



Development of a Random Forest model for forecasting allergenic pollen in North America

Fiona Lo^{a,*}, Cecilia M. Bitz^b, Jeremy J. Hess^{a,c,d}

^a Department of Environmental and Occupational Health Sciences, School of Public Health, University of Washington, United States of America

^b Department of Atmospheric Sciences, College of the Environment, University of Washington, United States of America

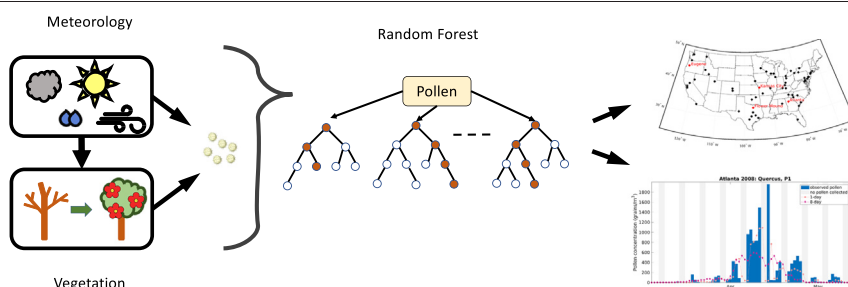
^c Department of Emergency Medicine, School of Medicine, University of Washington, United States of America

^d Department of Global Health, Schools of Medicine and Public Health, University of Washington, United States of America

HIGHLIGHTS

- Random Forest model is developed to predict pollen concentrations from environmental data.
- The forecast is skillful for 1–14 days for four pollen types in diverse eco-climate zones.
- The model produces skillful forecast in regions between stations with pollen observations used for training.
- Data augmentation greatly improves forecast skill.
- Skill of the forecast is limited by the spatiotemporal resolution of pollen observations.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:

Received 17 November 2020

Received in revised form 5 January 2021

Accepted 29 January 2021

Available online 4 February 2021

Editor: Pavlos Kassomenos

Keywords:

Allergy
Weather
Machine learning
Prediction
Hay fever

ABSTRACT

Pollen allergies have negative impacts on health. Information about airborne pollen concentration can improve symptom management by guiding choices affecting timing of medicines and pollen exposure. Observations provide accurate pollen concentrations at point locations. However, in the contiguous United States and southern Canada (CUSCC), observations are sparse, and sampling is often seasonal, intermittent or both. Modeling pollen concentration can fill in the gaps with estimates where direct observations are unavailable and also provide much-needed forecasts. The goal of this study is to develop and evaluate statistical models that predict daily pollen concentrations using a machine learning Random Forest algorithm. To evaluate our methods, we made retrospective forecasts of four pollen types (*Quercus*, *Cupressaceae*, *Ambrosia* and *Poaceae*), each in one of four CUSCC locations. Meteorological and vegetation conditions were input to the models at city and regional scales. A data augmentation technique was investigated and found to improve model skill. Models were also developed to forecast pollen in locations where there are no observations. Forecast skill in these models were found to be greater than in previous models. Nevertheless, the skill is limited by the spatiotemporal resolution of the pollen observations.

© 2021 Elsevier B.V. All rights reserved.

1. Introduction

Pollen is a common cause of seasonal allergic reactions, and allergic diseases impose a large and growing burden of disease. An estimate of up to 30% of the population in the United States (U.S.) suffer from

* Corresponding author at: Hans Rosling Center for Population Health, 3980 15th Ave NE, Seattle, WA 98195-1618, United States of America.
E-mail address: fionalo@uw.edu (F. Lo).

allergic rhinitis (Wheatley and Togias, 2015) and the prevalence is thought to be increasing with increasing airborne allergens and climate change (D'Amato et al., 2016). Allergic diseases impose a significant financial burden with direct costs of treatments and medications, and indirect costs from lower workplace productivity, adverse school performance, and reduced quality of life (Meltzer and Bukstein, 2011, Lamb et al., 2006, Marcotte, 2015, Nathan, 2007).

Accurate local pollen information, including forecasts, are useful for diagnosis and reduction of allergic symptoms. This information allows patients to self-manage symptoms and reduce the burden of disease (Medek et al., 2019). Because airborne pollen concentrations vary greatly in space and time, it is necessary to have many pollen observing sites. However, sampling pollen data on a large spatiotemporal scale is resource-intensive: identification of pollen requires expertise and counting pollen is time-consuming. In the U.S., the burden and cost of sampling pollen is borne largely by allergy clinics, but the service is not available at every clinic. In addition, the data is owned by individual clinics and not publicly available, making it difficult to obtain pollen data on an extensive spatial scale.

Models can bridge the gap of missing pollen data by simulating pollen concentrations and can also forecast pollen concentrations. There are two general approaches to pollen modeling: observation-based pollen models which use statistics to describe the relationships between pollen metrics and environmental variables, and process-based phenological models which use equations to represent known plant physiological responses to environmental variables. In addition to standard regression techniques, several different observation-based approaches have been used to model pollen concentrations, including neural networks, support vector machine models and machine learning Random Forest models (Valencia et al., 2019, Astray et al., 2016, Nowosad, 2016, Zewdie et al., 2019). We chose the Random Forest approach because it has been shown to be able to skillfully predict pollen (Nowosad et al., 2018; Liu et al., 2017; Zewdie et al., 2019), is complex enough to fit the data yet robust enough to perform well with new data, is computationally efficient and easy to interpret compared with other approaches. There are several additional reasons that machine learning, in particular Random Forest models, have an advantage over conventional regression methods. First, conventional models are designed by humans and are limited by our ability to fully comprehend the processes that influence pollen. Machine learning models are optimized to fit data without making assumptions and well-suited for predicting a complex nonlinear system such as pollen. Second, Random Forest models are able to handle many collinear predictors without overfitting (Breiman, 2001). Model developers often struggle to determine which variables are important and need to be included into a model. The Random Forest allows the numerous environmental variables that influence airborne pollen to be included in the model.

Pollen research has found many processes and correlations that link environmental variables to airborne pollen that should be included into a pollen model. The time scales of these relationships are important: meteorological parameters have been correlated with pollen on time scales ranging from days to years (Khwarahm et al., 2014; Ritenberga et al., 2018; Tseng et al., 2018). In general, time scales of a few weeks and longer affect the development of the plant and its readiness to flower, while shorter time scales of up to a few weeks affect the release and suspension of pollen in the atmosphere. Variables that are positively correlated with pollen concentration on a short time scale are temperature, solar radiation, sunshine duration, and wind speed, while variables that are negatively correlated are number of rain days and snow depth (i.e., Bruffaerts et al., 2018; Khwarahm et al., 2014; Ritenberga et al., 2018). Interestingly, humidity and rain have been found to have both positive and negative correlations with pollen concentration. These divergent results are possible because wet deposition can decrease the airborne pollen concentration (Vázquez et al., 2003); whereas, absorption of water on humid days can cause pollen to swell and fragment, leading to an increase in the concentration (Miguel

et al., 2006; Rathnayake et al., 2017). On a longer timescale, temperature-based metrics, such as growing degree days and chilling-day requirements are routinely used in process-based models to determine the phenological development of a plant and estimate the time of pollen release (e.g., Chuine and Belmonte, 2004; Caffarra et al., 2011; García-Mozo et al., 2008).

Pollen concentrations themselves can be a predictor of future pollen concentrations and useful to include in models. Pollen has a temporal autocorrelation on the order of days, so yesterday's pollen concentration is a good estimate for today's (Nowosad et al., 2015). Annual pollen production can also depend on the previous year's pollen production: some plants produce large quantities of pollen only once every two years (Ritenberga et al., 2018). Although including pollen as a predictor can improve a model's performance, pollen data is not always available. Therefore, developing a model that does not require pollen as a predictor is important for forecasting in locations without pollen observations.

Random Forest models have successfully predicted pollen concentrations using either, or both, pollen concentrations and meteorology as predictors. High levels of pollen in Poland have been forecasted using only pollen concentrations from other locations in Poland (Nowosad, 2016). Estimates of *Ambrosia* pollen in Tulsa, OK in the U.S. have been predicted using reanalysis meteorological and weather radar data from the same day (Zewdie et al., 2019; Liu et al., 2017).

The goal of this study is to build on previous work and develop Random Forest models to predict daily pollen concentrations for forecasts in the contiguous U.S. and southern Canada (CUSC). We develop and evaluate different versions of Random Forest models to predict daily pollen concentrations for four pollen types: *Quercus* (oak), *Cupressaceae* (cypress, cedar, juniper), *Ambrosia* (ragweed) and *Poaceae* (grass) over four locations in CUSC. The model uses satellite-based meteorological and vegetation variables, as well as pollen observations, as predictors to forecast pollen concentrations 1–14 days in advance. Multiple versions of the model, with and without pollen concentration as a predictor, and inclusion of regional information are evaluated to find the best performing model.

2. Materials and methods

2.1. Random Forest model

Random Forest is a supervised machine learning algorithm that utilizes an ensemble of decision trees to build a predictive model. We used the MATLAB (version 2019a) treebagger function in regression mode to develop our Random Forest model.

We used the Random Forest model to forecast daily mean pollen concentrations for each pollen type. We refer to the day in which a forecast is “issued” as day 0, forecasts in the future are positive days and days prior to the day of forecast issue are negative days. We evaluated forecasts with lead times between 1 and 14 days. The forecasts in this study are hindcasts, also known as retrospective forecasts, so they can be evaluated against observations. Nonetheless, the hindcasts are true forecasts in that data beyond day 0 is not used to make the forecast.

Following machine learning vocabulary, we will refer to predictor inputs into the Random Forest model as features and the output to be predicted as labels. The feature and label data were portioned into separate sets to train and test the model. To ensure that temporal characteristics of the seasonal cycle of pollen concentration are captured in the model, the data were partitioned by station-years. We used 5-fold cross-validation technique: the station-years were split into five groups, within each group 3–4 station-years were withheld as test data, and the model was trained on data from 12 to 13 station-years. The training and test data sets were randomly generated by MATLAB's *cvpartition* function. We developed a unique Random Forest model for each forecast lead time.

We used the Mean Absolute Error (MAE) (Eq. (1)) to evaluate the models' performance because it is a simple and easy to understand

metric and errors contribute equally whether they are large or small. MAE is also used to guide decisions on hyperparameters for the Random Forest model. The mean MAE of the 5-fold cross-validations were used for error analyses.

$$MAE = \frac{1}{n} \sum_i^n |\text{predicted}_i - \text{observed}_i| \quad (1)$$

Hyperparameters are parameters of the Random Forest model that can be tuned to optimize performance. We evaluated two hyperparameters: the number of decision trees and the number of variables randomly selected at each decision split. We tuned the number-of-trees hyperparameter by evaluating the performance of the model across a wide range and found that performance improvements were marginal for more than 100 trees (not shown). With computation efficiency in mind, we set the number of trees to 100. Similarly, the increasing number of variables to select at each decision split improved the model, but leveled off at 150, so we set this hyperparameter at 150.

2.2. Pollen data

Daily pollen concentrations (grains/m³) were obtained from National Allergy Bureau (NAB) American Academy of Allergy Asthma and Immunology's Aeroallergen Network. NAB stations are required to collect pollen samples at a minimum of 3 days per week and pollen counts are collected with a Burkard volumetric air sampler or a Rotorod rotation impaction sampler. For this study, we assume that pollen counting and reporting occurs daily and is immediately available for input into a model. However, in reality, many NAB stations use the standard 7-day Burkard sampler in which the pollen is counted only once a week. We note that daily counting is achievable as some NAB stations publish up-to-date daily pollen counts on their websites (e.g., [Atlanta Allergy & Asthma](#) website). Pollen concentrations are categorized into 43 pollen types. Based on their abundance and allergic potential (Lo et al., 2019), we analyzed the grass family (*Poaceae*), the ragweed genus (*Ambrosia*) and two tree pollen types: the oak genus (*Quercus*) and the cedar, cypress and juniper family (*Cupressaceae*). Sixty-two stations are located in the contiguous United States and one station

(London, ON) in Canada. For this study, we selected four locations based on measurement quality (i.e., least missing values), abundance of the pollen type at that location and diversity across the ecoclimatic zones in CUSSC. *Quercus* pollen was studied in Atlanta, GA; *Cupressaceae* pollen at Flower Mound, TX; *Poaceae* pollen at Eugene, OR and *Ambrosia* pollen at Kansas City, MO (Fig. 1).

Pollen concentration, prior to day 0, was included as a feature when available because the recent state of pollen is a good indicator of the near future state of pollen (Nowosad et al., 2015). Unfortunately, many stations do not sample pollen on weekends or over the winter, so there are many missing days of data. Random Forest models do not handle missing data, so if pollen is not collected for a given day, that day was not included into training the model. Pollen features that were included in the model were either (1) the previous day's pollen concentration or (2) the previous 7-day mean of available pollen concentrations.

Daily pollen concentrations can range from very small numbers (zero outside the pollen season) to very high numbers during the peak of the pollen season (~10⁴ grains/m³). One of the most important times to forecast is the beginning of the pollen season because a small pollen concentration can trigger allergy symptoms and allergy sufferers need to know when to start taking medications (de Weger et al., 2013). The beginning of the pollen season can vary by ~1–2 weeks (Lo et al., 2019), so the variability of pollen concentration around that time is very high. Because Random Forest uses the mean of decision trees to predict a daily pollen concentration, using absolute pollen concentration values can lead to large values dominating the prediction. To overcome this problem, we transformed the data using the 4th root of the pollen concentration, so the importance of large values is minimized, and small values are given more weight. The 4th root of the pollen concentrations was used in the models as a feature, and it was also predicted as a label. Daily pollen concentrations were then computed by taking the 4th exponent of the model outputs. Verifications were made by comparing predicted pollen concentrations with observed.

2.3. Meteorological data

We obtained temperature, precipitation, winds and humidity data from NASA's Modern-Era Retrospective analysis for Research and

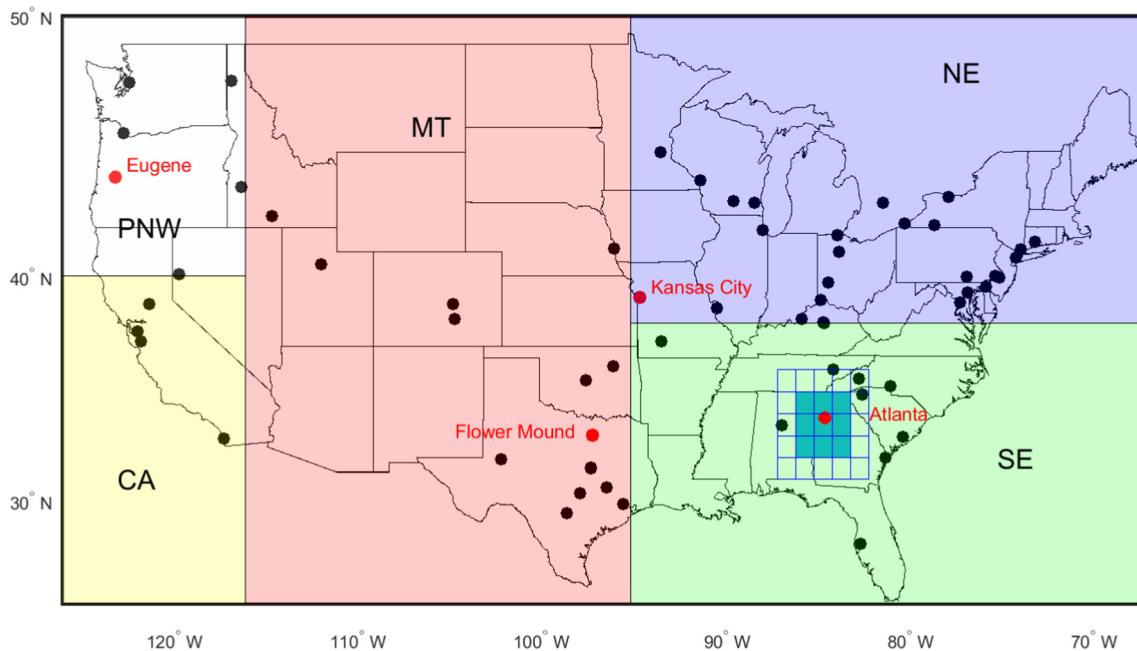


Fig. 1. Locations of NAB pollen stations: forecasted pollen locations are red dots; others are black dots. Geographic regions: Pacific Northwest (PNW, gray), California (CA, yellow), Mountain (MT, red), Northeast (NE, blue) and Southeast (SE, green). Domains used in the regional models: nine neighboring grid cells (teal shaded grids) and twenty-five grid cells (blue outlined grids) for Atlanta.

Applications version 2 (MERRA-2). We obtained daily mean temperature and maximum precipitation rate data from the MERRA-2 M2SDNXSLV dataset (GMAO, 2015a); and daily mean 2 m northward and eastward winds and 2 m specific humidity are from the MERRA-2 M2T1NXSLV dataset (GMAO, 2015b). We linearly interpolated the data's spatial resolution from $0.5^\circ \times 0.625^\circ$ to a $1^\circ \times 1^\circ$ grid.

Growing degree days (GDD) and chilling-day requirements are quantitative estimates of the energy acquired by a plant in order to overcome winter dormancy. Commonly used equations for GDD and chilling-day requirements (Eq. (2)) are calculated by accumulating temperatures from a set date over or under a base temperature. Unfortunately, base temperatures and dates can only be determined empirically and are specific to plant species and locations, so many of these values are not known. To overcome this problem and to take advantage of the fact that Random Forest model can include many features without overfitting, we calculated multiple GDD and chilling-day requirements and included them all as features. They were calculated by accumulating daily temperatures, t_i , over or under a base temperature for 30 and 90 days prior to the forecast day at lags on the monthly and seasonal time scale to cover periods that are potentially important to the pollen type being modeled (details in Section 2.6). We calculated GDD and chilling-day requirements with three different base temperatures, t_{base} , of 0, 5 and 10 deg. °C.

$$\begin{aligned} GDD &= \sum_i^{30,90\text{days}} t_i - t_{base}, \text{ if } t_i > t_{base} \\ \text{chilling} &= \sum_i^{30,90\text{days}} t_i - t_{base}, \text{ if } t < t_{base} \end{aligned} \quad (2)$$

We obtained daily mean downward solar radiation at the surface from NASA's Clouds and the Earth's Radiant Energy System (CERES): we used the adjusted shortwave downward flux at the surface for all-sky conditions from the SYN1deg-Day dataset (Rutan et al., 2015).

2.4. Vegetation data

We obtained vegetation data from NASA's Moderate Resolution Imaging Spectroradiometer (MODIS/Terra+Aqua) satellite data. The vegetation data were downloaded in various indicated spatial resolutions and interpolated to a $1^\circ \times 1^\circ$ grid.

We obtained data on the Leaf Area Index (LAI), the area of green leaf per unit ground surface area, and the Fraction of Photosynthetically Active Radiation (FPAR), the fraction of incoming solar radiation in the 400–700 nm range that is used by plants, from the MCD15A3H v006 dataset (Myneni et al., 2015). Due to the cloud cover and noise, this dataset is a 4-day composite data set that uses the best value in a 500 m pixel over the 4 days.

We obtained Daily Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI) data from the MOD13C1 v006 dataset (Didan, 2015). NDVI and EVI are both dimensionless indices that estimate the density of greenness on land. But EVI corrects for distortions from aerosols and ground cover. This dataset uses the best cloud free value in a 5600 m pixel over a 16-day period and was downloaded in the $0.05^\circ \times 0.05^\circ$ Climate Model Grid.

We included LAI, FPAR, NDVI and EVI as predictors because the increasing or greening of vegetation indicates the advent of spring. Because we are interested in the seasonal change in vegetation, and to reduce the noise in the data, we used a 60-day running mean of LAI, NDVI, EVI and FPAR.

Greenup is the date at which the onset of greenness occurs, identified by EVI. We obtained greenup data from the Cover Dynamics MCD12Q2 v006 data product at a 500 m pixel spatial resolution. When more than one growing cycle were identified per year, we use the first greenup date in the calendar year. Data were available only up until 2017. Mean greenup date over 2003–2017 was used as the greenup date for 2018.

2.5. Geographic data

To account for location and time, we incorporated latitude, longitude, day of year and year as features.

We calculated the number of daylight hours in a day (DL) at specific latitudes following equations in Caffarra et al. (2011) (Eq. (3)).

$$a = -\tan\left(\frac{\varphi\pi}{180}\right)\tan\left[\left(-\frac{23.45\pi}{180}\right)\cos\left(\frac{2\pi(t+10)}{365}\right)\right] \quad (3)$$

where t is the day of year, and φ is the latitude of the site in decimal degrees.

if $a > 1$, then $DL(t) = 0$.

if $a < -1$, then $DL(t) = 24$

if $-1 \leq a \leq 1$, then $DL(t) = (24/\pi) \arccos(a)$,

2.6. Time scale of features

The timescale of a variable can have differing effects on the pollen concentration. For example, temperatures are positively correlated with pollen concentration on a given day, and a warm winter can lead to an earlier onset of the pollen season. To encompass the information from many timescales into the models, we input meteorological and vegetative quantities as daily, weekly, monthly, seasonal and yearly means. In addition, means at various time lag times were also included as inputs (Table 1). We input all variables as daily means with lags of −6 to 0 days (i.e., 7 daily means for each variable); weekly inputs as the mean of days −6 to 0; and monthly inputs as the mean of days −29 to 0. We also included monthly means with lags of −30, −60, −90, −120 and 150 days as features in the models (e.g., a monthly mean with 60-day lag is the mean of days −89 to −60). Seasonal means are computed from days −89 to 0. Seasonal means with lags of −90 and −180 days were also included. Yearly means are computed from days −364 to day 0. The total number of features we input is 252 or 253 depending if pollen concentration was included as a feature (see Section 3.1). Table 1 lists features that were considered as candidates, although not every model included every variable in the list.

Many features are collinear, and the model was checked for overfitting by evaluating test error when increasing complexity of the model. We varied model complexity by modifying the number of observations in the terminal tree leaf, a hyperparameter in the Random Forest model. A smaller number of observations implies a deeper tree and a more complex model. We found that with increasing complexity, test error stays the same, implying that the model is not overfit (not shown).

3. Results and discussion

3.1. City models

Three versions of the Random Forest forecast models with different pollen features were evaluated. The first version, P1, uses the previous day pollen concentration (day 0) as a feature; the second version, P7, uses the previous 7-day mean pollen concentration (mean of day −6 to 0) as a feature; and, the third version, P0, does not include any pollen concentration as a feature. All other features were identical among these model versions (Table 1). The models use the meteorological and vegetation data from the $1^\circ \times 1^\circ$ grid cell in which the pollen station is located. These models are the city models. Because the Random Forest model has inherent randomness in the model, 20 iterations were run for each model to determine the variability of MAE.

Past pollen is a good indication of near-future pollen concentration as demonstrated by high 1-day lag autocorrelations of the four pollen types (Fig. 2). The autocorrelations generally decrease with increasing lags, thus, we hypothesized that the P1 models would outperform the P7 models and the models that included pollen as a feature (P1 and P7) would outperform the models that didn't include pollen (P0).

Table 1

Features of the Random Forest models are labeled 'y' (in green). Quantities not used are labeled 'n' (in red), and those unavailable or irrelevant are labeled 'u' (in red). All features indicated with 'y' except for pollen are input in every version of the model.

Input variable	Daily mean, Day 0	Daily mean, lags of 1 to 6 days	7-day mean (day -6 to day 0)	30-day mean (day -29 to day 0)	30-day mean, lags of 30,60,90,120,150 days	90-day mean (day -89 to day 0)	90-day mean, lags of 90,180 days	365-day mean or once a year	Time invariant
Day of year	y	u	u	u	u	u	u	u	u
Year	u	u	u	u	u	u	u	y	u
Latitude	u	u	u	u	u	u	u	u	y
Longitude	u	u	u	u	u	u	u	u	y
Length of daylight	y	u	u	u	u	u	u	u	u
Pollen	y	n	y	n	n	n	n	n	u
Temperature	y	y	y	y	y	y	y	y	u
Precipitation	y	y	y	y	y	y	y	y	u
U wind	y	y	y	y	y	y	y	y	u
V wind	y	y	y	y	y	y	y	y	u
Humidity	y	y	y	y	y	y	y	y	u
Shortwave	y	y	y	y	y	y	y	y	u
FPAR	y	y	y	y	y	y	y	y	u
LAI	y	y	y	y	n	n	n	n	u
NDVI	y	y	y	y	n	n	n	n	u
EVI	y	y	y	y	n	n	n	n	u
Green-up date	u	u	u	u	u	u	u	y	u
GDD ($t_{base}=0, 5, 10^{\circ}\text{C}$)	n	n	n	y	y	y	y	y	u
Chill ($t_{base}=0, 5, 10^{\circ}\text{C}$)	n	n	n	y	y	y	y	y	u

However, we found there is no significant difference in MAE between the three versions (Fig. 3). This result may be explained by the challenges that arise from days of missing pollen data at these stations.

For some station-years, pollen observations were only sampled on weekdays, e.g., Atlanta in 2008 and Kansas City in 2014 (Fig. 4a, d). The 1-day P1 forecast relies heavily on the previous day's pollen concentration (e.g., Apr 1–5 for Atlanta, Fig. 4a). When pollen concentrations were unavailable on the previous day, the P1 models performed poorly and predicted very low pollen concentrations (i.e., Aug 30–Sep 1 for

Kansas City, Fig. 4d). The P7 models used the mean of the available previous 7 days of pollen concentrations as a feature, and seldom were all 7 days missing during the pollen season. Nonetheless, the mean of the previous 7 days was a poorer predictor than the previous day alone, as the P7 forecast was less accurate than the P1 forecast for days in which pollen concentrations were available on the previous day (Fig. 4). Generally, the P7 pollen concentration forecast was lower than the P1 forecast and both were often lower than the observed pollen concentration.

We developed separate models for each forecast lead time (1 to 14 days) and found that the MAE did not increase significantly at longer lead times (Fig. 3). Each model was optimized specifically for its forecast lead time and so feature importance differs depending on the lead time. The importance of pollen as a feature decreases with lead time (Fig. 2). For example, pollen concentration at day 0 was the most important feature for the *Quercus* P1 1-day forecast but was only the 3rd most important feature for the 8-day forecast (not shown). The P1 8-day forecast model relied less on pollen as a feature and, as a result, handled missing weekend data better than the 1-day forecast (e.g., Apr 4–7 in Fig. 5).

A skillful forecast must, at a minimum, beat standard benchmarks of persistence (also known as a naïve forecast) and the daily climate mean (also known as climatology). A persistence forecast assumes that future conditions are identical to present conditions; therefore, a persistence forecast at any lead time is the observed value at day 0. How do the models compare to a forecast of persistence? The city models at Atlanta, Flower Mound and Eugene outperformed persistence for all forecast lead times (Fig. 3a, b, c). At Kansas City, persistence at 1- and 2-day lead times beat the city models, but beyond 2 days, the city models have a better forecast (Fig. 3d). The periodic dip in MAE for persistence at 7- and 14-day forecast lead times at Atlanta and Eugene are an artifact of the timing of the pollen observations: these stations collect data on weekdays and only sometimes on weekends during the pollen season, which leads to a minimum in MAE at forecast lead times that are integer multiples of 7 days (Fig. 3a, b).

How do the city models compare to a forecast of the daily climate mean? The daily climate mean at each station was calculated as the

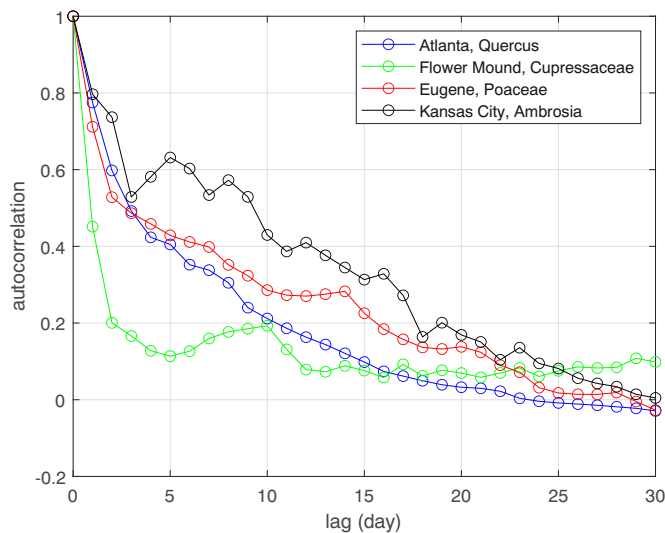


Fig. 2. Lag autocorrelations for observed pollen concentrations for *Quercus* at Atlanta, GA (blue), *Cupressaceae* at Flower Mound, TX (green), *Poaceae* at Eugene, OR (red) and *Ambrosia* at Kansas City (black).

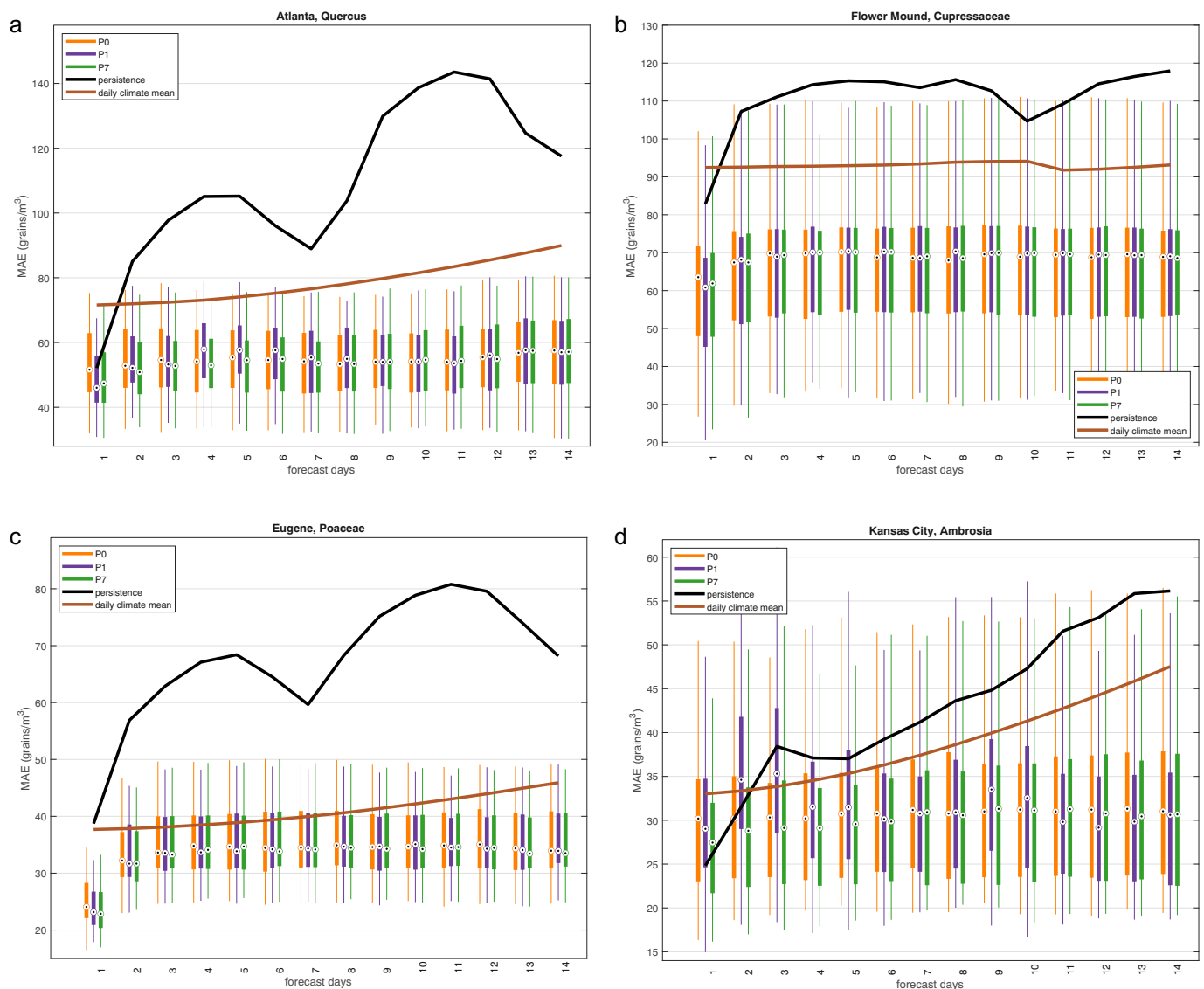


Fig. 3. Mean Absolute Error (MAE) in grains/m³ for 1–14 day lead forecast with models P0 (orange), P1 (purple) and P7 (green) compared with benchmark forecasts of daily climate mean (brown line) and persistence (black line). Box and whisker for (a) *Quercus* pollen at Atlanta, GA (b) *Cupressaceae* pollen at Flower Mound, TX (c) *Poaceae* pollen at Eugene, OR and (d) *Ambrosia* pollen at Kansas City, MO show median (black dot); 25th and 75th percentile (bottom and top of box); and range of all MAE (whiskers).

mean of the pollen concentration over all available years of data at each calendar day, and then smoothed with a 30-day running mean. Forecasts of the tree pollen types, *Cupressaceae* and *Quercus*, performed significantly better than the daily climate normal at all forecast lead times (Fig. 3a, c). We speculate that the models beat the daily climate mean because the models can reproduce interannual variability of the pollen season by the inclusion of monthly and seasonal time scales associated with timing of flowering and pollen release (i.e., growing degree days and chilling days). The forecast models for *Poaceae* and *Ambrosia* outperformed the daily climate normal, but not significantly. Growing degree and chilling days were developed for describing the phenology of trees and are not appropriate to apply to *Poaceae* and *Ambrosia*. The physical processes that affect the phenology of *Poaceae* and *Ambrosia* were not included into the model.

3.2. Regional models

Do models perform better if environmental information is included from the surrounding region? We hypothesized that using upstream meteorological conditions might improve the models due to advection

of weather. To test this hypothesis, the models were adapted to include features on a regional domain to forecast a given station at the domain center. This regional modeling approach was evaluated on two domains: one with nine (P0grid9) and another with twenty-five (P0grid25) $1^\circ \times 1^\circ$ grid cells (Fig. 1). Pollen data were unavailable on the neighboring grid cells and so were excluded as features. The total number of model features is a multiple of the grid cells (i.e., P0 has 252 features; P0grid9 has 252×9 features; P0grid25 has 252×25 features).

Fig. 6 shows that the MAE for P0grid9 and P0grid25 were almost the same for all pollen types at all forecast lead times, so adding additional grid cells beyond the closest 9 cells did not increase the performance of the models. Comparing P0 with P0grid9 and P0grid25, we found that the regional models were no better at forecasting pollen than the city models (Fig. 6).

3.3. Data-augmented models

We investigated data augmentation as a strategy to increase the performance of the models. Data augmentation is a method used in

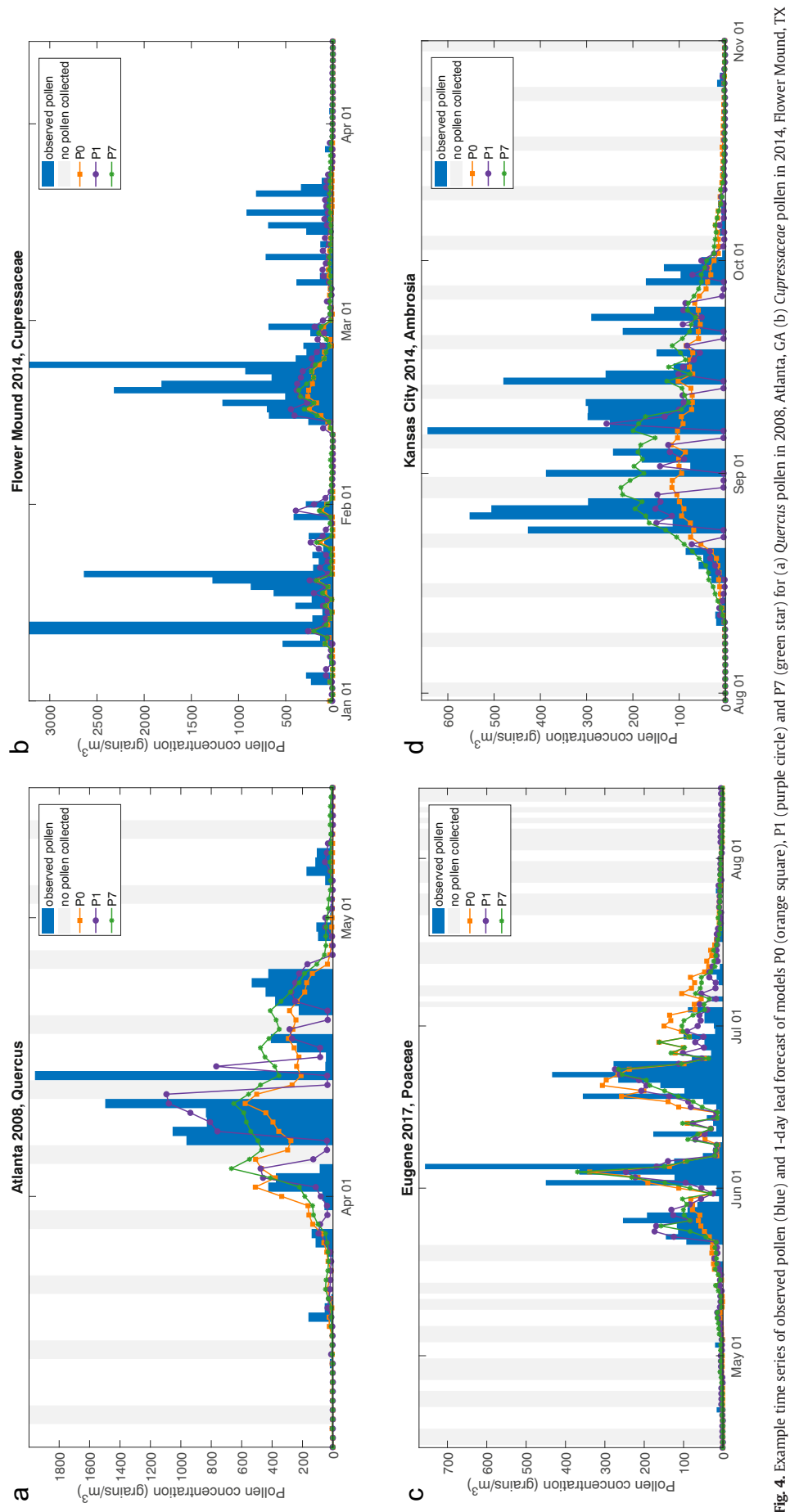


Fig. 4. Example time series of observed pollen (blue) and 1-day lead forecast of models P0 (orange circle), P1 (purple square), and P7 (green star) for (a) *Quercus* pollen in 2008, Atlanta, GA (b) *Cupressaceae* pollen in 2014, Flower Mound, TX (c) *Poaceae* pollen in 2017, Eugene, OR and (d) *Ambrosia* pollen in 2014, Kansas City, MO.

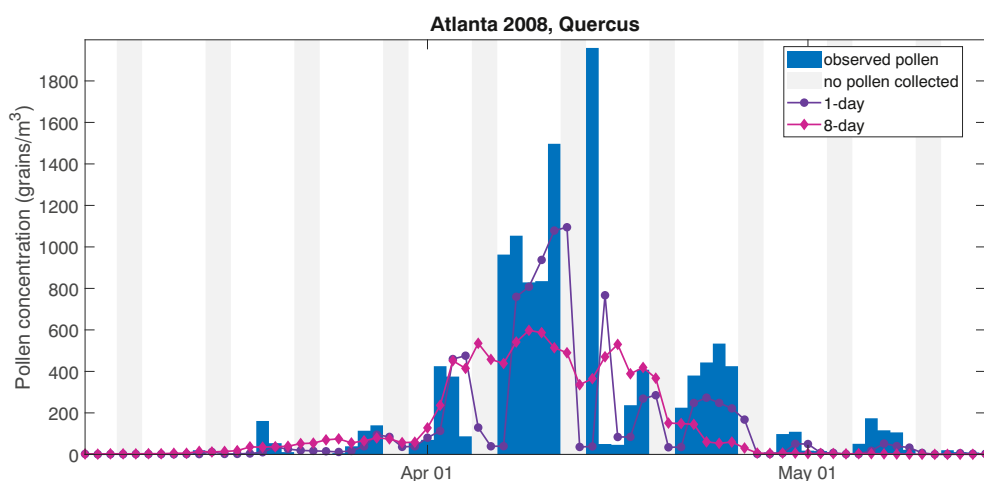


Fig. 5. Example time series of observed pollen (blue) and P1 model for 1-day (purple circle) and 8-day (magenta diamond) lead forecast, *Quercus* pollen in 2008, Atlanta, GA.

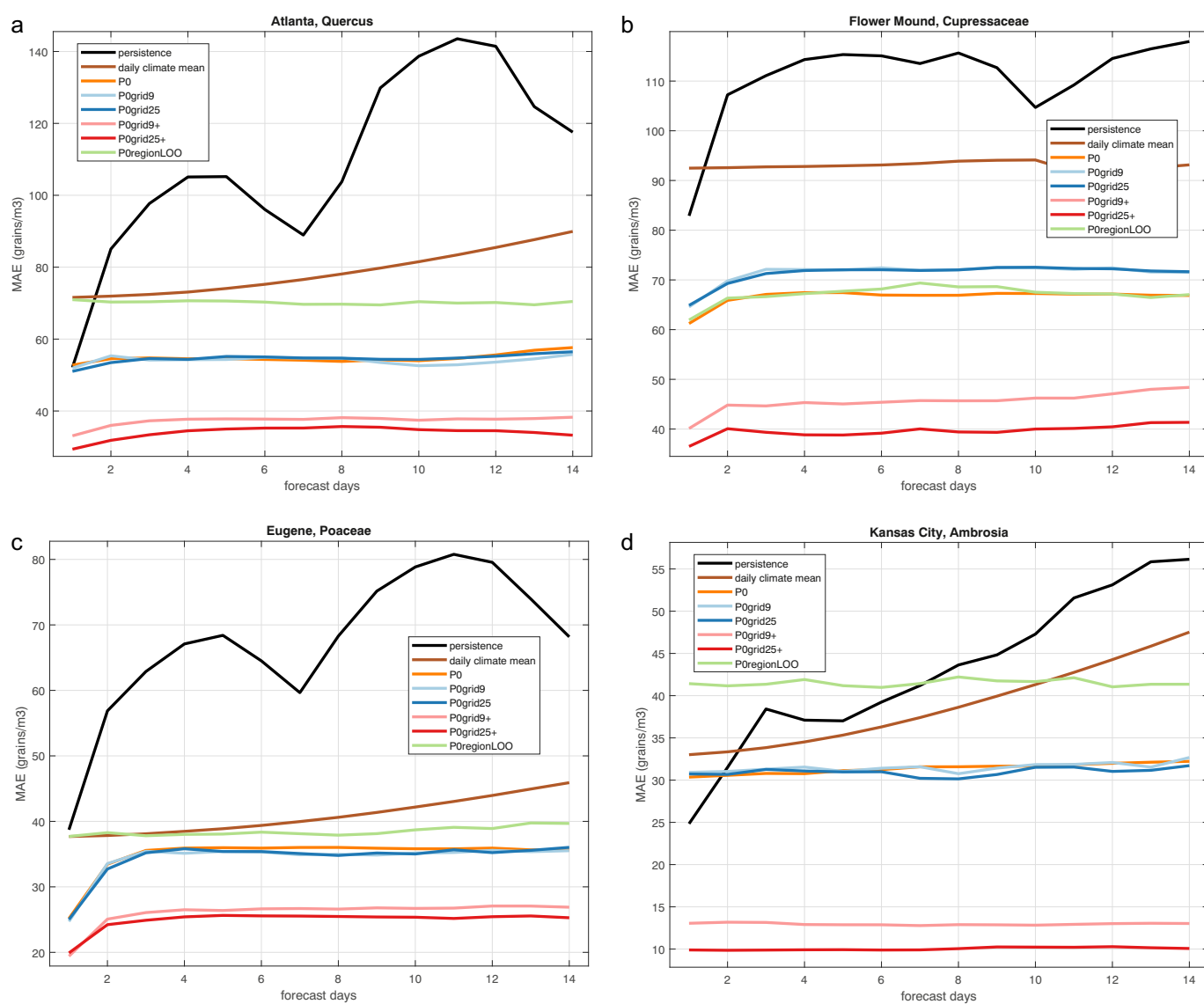


Fig. 6. Mean Absolute Error for 1–14 day lead forecast with models P0 (orange), P0grid9 (light blue), P0grid25 (dark blue), P0grid9+ (pink), P0grid25+ (red), P0regionLOO (green), daily climate mean (brown) and persistence (black). (a) *Quercus* at Atlanta, GA (b) *Cupressaceae* at Flower Mound, TX (c) *Poaceae* at Eugene, OR and (d) *Ambrosia* at Kansas City, MO.

machine learning to increase the amount of data available for training models without collecting new data. As in the previous section, the models included feature data from domains with 9 and 25 grid cells, called P0grid9+ and P0grid25+, respectively. But these models were developed by assuming that each grid cell contributes equally to influence pollen concentration at the domain center. The models are constructed by imagining a station sits at each grid, with the meteorological and vegetation feature data unique to the grid cell but with the pollen observations label data from the actual station at the center of the domain. In practice, data at each grid cell was strung together with an offset in the time dimension, and the total number of time steps in the model is increased by a multiple of the number of grid cells. For example, the P0grid9+ model that is trained on 12 years of data (with 365 days in a year) will use $12 \times 365 \times 9$ times steps in the model. The same number of features as the P0 models, 252, were used in these models.

P0grid25+ outperformed or was as good as P0grid9+ at 1–14 day forecast lead times at all locations (Fig. 6). Both data-augmented models (P0grid9+, P0grid25+) significantly outperformed the city and regional models at all forecast lead times. Because the models benefit from data augmentation, we conclude that increasing the number of days of observed pollen data at a given station or increasing the number of stations would improve model accuracy.

The data augmented models outperformed the city and regional models in this study. As our best performing model, we also computed the coefficient of determination R^2 so that we could compare results with other studies. For a 1-day forecast, P0grid9+ had R^2 values between 0.64 and 0.88 (0.77 for oak pollen at Atlanta oak; 0.76 for grass pollen at Eugene; 0.64 cedar pollen at Flower Mound and 0.88 for ragweed pollen at Kansas City). Two other recent studies used Random Forest models to predict pollen: they used meteorology to predict pollen for the same day, a 0-day forecast. They reported R^2 values of between 0.12 and 0.51 (Nowosad et al., 2018) and a value of 0.37 (Zewdie et al., 2019). Although the P0grid9+ model is forecasting for 1-day, a more difficult prediction than 0-day forecast, the P0grid9+ performs better than other Random Forest pollen models.

3.4. Forecasting in locations without pollen observations

We were interested in using the model to forecast pollen in locations where no pollen observations were collected. To test this potential, a version of the model (P0regionLOO) was developed using data-denial Leave-One-Out cross-validation technique, with the test station as the forecast location. We assessed the ability of the model to forecast target pollen types at test stations within four of the five geographic regions in CUSSC: *Poaceae* in the Pacific Northwest, *Cupressaceae* in the Mountain region, *Ambrosia* in the Northeast and *Quercus* in the Southeast (Fig. 1). A model for each region was trained on the NAB stations within a region with the exception of the test station. The model was then applied to the features at the test station to predict the pollen concentration at the test station.

At Atlanta and Eugene, the P0regionLOO models performed better than persistence and the daily climate mean for forecast lead times longer than two days but did not perform well compared with the city and regional models (Fig. 6a, c). At Kansas City, the P0regionLOO models did not give skillful forecasts at shorter forecast lead times; the models performed better than persistence only for forecasts longer than 7 days and performed better than the daily climate mean for forecasts longer than 10 days (Fig. 6d). Kansas City is located on the edge of the Northeast region (Fig. 1) and the models are trained on other stations in the Northeast, which may not be representative of Kansas City. At Flower Mound, the P0regionLOO models performed as well as the P0 city models, better than the regional models (P0grid9, P0grid25), but not as well as the data-augmented models (P0grid9+, P0grid25+) (Fig. 6b). There is a relatively dense network of NAB stations near Flower Mound in Texas and Oklahoma (Fig. 1), and these stations have very few missing days and relatively long records. The high spatial and temporal resolution

of these stations are able to represent pollen concentration at the Flower Mound station well, and results in the P0regionLOO models performing as well as the P0 city models.

We recognize that the geographic regions are large and within each region the target pollen type maybe comprised of multiple species. This diversity of species may contribute to the decrease in forecast skill.

4. Conclusion

We developed Random Forest models to forecast daily pollen concentrations using meteorology, vegetation and pollen concentration as features. Using this framework, we generated skillful city, regional and data-augmented models for four pollen types and locations in CUSSC for 1–14 day forecast lead times. For forecasts longer than 1-day, the models performed better than persistence and daily climate means. The data-augmented models were the best performing models.

We found that including pollen observations as features (P1, P7 city models) did not significantly improve the performance of the forecast over excluding pollen as a feature (P0 city models). P1 city models have the potential to be more skillful than P7 and P0 city models because there is a high 1-day lag autocorrelation for observed pollen concentrations, but the potential is not realized due to a lack of consistent daily pollen sampling. The P7 city models were designed to compensate for the missing data by using a 7-day mean of available pollen data, but because autocorrelation of pollen concentration drops rapidly after 1 day, the inclusion of 7-day mean as a feature was not able to significantly improve in model performance.

Limits in the forecast skill of the models can be partly attributed to the incomplete spatial and temporal resolution of the pollen observations. Data augmentation can compensate for this lack of data as evidenced by significant increases in performance of the data-augmented models (P0grid9+ and P0grid25+) over the other approaches. P0grid25+ performed better than P0grid9+, implying there may be a quantifiable relationship between the amount of pollen data used to train the model and the performance of the model. However, the data-augmented models take more time to run than the other models, so for practical purposes, computation time should also be considered. Future work is needed to find the optimal amount of data augmentation for a given NAB pollen data set, while considering the tradeoffs between performance and efficiency.

Because the spatial resolution of the NAB pollen data is limited, we developed a model to forecast in locations with no pollen observations (P0regionLOO). We demonstrated that using available pollen stations in a geographic region can yield forecasts as good as a P0 model (i.e., at Flower Mound, Fig. 6b). However, the skill of the P0regionLOO is dependent on the temporal resolution of the observed pollen data and spatial distance between forecast location and training stations within the region. Grouping stations by ecoclimatic zones might improve the model and should be investigated in future research.

There are some limitations to the Random Forest model framework. The model is based on randomness, so only a random subset of the data is used to train the model and produce a forecast. One forecast may not be representative. Generating an ensemble of forecasts and using the most probable forecast from the ensemble may be a more accurate method to forecast. Another constraint of this framework is that Random Forest models are only capable of predicting within the limits of the trained data. If future climate change moves the meteorology or pollen concentration into a new regime that has not been observed, this model will not be able to predict it.

There are many benefits of this Random Forest model framework: (1) new data can be easily incorporated into the model so that information from updated daily pollen concentrations and meteorological observations can be immediately used to generate a better forecast; (2) the model is flexible and can be modified easily to forecast at different locations or different pollen types; and, (3) this model can also be used to forecast in locations that have no pollen observations.

Funding

This work was supported by the National Aeronautics and Space Administration (NASA), United States of America, Grant 15-HAQST15-0025.

CRediT authorship contribution statement

Fiona Lo: Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization, Funding acquisition. **Cecilia M. Bitz:** Conceptualization, Methodology, Resources, Writing – review & editing, Funding acquisition. **Jeremy Hess:** Conceptualization, Methodology, Data curation, Writing – review & editing, Resources, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank the National Allergy Bureau for providing pollen data from the following stations (the associated professionals and clinics for each station are listed below).

Stanley M Fineman, MD MBA FAAAAI, Allergy and Asthma Clinic, Marietta (Atlanta), GA. Sheila Amar, MD, FAAAAI, FAAAAI, Allergy & Asthma Center of Georgetown, Austin, TX. Jonathon Matz, MD, FAAAAI, & David Golden, MD, FAAAAI, Baltimore, MD. Linda Ford, MD, FAAAAI, The Asthma and Allergy Center, PC, Bellevue, NE. David Weldon, MD, FAAAAI, FAAAAI, Scott & White Clinic, College Station, TX. Robert Nathan, MD, FAAAAI, & Daniel Soteres, MD, MPH, FAAAAI, Asthma and Allergy Associates, PC, Colorado Springs, CO. Donald Pulver, MD, FAAAAI, Allergy, Asthma & Immunology of Rochester, Rochester, NY. Brian Huxtable, RAPCA, Dayton, OH. Duane Harris, MD, FAAAAI, Intermountain Allergy & Asthma Clinic, Draper, UT. Philip Gallagher, MD, FAAAAI, Allergy & Asthma Associates of Northeastern Pennsylvania, Erie, PA. Kraig Jacobson, MD, FAAAAI, Allergy & Asthma Research Group, Eugene, OR. Neil Kao, MD, FAAAAI, Allergic Disease and Asthma Center, Greenville, SC. Tony Huynh, City of Houston, Houston, TX. Jay Portnoy, MD, FAAAAI, Children's Mercy Hospital, Kansas City, MO. James Anderson, MLT, OSHTech, London, ON. Robert Bush, MD, FAAAAI, University of Wisconsin Medical School, Madison, WI. Joseph Leija, MD, FAAAAI, Melrose Park, IL. Harold Kaiser, MD, FAAAAI, Clinical Research Institute, Minneapolis, MN. Warren Filley, MD, FAAAAI, OK Allergy Asthma Clinic, Oklahoma City, OK. Martha Tarpay, MD, Allergy & Asthma Center, Oklahoma City, OK. Amy L. Darter, MD FAAAAI, Oklahoma Institute of Allergy & Asthma, Oklahoma City, OK. Wayne Wilhelm, Saint Louis County Health Department, St. Louis, MO. Robert Gomez, Wilford Hall Ambulatory Surgical Center, San Antonio, TX. Erika Gonzalez-Reyes, MD, FAAAAI, **Sylvana Research Associates** San Antonio, TX. Alan Goldsobel, MD, FAAAAI, & James Wolfe, MD, FAAAAI, Allergy and Asthma Associates of Northern California, San Jose, CA. Frank Virant, MD, FAAAAI, Northwest Asthma & Allergy Center, Seattle, WA. Rhizza Adams, Springfield-Greene County Health Department, Springfield, MO. James Love, Jr., MD, PhD, FAAAAI, Allergy Clinic of Tulsa, Tulsa, OK. Richard Henry, MD, Asthma & Allergy of Idaho, Twin Falls, ID. N.J. Amar, MD FAAAAI, **Allergy and Asthma Center**, Waco, TX. Pramila K. Daftary, MD, FAAAAI, Allergy & Asthma Care of Waco, Waco, TX. Susan E. Kosisky, MHA, US Army Garrison-Forest Glen, Silver Spring, MD (Washington D.C.) Christopher Randolph, MD, FAAAAI, Waterbury, CT. Michael Nickels, MD, PhD, Allergy and Asthma Consultants, Inc., York, PA. Sunil P. Perera, MD FAAAAI, Allergy Medical Group of the North Area, Roseville, CA. James L. Sublett, MD FAAAAI, Family Allergy &

Asthma, Louisville, KY. Ghassan S Safadi, MD FAAAAI, Toledo Allergy, Sylvania, OH. Richard E. Collins III, MD, Carolina Asthma and Allergy Center, Charlotte, NC. Shari Montandon, DO, Asthma and Allergy of Idaho, Coeur d'Alene, ID. Ghassan S. Safadi, MD FAAAAI, Findlay, OH. Michael Miller, MD FAAAAI, Knoxville, TN. Mary Morris, MD, Onalaska (La-Crosse), WI. Mark H. Moss, MD, CSC, Madison, WI. John D. Bray, MD FAAAAI, Midland, TX. Donald J. Dvorin, MD, FAAAAI, Mount Laurel, NJ. Fred H. Lewis, MD FAAAAI, Olean, NY. Donald J. Dvorin, MD, FAAAAI, Philadelphia, PA. Allyson Tevzian, MD, Pleasanton, CA. Leonard Shapiro, MD FAAAAI, Sparks (Reno), NV. Mark C. Glaum, MD PhD FAAAAI, USF College of Medicine, Tampa, FL. Susan E. Kosisky, MHA, Silver Spring, MD. David Cypcar, MD FAAAAI, Asheville, NC. Andrew E. Davidson, MD FAAAAI, Charleston, SC. Marie H. Fitzgerald, MD FAAAAI, Family Allergy and Asthma Care, Flower Mound, TX. Beth A. Miller, MD, Lexington, KY. Rajiv Arora, MD FAAAAI, Lexington, KY. Brad H. Goodman, MD FAAAAI & Bruce D. Finkel, MD FAAAAI, Savannah, GA. Raj Srinivasan, MD, Vancouver, WA. Guy S Robinson, PhD, Armonk, NY. Weily Soong, MD FAAAAI, Birmingham, AL. Neetu Talreja, MD, Boise, ID. Guy Robinson, PhD, New York, NY. Robert T. Reid III, San Diego, CA.

References

- Astray, G., Fernández-González, M., Rodríguez-Rajo, F.J., López, D., Mejuto, J.C., 2016. Airborne castanea pollen forecasting model for ecological and allergological implementation. *Sci. Total Environ.* 548–549, 110–121. <https://doi.org/10.1016/j.scitotenv.2016.01.035>.
- Atlanta Allergy & Asthma. Atlanta Allergy Pollen Count and Model Activity. Accessed: Dec 24, 2020, http://www.atlantaallergy.com/pollen_counts.
- Breiman, L., 2001. Random forests. *Mach. Learn.* 45 (1), 5–32. <https://doi.org/10.1023/A:1010933404324>.
- Bruffaerts, N., Smedt, T.D., Delcloo, A., Simons, K., Hoebek, L., Verstraeten, C., Van Nieuwenhuysse, A., Packeu, A., Hendrickx, M., 2018. Comparative long-term trend analysis of daily weather conditions with daily pollen concentrations in Brussels, Belgium. *Int. J. Biometeorol.* 62 (3), 483–491. <https://doi.org/10.1007/s00484-017-1457-3>.
- Caffarra, A., Donnelly, A., Chuine, I., 2011. Modelling the timing of *Betula pubescens* budburst. II. Integrating complex effects of photoperiod into process-based models. *Clim. Res.* 46 (2), 159–170. <https://doi.org/10.3354/cr00983>.
- Chuine, I., Belmonte, J., 2004. Improving prophylaxis for pollen allergies: predicting the time course of the pollen load of the atmosphere of major allergenic plants in France and Spain. *Grana* 43 (2), 65–80. <https://doi.org/10.1080/0017130410019163>.
- D'Amato, G., Pawankar, R., Vitale, C., Lanza, M., Molino, A., Stanziola, A., Sanduzzi, A., Vatrella, A., D'Amato, M., 2016. Climate change and air pollution: effects on respiratory allergy. *Allergy, Asthma Immunol. Res.* 8 (5), 391–395. <https://doi.org/10.4168/aaair.2016.8.5.391>.
- de Weger, L.A., Bergmann, K.C., Rantio-Lehtimäki, A., Dahl, A., Buters, J., De' champ, C., et al., 2013. Impact of pollen. In: Sofiev, M., Bergmann, K.-C. (Eds.), *Allergenic Pollen: A Review of the Production, Release, Distribution and Health Impacts*. Springer, Dordrecht, pp. 161–215.
- Didan, K. (2015). MOD13C1 MODIS/Terra Vegetation Indices 16-Day L3 Global 0.05Deg CMG V006 [Data set]. NASA EOSDIS Land Processes DAAC. Accessed 30 May 2019 from <https://doi.org/10.5067/MODIS/MOD13C1.006>.
- García-Mozo, H., Chuine, I., Aira, M.J., Belmonte, J., Bermejo, D., Díaz de la Guardia, C., Elvira, B., Gutierrez, M., Rodriguez-Rajo, J., Ruiz, L., Trigo, M.M., Tormo, R., Valencia, R., Galán, C., 2008. Regional phenological models for forecasting the start and peak of the Quercus pollen season in Spain. *Agric. For. Meteorol.* 148 (3), 372–380. <https://doi.org/10.1016/j.agrformet.2007.09.013>.
- Global Modeling and Assimilation Office (GMAO) (2015a), MERRA-2 statD_2d_slv_Nx: 2d, Daily, Aggregated Statistics, Single-Level, Assimilation, Single-Level Diagnostics V5.12.4, Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC), Accessed: May 20, 2019, <https://doi.org/10.5067/9SC1VNTGWV3>.
- Global Modeling and Assimilation Office (GMAO) (2015b), MERRA-2 tavg1_2d_slv_Nx: 2d, 1-Hourly, Time-Averaged, Single-Level, Assimilation, Single-Level Diagnostics V5.12.4, Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC), Accessed: Mar 28 2019, <https://doi.org/10.5067/VJAFPL1CSIV>.
- Khwarahm, N., Dash, J., Atkinson, P.M., Newnham, R.M., Skjøth, C.A., Adams-Groom, B., Caulton, E., Head, K., 2014. Exploring the spatio-temporal relationship between two key aeroallergens and meteorological variables in the United Kingdom. *Int. J. Biometeorol.* 58 (4), 529–545. <https://doi.org/10.1007/s00484-013-0739-7>.
- Lamb, C.E., Ratner, P.H., Johnson, C.E., Ambegaonkar, A.J., Joshi, A.V., Day, D., Sampson, N., Eng, B., 2006. Economic impact of workplace productivity losses due to allergic rhinitis compared with select medical conditions in the United States from an employer perspective. *Curr. Med. Res. Opin.* 22 (6), 1203–1210. <https://doi.org/10.1185/030079906X112552>.
- Liu, X., Wu, D., Zewdie, G.K., Wijerante, L., Timms, C.I., Riley, A., Levettin, E., Lary, D.J., 2017. Using machine learning to estimate atmospheric Ambrosia pollen concentrations in Tulsa, OK. *Environ. Health Insights* 11. <https://doi.org/10.1177/1178630217699399>.

- Lo, F., Bitz, C.M., Battisti, D.S., Hess, J.J., 2019. Pollen calendars and maps of allergenic pollen in North America. *Aerobiologia*. <https://doi.org/10.1007/s10453-019-09601-2>.
- Marcotte, D.E., 2015. Allergy test: seasonal allergens and performance in school. *J. Health Econ.* 40, 132–140. <https://doi.org/10.1016/j.jhealeco.2015.01.002>.
- Medek, D.E., Simunovic, M., Erbas, B., Katelaris, C.H., Lampugnani, E.R., Huete, A., Beggs, P.J., Davies, J.M., 2019. Enabling self-management of pollen allergies: a pre-season questionnaire evaluating the perceived benefit of providing local pollen information. *Aerobiologia*. <https://doi.org/10.1007/s10453-019-09602-1>.
- Meltzer, E.O., Bukstein, D.A., 2011. The economic impact of allergic rhinitis and current guidelines for treatment. *Ann. Allergy Asthma Immunol.* 106 (2), S12–S16. <https://doi.org/10.1016/j.anai.2010.10.014>.
- Miguel, A.G., Taylor, P.E., House, J., Glovsky, M.M., Flagan, R.C., 2006. Meteorological influences on Respirable fragment release from Chinese elm pollen. *Aerosol Sci. Technol.* 40 (9), 690–696. <https://doi.org/10.1080/02786820600798869>.
- Myneni, R., Knyazikhin, Y., Park, T. (2015). MCD15A3H MODIS/Terra+Aqua Leaf Area Index/FPAR 4-day L4 Global 500m SIN Grid V006. NASA EOSDIS Land Processes DAAC. Accessed 9 May 2019 from doi:<https://doi.org/10.5067/MODIS/MCD15A3H.006>.
- Nathan, R.A., 2007. The burden of allergic rhinitis. *Allergy and Asthma Proceedings* 28 (1), 3–9. <https://doi.org/10.2500/aap.2007.28.2934>.
- Nowosad, J., 2016. Spatiotemporal models for predicting high pollen concentration level of *Corylus*, *Alnus*, and *Betula*. *Int. J. Biometeorol.* 60, 843–855. <https://doi.org/10.1007/s00484-015-1077-8>.
- Nowosad, J., Stach, A., Kasprzyk, I., Grewling, Ł., Latałowa, M., Puc, M., Myszkowska, D., Weryszko-Chmielewska, E., Piotrowska-Weryszko, K., Chłopek, K., Majkowska-Wojciechowska, B., Uruska, A., 2015. Temporal and spatiotemporal autocorrelation of daily concentrations of *Alnus*, *Betula*, and *Corylus* pollen in Poland. *Aerobiologia* 31 (2), 159–177. <https://doi.org/10.1007/s10453-014-9354-2>.
- Nowosad, J., Stach, A., Kasprzyk, I., Chłopek, K., Dąbrowska-Zapart, K., Grewling, Ł., Latałowa, M., Pędziszewska, A., Majkowska-Wojciechowska, B., Myszkowska, D., Piotrowska-Weryszko, K., Weryszko-Chmielewska, E., Puc, M., Rapiejko, P., Stosik, T., 2018. Statistical techniques for modeling of *Corylus*, *Alnus*, and *Betula* pollen concentration in the air. *Aerobiologia* 34 (3), 301–313. <https://doi.org/10.1007/s10453-018-9514-x>.
- Rathnayake, C.M., Metwali, N., Jayarathne, T., Kettler, J., Huang, Y., Thorne, P.S., O'Shaughnessy, P.T., Stone, E.A., 2017. Influence of rain on the abundance of bioaerosols in fine and coarse particles. *Atmos. Chem. Phys.* 17 (3), 2459–2475. <https://doi.org/10.5194/acp-17-2459-2017>.
- Ritenberga, O., Sofiev, M., Siljamo, P., Saarto, A., Dahl, A., Ekebom, A., Sauliene, I., Shalaboda, V., Severova, E., Hoebeke, L., Ramfjord, H. (2018). A statistical model for predicting the inter-annual variability of birch pollen abundance in northern and North-Eastern Europe. *Sci. Total Environ.*, 615, 228–239. doi:<https://doi.org/10.1016/j.scitotenv.2017.09.061>.
- Rutan, D.A., Kato, S., Doelling, D.R., Röse, F.G., Nguyen, L.T., Caldwell, T.E., Loeb, N.G., 2015. CERES synoptic product: methodology and validation of surface radiant flux. *J. Atmos. Ocean. Technol.* 32, 1121–1143. <https://doi.org/10.1175/JTECH-D-14-00165.1>.
- Tseng, Y.-T., Kawashima, S., Kobayashi, S., Takeuchi, S., Nakamura, K., 2018. Algorithm for forecasting the total amount of airborne birch pollen from meteorological conditions of previous years. *Agric. For. Meteorol.* 249, 35–43. <https://doi.org/10.1016/j.agrformet.2017.11.021>.
- Valencia, J.A., Astray, G., Fernández-González, M., Aira, M.J., Rodríguez-Rajo, F.J., 2019. Assessment of neural networks and time series analysis to forecast airborne *Parietaria* pollen presence in the Atlantic coastal regions. *Int. J. Biometeorol.* 63, 735–745. <https://doi.org/10.1007/s00484-019-01688-z>.
- Vázquez, L.M., Galán, C., Domínguez-Vilches, E., 2003. Influence of meteorological parameters on *Olea* pollen concentrations in Córdoba (South-Western Spain). *Int. J. Biometeorol.* 48 (2), 83–90. <https://doi.org/10.1007/s00484-003-0187-x>.
- Wheatley, L.M., Togias, A., 2015. Allergic rhinitis. *N. Engl. J. Med.* 372, 456–463. <https://doi.org/10.1056/NEJMcp1412282>.
- Zewdie, G.K., Liu, X., Wu, D., Lary, D., Levettin, 2019. Applying machine learning to forecast daily Ambrosia pollen using environmental and NEXRAD parameters. *E. Environ Monit Assess* 191 (Suppl. 2), 261. <https://doi.org/10.1007/s10661-019-7428-x>.