola, N., Patt, A. & Wohland, J. Definitions and implications of climate-neutral aviation. *Nat. Clim. Change* **12**, 761–767 (2022).
9. Lee, D. S. *International Aviation and the Paris Agreement Temperature Goals* https://assets.publishing.service.gov.uk/media/5d19c520e5274a08e337b235/international-aviation-paris-agreement.pdf (2018).
10. Schumann, U. Formation, properties and climatic effects of contrails. *C. R. Physique* **6**, 549–565 (2005).

11. Avila, D., Sherry, L. & Thompson, T. Reducing global warming by airline contrail avoidance: a case study of annual benefits for the contiguous United States. *Transp. Res. Interdiscip. Perspect.* **2**, 100033 (2019).

12. Gauss, M., Isaksen, I. S. A., Lee, D. S. & Søvde, O. A. Impact of aircraft $NO_x$ emissions emissions on the atmosphere—tradeoffs to reduce the impact. *Atmos. Chem. Phys.* **6**, 1529–1548 (2006).

13. Gilmore, C. K., Barrett, S. R. H., Koo, J. & Wang, Q. Temporal and spatial variability in the aviation $NO_x$-related $O_3$ impact. *Environ. Res. Lett.* **8**, 034027 (2013).

14. Köhler, M. O., Rädel, G., Shine, K. P., Rogers, H. L. & Pyle, J. A. Latitudinal variation of the effect of aviation $NO_x$ emissions on atmospheric ozone and methane and related climate metrics. *Atmos. Environ.* **64**, 1–9 (2013).

15. Freeman, S., Lee, D. S., Lim, L. L., Skowron, A. & De León, R. R. Trading off aircraft fuel burn and $NO_x$ emissions for optimal climate policy. *Environ. Sci. Technol.* **52**, 2498–2505 (2018).

16. Terrenoire, E. et al. Impact of present and future aircraft $NO_x$ and aerosol emissions on atmospheric composition and associated direct radiative forcing of climate. *Atmos. Chem. Phys.* **22**, 11987–12023 (2022).

17. Skowron, A., Lee, D. S., De León, R. R., Lim, L. L. & Owen, B. Greater fuel efficiency is potentially preferable to reducing $NO_x$ emissions for aviation's climate impacts. *Nat. Commun.* **12**, 564 (2021).

18. Morris, G. A., Rosenfield, J. E., Schoeberl, M. R. & Jackman, C. H. Potential impact of subsonic and supersonic aircraft exhaust on water vapor in the lower stratosphere assessed via a trajectory model. *J. Geophys. Res. Atmos.* **108**, 4443 (2003).

19. Wilcox, L., Shine, K. & Hoskins, B. Radiative forcing due to aviation water vapour emissions. *Atmos. Environ.* **63**, 1–13 (2012).

20. Wang, H. et al. Global tropospheric ozone trends, attributions, and radiative impacts in 1995–2017: an Integrated Analysis using Aircraft (IAGOS) observations, ozonesonde, and multi-decadal chemical model simulations. *Atmos. Chem. Phys.* **22**, 13753–13782 (2022).

21. Eastham, S. D., Chossière, G. P., Speth, R. L., Jacob, D. J. & Barrett, S. R. H. Global impacts of aviation on air quality evaluated at high resolution. *Atmos. Chem. Phys.* **24**, 2687–2703 (2024).

22. Niklaß, M. et al. Potential to reduce the climate impact of aviation by climate restricted airspaces. *Transport Policy* **83**, 102–110 (2019).

23. Matthes, S. et al. *Climate Optimized Air Transport* 727–746 (Springer, Berlin, Heidelberg, 2012).

24. Grewe, V. et al. Aircraft routing with minimal climate impact: the REACT4c climate cost function modelling approach (v1.0). *Geosci. Model Dev.* **7**, 175–201 (2014).

25. Frömming, C. et al. Aviation-induced radiative forcing and surface temperature change in dependency of the emission altitude. *J. Geophys. Res. Atmos.* **117**, D19104 (2012).

26. Dahlmann, K. et al. Climate-compatible air transport system—climate impact mitigation potential for actual and future aircraft. *Aerospace* **3**, 38 (2016).

27. Matthes, S. et al. Mitigation of non-$CO_2$ aviation's climate impact by changing cruise altitudes. *Aerospace* **8**, 36 (2021).

28. Bauer, P., Thorpe, A. & Brunet, G. The quiet revolution of numerical weather prediction. *Nature* **525**, 47–55 (2015).

29. Cho, D., Yoo, C., Im, J. & Cha, D. Comparative assessment of various machine learning-based bias correction methods for numerical weather prediction model forecasts of extreme air temperatures in urban areas. *Earth Space Sci.* **7**, e2019EA000740 (2020).

30. van Manen, J. & Grewe, V. Algorithmic climate change functions for the use in eco-efficient flight planning. *Transp. Res. Part D: Transport Environ.* **67**, 388–405 (2019).

31. Rao, P. et al. Case study for testing the validity of $NO_x$–ozone algorithmic climate change functions for optimising flight trajectories. *Aerospace* **9**, 231 (2022).

32. Frömming, C. et al. Influence of weather situation on non-$CO_2$ aviation climate effects: the react4c climate change functions. *Atmos. Chem. Phys.* **21**, 9151–9172 (2021).

33. Maruhashi, J., Grewe, V., Frömming, C., Jöckel, P. & Dedoussi, I. C. Transport patterns of global aviation $NO_x$ and their short-term $O_3$ radiative forcing—a machine learning approach. *Atmos. Chem. Phys.* **22**, 14253–14282 (2022).

34. Maruhashi, J., Mertens, M., Grewe, V. & Dedoussi, I. C. A multi-method assessment of the regional sensitivities between flight altitude and short-term $O_3$ climate warming from aircraft $NO_x$ emissions. *Environ. Res. Lett.* **19**, 054007 (2024).

35. Grewe, V. & Stenke, A. AirClim: an efficient tool for climate evaluation of aircraft technology. *Atmos. Chem. Phys.* **8**, 4621–4639 (2008).

36. Stevenson, D. S. et al. Radiative forcing from aircraft $NO_x$ emissions: Mechanisms and seasonal dependence. *J. Geophys. Res. Atmos.* **109**, D17307 (2004).

37. Neal, R. M. *Bayesian Learning for Neural Networks* (Springer, New York, 1996).

38. Gray, L. J. et al. Solar influences on climate. *Rev. Geophys.* **48**, RG4001 (2010).

39. Rosanka, S., Frömming, C. & Grewe, V. The impact of weather patterns and related transport processes on aviation's contribution to ozone and methane concentrations from $NO_x$ emissions. *Atmos. Chem. Phys.* **20**, 12347–12361 (2020).

40. Saul, A. D., Hensman, J., Vehtari, A. & Lawrence, N. D. Chained Gaussian processes. In *Proc. 19th International Conference on Artificial Intelligence and Statistics*, (eds Gretton, A. & Robert, C. C.) Vol. 51 of *Proc. Machine Learning Research* 1431–1440 (PMLR, 2016).

41. Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning* (The MIT Press, 2006).

42. Yamashita, H. et al. Newly developed aircraft routing options for air traffic simulation in the chemistry—climate model EMAC 2.53: AirTraf 2.0. *Geosci. Model Dev.* **13**, 4869–4890 (2020).

43. Dahlmann, K., Grewe, V., Matthes, S. & Yamashita, H. Climate assessment of single flights: deduction of route specific equivalent $CO_2$ emissions. *Int. J. Sustain. Transp.* **17**, 29–40 (2023).

44. Kullback, S. & Leibler, R. A. On information and sufficiency. *Ann. Math. Stat.* **22**, 79–86 (1951).

45. Eurocontrol. *Aviation Data for Research* (Eurocontrol, 2023).

46. Jöckel, P. et al. Development cycle 2 of the modular earth submodel system (messy2). *Geosci. Model Dev.* **3**, 717–752 (2010).

47. Jöckel, P. et al. Earth system chemistry integrated modelling (ESCiMo) with the modular earth submodel system (MESSy) version 2.51. *Geosci. Model Dev.* **9**, 1153–1200 (2016).

48. Jöckel, P. et al. The atmospheric chemistry general circulation model ECHAM5/MESSy1: consistent simulation of ozone from the surface to the mesosphere. *Atmos. Chem. Phys.* **6**, 5067–5104 (2006).

49. Emmons, L. K. et al. Data composites of airborne observations of tropospheric ozone and its precursors. *J. Geophys. Res.: Atmos.* **105**, 20497–20538 (2000).

50. Søvde, O. A. et al. Aircraft emission mitigation by changing route altitude: a multi-model estimate of aircraft $NO_x$ emission impact on $O_3$ photochemistry. *Atmos. Environ.* **95**, 468–479 (2014).

51. Köhler, M. O. et al. Impact of perturbations to nitrogen oxide emissions from global aviation. *J. Geophys. Res.* **113**, D19303 (2008).

52. Mertens, M., Grewe, V., Rieger, V. S. & Jöckel, P. Revisiting the contribution of land transport and shipping emissions to tropospheric ozone. *Atmos. Chem. Phys.* **18**, 5567–5588 (2018).

53. Thunis, P. et al. Source apportionment to support air quality planning: strengths and weaknesses of existing approaches. *Environ. Int.* **130**, 104825 (2019).

54. Dedoussi, I. C., Eastham, S. D., Monier, E. & Barrett, S. R. H. Premature mortality related to United States cross-state air pollution. *Nature* **578**, 261–265 (2020).

55. Lee, D. S. et al. Transport impacts on atmosphere and climate: aviation. *Atmos. Environ.* **44**, 4678–4734 (2010).

56. Holmes, C. D., Tang, Q. & Prather, M. J. Uncertainties in climate assessment for the case of aviation NO. *Proc. Natl Acad. Sci. USA* **108**, 10997–11002 (2011).

57. Myhre, G. et al. Intercomparison of radiative forcing calculations of stratospheric water vapour and contrails. *Meteorol. Z.* **18**, 585–596 (2009).

58. Cameron, M. A., Jacobson, M. Z., Naiman, A. D. & Lele, S. K. Effects of plume-scale versus grid-scale treatment of aircraft exhaust photochemistry. *Geophys. Res. Lett.* **40**, 5815–5820 (2013).

59. Fritz, T. M., Eastham, S. D., Speth, R. L. & Barrett, S. R. H. The role of plume-scale processes in long-term impacts of aircraft emissions. *Atmos. Chem. Phys.* **20**, 5697–5727 (2020).

60. Simorgh, A. et al. Robust 4d climate-optimal flight planning in structured airspace using parallelized simulation on GPUS: Roost v1.0. *Geosci. Model Dev.* **16**, 3723–3748 (2023).

61. Tenenbaum, J. B. A global geometric framework for nonlinear dimensionality reduction. *Science* **290**, 2319–2323 (2000).

62. Wang, Y., Yao, H. & Zhao, S. Auto-encoder based dimensionality reduction. *Neurocomputing* **184**, 232–242 (2016).

63. Zhou, X.-H., Han, J. & Xiao, H. Learning nonlocal constitutive models with neural networks. *Comput. Methods Appl. Mech. Eng.* **384**, 113927 (2021).

64. Wild, O., Prather, M. J. & Akimoto, H. Indirect long-term global radiative cooling from $NO_x$ emissions. *Geophys. Res. Lett.* **28**, 1719–1722 (2001).

65. Stevenson, D. S. et al. Tropospheric ozone changes, radiative forcing and attribution to emissions in the atmospheric chemistry and climate model intercomparison project (accmip). *Atmos. Chem. Phys.* **13**, 3063–3085 (2013).

66. Grewe, V., Dameris, M., Fichter, C. & Lee, D. S. Impact of aircraft $NO_x$ emissions. Part 2: effects of lowering the flight altitude. *Meteorol. Z.* **11**, 197–205 (2002).

67. Roeckner, E. et al. Sensitivity of simulated climate to horizontal and vertical resolution in the ECHAM5 atmosphere model. *J. Clim.* **19**, 3771–3791 (2006).

68. Brinkop, S. & Jöckel, P. Attila 4.0: Lagrangian advective and convective transport of passive tracers within the echam5/messy (2.53.0) chemistry-climate model. *Geosci. Model Dev.* **12**, 1991–2008 (2019).

69. Dietmüller, S. et al. A new radiation infrastructure for the modular earth submodel system (MESSy, based on version 2.51). *Geosci. Model Dev.* **9**, 2209–2222 (2016).

70. Stuber, N., Sausen, R. & Ponater, M. Stratosphere adjusted radiative forcing calculations in a comprehensive climate model. *Theor. Appl. Climatol.* **68**, 125–135 (2001).

71. Lloyd, S. Least squares quantization in pcm. *IEEE Trans. Inf. Theory* **28**, 129–137 (1982).

72. Matthews, A. Gd. G. et al. GPflow: a Gaussian process library using TensorFlow. *J. Mach. Learn. Res.* **18**, 1–6 (2017).

73. Harris, C. R. et al. Array programming with numpy. *Nature* **585**, 357–362 (2020).

74. McKinney, W. Data structures for statistical computing in Python. In *Proc. 9th Python in Science Conference* 56–61 (eds van der Walt, S. J. & Jarrod Millman, K.) (*Proceedings of the Python in Science Conference* (SciPy), 2010).

75. Hunter, J. D. Matplotlib: a 2d graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

76. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

77. Maruhashi, J., Mertens, M., Grewe, V. & Dedoussi, I. C. *Supplementary Dataset for "a Multi-method Assessment of the Regional Sensitivities Between Flight Altitude and Short-term $O_3$ Climate Warming from Aircraft $NO_x$ Emissions"* https://doi.org/10.4121/56327667-69F1-4340-BE45-9F9A6BD80584.V1 (2024).

78. Rao, P. et al. *Data from: The Ozone Radiative Forcing of Nitrogen Oxide Emissions from Aviation can be Estimated Using a Probabilistic Approach* https://doi.org/10.5281/ZENODO.10546876 (2024).

## Author contributions
P.R. and R.D. conceptualised this study. J.M. prepared and provided the data required for the Gaussian process models. P.R. and D.S. developed the computer code. P.R. trained and tested the Gaussian process models. R.D. and V.G. provided the necessary supervision. R.D., V.G., C.F., and I.D. gave important advice for writing the manuscript and visualising many results. P.R. wrote the manuscript. All authors were involved in discussions and reviewing the manuscript.

## Competing interests
The authors declare no competing interests.

## Additional information
**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s43247-024-01691-2.

**Correspondence** and requests for materials should be addressed to Pratik Rao.

**Peer review information** *Communications Earth & Environment* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editors: Mengze Li, Heike Langenberg. A peer review file is available.

**Reprints and permissions information** is available at http://www.nature.com/reprints

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.