Towards emulating an explicit organic chemistry mechanism with random forest models

Camille Mouchel-Vallon[1] and Alma Hodzic[2]

[1]Laboratoire d'Aérologie, Université de Toulouse, CNRS, UPS, Toulouse, France
[2]Atmospheric Chemistry Observations and Modeling, National Center for Atmospheric Research, Boulder, CO 80301, USA

Key Points:

- Random forests were trained to reproduce toluene and dodecane SOA formation calculated by an explicit organic chemistry mechanism
- The random forests performances are highly sensitive to the chemical regime
- The number of predictors cannot be too high and their selection depends on the precursor

Corresponding author: C. Mouchel-Vallon, camille.mouchel-vallon@aero.obs-mip.fr

Abstract

Predicting secondary organic aerosol (SOA) formation relies either on extremely detailed, numerically expensive models accounting for the condensation of individual species or on extremely simplified, numerically affordable models parameterizing SOA formation for large-scale simulations. In this work, we explore the possibility of creating a random forest to reproduce the behavior of a detailed atmospheric organic chemistry model at a fraction of the numerical cost. A comprehensive dataset was created based on thousands of individual detailed simulations, randomly initialized to account for the variety of atmospheric chemical environments. Recurrent random forests were trained to predict organic matter formation from dodecane and toluene precursors, and the partitioning between gas and particle phases. Validation tests show that the random forests perform well without any divergence over 10 days of simulations. The distribution of errors shows that the sampling of initial conditions for the training simulations needs to focus on chemical regimes where SOA production is the most sensitive. Sensitivity tests show that specializing multiple random forests for a specific chemical regime is not more efficient than training a single general random forest for the entire dataset. The most important predictors are those providing information about the chemical regime, oxidants levels and existing organic mass. The choice of predictors is crucial as using too many unimportant predictors reduces the performances of the random forests.

Plain Language Summary

Organic compounds constitute a significant fraction of atmospheric particles and thus have an impact on health and climate. Predicting the contribution of organic compounds to atmospheric particles is extremely complex because of the very large number of different chemical species potentially condensing into the aerosol phase. Air quality and climate models usually rely on simplified, empirical approaches to predict organic aerosol mass concentrations, based on laboratory experiments. In this work, we apply a machine learning approach to construct a tool that behaves like the most detailed organic chemistry model, for a numerical cost affordable by air quality and climate models. Building upon this method, it will be possible to bring the complexities of organic chemistry to large-scale models.

1 Introduction

Secondary organic aerosol (SOA) constitutes a major fraction of atmospheric particles worldwide. It is composed of a multitude of organic compounds (*e.g.* Kourtchev et al., 2016). Our current understanding and modeling of SOA formation processes are highly uncertain (Pai et al., 2020) and involve representing the complex interplay between gas-phase oxidation and condensation of semi- and low-volatile organic species. SOA models need to include processes such as (i) the multi-step oxidation of the large variety of organic compounds emitted naturally and by human activities, (ii) the condensation of semi-volatile species to the particle phase and (iii) the heterogeneous and in-particle reactivity of condensed species. This complexity can only be represented in models that explicitly account for aerosol physico-chemical processes. In these so-called explicit models, the aim is to represent the fate of each individual chemical species through individual reactions, which can number in the 10^9 range (*e.g.* Aumont et al., 2005). The Generator of Explicit Chemistry and Kinetics for Organics in the Atmosphere (Aumont et al., 2005) is an example of such a model able to generate chemical mechanisms that explicitly describe the oxidation of organic compounds in the atmosphere, as well as their condensation into the particle phase (Camredon et al., 2007). It has previously been used to study SOA formation in various settings such as atmospheric chamber experiments (La et al., 2016), sensitivity studies (Valorso et al., 2011; Aumont et al., 2012; Hodzic

et al., 2015) and urban plume modeling (Lee-Taylor et al., 2015; Mouchel-Vallon et al., 2020).

Because of their size, explicit mechanisms like those generated by GECKO-A cannot be used in 3D air quality models, which rely on empirical SOA parameterizations. The volatility basis set (VBS, Donahue et al., 2006) and its derivatives (*e.g.,* Donahue et al., 2011; Cappa & Wilson, 2012) are the most prevalent representations of SOA chemistry in this field. In such models, simplifications are made to represent SOA formation by grouping organic species of similar properties into discretized bins, *i.e.,* volatility or oxidation state are the typically chosen properties for the VBS bins. This approach has been presented by Pankow et al. (2015) as the *anonymized* view of SOA modeling, as opposed to the *molecular* view of SOA modeling used by explicit models.

Previous attempts have been made to bring the molecular view of SOA modeling to 3D models. Li et al. (2015) included the near explicit Master Chemical Mechanism (MCM v3.2, Saunders et al., 2003; Jenkin et al., 2003), in the Community Multiscale Air Quality model (CMAQ, Foley et al., 2010). MCM is a *near* explicit mechanism, as some simplifications are made to simplify its development such as removing unlikely reaction channels and simplifying the oxidation of minor or unknown products. At the time MCM therefore used approximately 17 000 reactions involving approximately 6000 species to represent the progressive oxidation of 142 primary hydrocarbons. Although the implementation of MCM in CMAQ was able to reproduce reasonably well the observed SOA surface concentrations over eastern US for a case study, this approach did not have further applications to our knowledge, and was limited by the considerable computational cost required to run regional scale simulations.

Lannuque et al. (2018) created VBS-GECKO, an empirical VBS parameterization where the stoichiometric coefficients were optimized to fit data produced from GECKO-A runs instead of being fitted to empirical data. Their method had the advantage of parameterizing the model over the multi-day simulated aging of SOA, which cannot be obtained from the shorter chamber studies used to derive traditional VBS parameterizations. Lannuque et al. (2020) ran VBS-GECKO in an air quality model (Menut et al., 2013) and showed that VBS-GECKO was producing more SOA in the summer over Europe compared to the traditional SOA parameterization that is based on laboratory data (Couvidat et al., 2012). Because, in essence, their resulting model was a linear combination of multiple VBS produced for different levels of pollution, it relied on the assumption that atmospheric chemistry behaves linearly between the selected chemical regimes. As a result, VBS-GECKO may have been applied outside of its application domain.

Here, we propose to use a machine learning (ML) approach to bring the molecular view to 3D chemistry-climate models across a range of chemical regimes representative of tropospheric conditions. Machine learning techniques have been applied previously for air quality forecasts (Liao et al., 2020) demonstrating that it is possible to run a trained artificial intelligence in a 3D model. Keller and Evans (2019) used the GEOS-Chem chemical mechanism solver to train multiple random forests that were then able to emulate the chemical solver behavior in that same model for various pollutant, for a fraction of the computational cost of the default GEOS-Chem model. Kelp et al. (2020) improved on this method by using a unique neural network to predict 20 chemical species. They implemented it in GEOS-Chem (Kelp et al., 2022), achieving stable one-year simulations for ozone prediction with less than 10 % bias compared to the reference and reducing computational times by a factor of five. The motivation of these previous studies stems from reducing the costs of calculating chemistry, that is usually taking from 50 % to 90 % of the computational costs of running global chemistry models such as GEOS-Chem (Keller & Evans, 2019).

Schreck et al. (2022) recently presented a neural network approach to emulate the behavior of idealized GECKO-A simulations for the SOA formation following the oxi-

Table 1. Environmental and chemistry parameters used for generating the GECKO-A box-model dataset.

Parameter	Range	Parameter	Range
Latitude [°]	−80–80	Relative humidity [%]	3–102
Temperature [K]	216–313	Atmospheric pressure [atm]	0.5–1.02
Preexisting aerosol seed [$\mu g\,m^{-3}$]	0.03–340	Initial NO_x [ppb]	10^{-4}–42
Initial precursor [ppb]	0–16	Initial CO [ppb]	33–1012
Initial O_3 [ppb]	1–100	NO Emission [$molec\,cm^{-2}\,s^{-1}$]	10^{7}–10^{9}

dation of three individual precursors reacting with OH under varied environmental conditions. While this work showed the ability of neural networks to reproduce idealized oxidation situations, it showed their limitations when extrapolated to realistic simulations with diurnally varying conditions. The results indicate that this type of system needs to be trained with a dataset representative of the conditions in which it will be applied.

In this paper, we train a random forest on a dataset constructed with multiple GECKO-A simulations, with the primary aim of predicting SOA mass from the oxidation of toluene and dodecane for realistic atmospheric conditions over a range of chemical regimes covering daytime and nighttime oxidation by the main oxidants (OH, O_3 and NO_3). Our objective is to build an empirical SOA model that is able to reproduce the aerosol mass that a complex, explicit mechanism would predict, at a numerical cost that is comparable to that of reduced chemical mechanisms currently used in large-scale models.

2 Methods

2.1 Reference Organic Chemistry Mechanisms

The Generator for Explicit Chemistry of Organics in the Atmosphere (GECKO-A, Aumont et al., 2005) is a software tool allowing the automatic generation of detailed multi-generational organic chemistry mechanisms. It is based on state-of-the-art knowledge of atmospheric organic chemistry and structure activity relationships (SAR) (*e.g.,* Atkinson, 1997; Raventos-Duran et al., 2010) to estimate unknown reaction kinetics and thermodynamics. It calculates the gas-particle partitioning of individual organic species based on estimates of their volatility (Valorso et al., 2011). In the present study, chemical mechanisms were generated for the oxidation by OH, NO_3 and O_3 of toluene (C_7H_8) and dodecane ($C_{12}H_{26}$), two compounds emitted by anthropogenic activities. Currently, GECKO-A only includes gas-phase oxidation and condensation of semi-volatile organic compounds. There are no heterogeneous processes and no aerosol phase processes (*e.g.,* oligomerization) included in the model. The resulting toluene full oxidation mechanism contains 8560 species involved in 47 349 chemical reactions, including 4536 gas-aerosol equilibrium reactions. The oxidation mechanism for dodecane was completed to the 4^{th} generation. This was done in order to reduce the generated mechanism to a manageable size for the purpose of this work. The resulting dodecane oxidation mechanism contains 75 745 species involved in 465 751 chemical reactions, including 20 968 gas-aerosol equilibrium reactions.

2.2 Dataset Construction

To create the dataset used to train and evaluate the random forest model, we ran a large set of simulations for toluene and dodecane. Each simulation is performed for 230 hours with a uniformly sampled set of randomly chosen initial conditions and external forcing (see Tab. 1). Temperature, relative humidity and atmospheric pressure are se-

lected in ranges typical of values found in the lower troposphere. This ensures that the SOA emulator will have the correct sensitivity to changes in these parameters through the effects of (i) temperature on reaction rates and SOA evaporation, (ii) relative humidity on OH formation and (iii) pressure on third-body reaction rates. Initial concentrations of precursors, NO_x and CO are randomly picked to cover a wide range of chemical regimes. The latitude is varied from 80°S to 80°N to ensure that the model does not fit to a specific diurnal cycle. All model simulations start at 10:00am UTC and simulate a diurnal light cycle defined by the chosen latitude. The model time-step length is 5 minutes. After initialization, the precursor, NO_x, O_3, and CO freely react without constraints. The other external forcing (temperature, relative humidity, pressure, latitude, NO emissions, seed) are maintained constant for the whole simulation. The simulated photochemistry leads to the multi-generational formation of semi- and low-volatile secondary organic compounds that can condense to form SOA.

2.2.1 Outcomes

In this work, the aim is to predict the distribution of organic species between gas and aerosol phases. In order to build a flexible approach that will allow future developments such as adding additional phases (*e.g.,* aqueous phase) and predicting organic matters properties (*e.g.,* solubility for deposition of organic vapors and particles), the first chosen outcome is the total organic mass m_t [$\mu g\,m^{-3}$]:

$$m_t = m_g + m_a \tag{1}$$

m_g and m_a are respectively the total gas- and particle-phase organic mass. The goal is to have only one outcome m_t in mass concentration units ($\mu g\,m^{-3}$) and predict the contributing phases to the total mass as fractions of this m_t. We arbitrarily chose to predict gaseous mass fraction γ ($m_g = \gamma \times m_t$). m_a can then be derived as $m_a = (1 - \gamma) \times m_t$.

Following the method of Keller and Evans (2019), we also established a variance criterion to decide whether the random forest should predict the value of the predictor or its trend. For stability and better performances, stable and unstable outcomes are identified. This classification is based on the standard deviation of the ratio between post- and pre- numerical time-step solve value. This ratio was calculated for each outcome and each time-step on the whole training dataset. The standard deviation of these ratios was calculated and if the value of this standard deviation is below a threshold of 0.07, the outcome is classified as stable and its trend is predicted. Otherwise, the outcome is unstable and its direct value is predicted. For the two outcomes used in this work, we found that the value of the total mass outcome m_t is unstable and needs to be directly predicted, while the gas phase fraction γ is stable and its trend is predicted.

2.2.2 Predictors

We selected the predictors based on parameters relevant to SOA formation (see Tab. 2). The concentration of the precursor as well as the main daytime oxidants (OH, O_3) and the aerosol seed concentrations have been chosen for their key role in SOA formation. The pressure and temperature modulate the kinetics that control gas-phase oxidation. Temperature is also very important for the condensation of vapors. The solar zenith angle and the photolysis rate of NO_2 represent the influence of the diurnal cycle on gas kinetics.

At each time-step of the GECKO-A simulations, predictors are computed before chemistry is integrated, and outcomes are computed after the solver has finished. 4000 simulations ($\approx 1 \times 10^6$ time-steps) where produced for each precursor, resulting in a total of 8000 simulations ($\approx 2 \times 10^6$ time-steps).

Table 2. List of predictors and outcomes used for training the random forests.

Predictors [units]	Outcomes [units] (prediction method)
Temperature [K]	Total organic mass [$\mu g\,m^{-3}$] (direct)
Water vapor concentration [molec/cm^3]	Organic gaseous fraction [dimensionless] (trend)
Pressure [atm]	
Solar zenith angle [deg]	
J_{NO_2} [s^{-1}]	
NO [molec/cm^3]	
NO_2 [molec/cm^3]	
s O_3 [molec/cm^3]	
OH [molec/cm^3]	
H_2O_2 [molec/cm^3]	
CH_2O [molec/cm^3]	
Aerosol seed mass [$\mu g\,m^{-3}$]	
Total organic mass [$\mu g\,m^{-3}$]	
Organic gaseous fraction [dimensionless]	
Organic aerosol fraction [dimensionless]	
Precursor [molec/cm^3]	

The collection of all predictor and outcome values at every time-step therefore constitutes a dataset representative of what the integration of an explicit chemical scheme would produce for a given predictor over one time-step. The random forest has to be recurrent to autonomously predict chemistry over multiple time-steps as both outcomes (*i.e.,* total organic mass and its gas phase fraction) are also predictors.

2.3 Random Forest Regression

We use Python libraries *scikit-learn* (Pedregosa et al., 2011) to fit the random forests, and *dask* (Dask Development Team, 2016) to handle parallelization of the code on the NCAR CISL supercomputers (Computational and Information Systems Laboratory, 2017).

Keller and Evans (2019) trained individual random forests for each of their outcomes, and integrated the random forest model within GEOS-Chem at each model time-step. We chose to train only one random forest that predicts all outcomes simultaneously, *i.e.,* each tree predicts a vector of values instead of predicting single values. This approach has the advantage of reducing model bias by implicitly accounting for correlations between outcomes. As the dataset contains values spanning many orders of magnitude, it was successively log-transformed, power-transformed and normalized to map the data as close as possible to a Gaussian distribution. The training dataset was shuffled prior to the regression procedure to avoid bias related to the random forest learning a specific diurnal cycle.

Hyperparameters for the random forest were tuned automatically during the random forest training with the *scikit-learn* library. The number of decision trees is the most important hyperparameter because it impacts both the numerical cost of running the random forest as well as the quality of the random forest. Random forest's training configurations were tested with 10, 50, 75 and 100 decision trees. The random forest hyperparameters optimization consistently selected 50 trees, which is the same number of decision trees that were selected in Keller and Evans (2019).

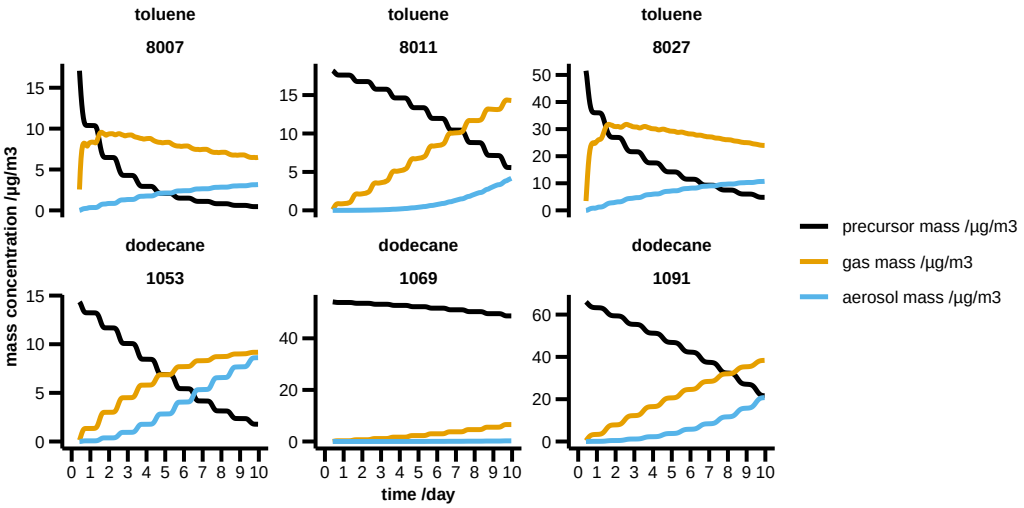

Figure 1. Time evolution of precursor oxidation, organic gas and organic aerosol formation from toluene and dodecane oxidation for representative training simulations.

3 Results

3.1 Training Dataset Characterization

Fig. 1 depicts the typical time evolution of a few randomly selected simulations from the validation dataset. As expected, the precursor is progressively oxidized during the ten days of the simulation. The kinetics of this decay depends on the concentration of the precursors' oxidants: OH (daytime) and NO_3 (nighttime) for dodecane, OH, O_3 (daytime) and NO_3 (nighttime) for toluene. Given that concentrations of these oxidants on randomly selected initial and environmental conditions in each simulation, the decay kinetics vary for each simulation. The oxidation of the precursor leads to the progressive formation of gaseous organic compounds. Depending on the availability of oxidants, the formation of these secondary organic compounds can peak early in the simulation as in simulations 8007 and 8027 in Fig. 1. On the other hand, the evolution displays a characteristic stepwise diurnal profile, with the organic mass increasing during daytime when photochemistry can take place. After the peak of the quicker oxidation simulations, the total organic mass decreases because the oxidation products are ultimately lost to the terminal CO_2 formation step. As their oxidation progresses, the secondary gaseous compounds become more oxidized and are able to condense onto the pre-existing aerosol seed, forming secondary organic aerosol (SOA). SOA formation therefore highly depends on the availability of oxidants. For instance, simulation 1069 displays a typical case of a slow precursor decay causing the slow formation of secondary organic compounds with almost no SOA production.

As shown in Fig. 1 on a few sample simulations, this work is aimed at reproducing a large variety of situations, with the SOA formation behavior that non-linearly depends on multiple parameters. Figure 2 depicts the distribution of SOA mass yield as a function of key parameters describing the chemical regimes controlling SOA formation.

The precursor controls the total amount of organic carbon that is available to form SOA. Both dodecane and toluene SOA yields are not constant as a function of the precursor concentration. This is a typical illustration of the non-linearity of SOA formation and atmospheric chemistry in general. As precursor mixing ratios increase, the precursor becomes a significant competitor for oxidants, slowing the formation of later gener-

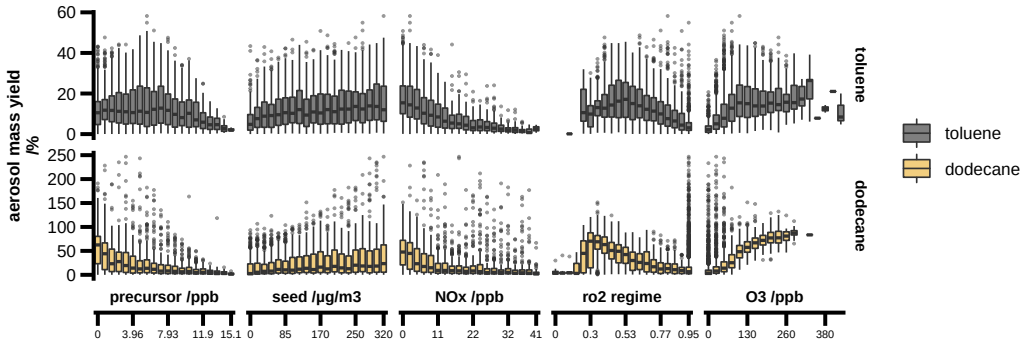

Figure 2. Dependence of average organic aerosol mass and aerosol mass yield as a function of the chemical environment.

ations organic compounds that are more likely to efficiently contribute to SOA. This hypothesis is confirmed by the fact that the dodecane SOA yield decay starts for lower precursor mixing ratios, which is explained by the faster reaction rate of dodecane with the main oxidant OH compared to toluene ($k_{\text{dodecane + OH,298}} = 1.32{\times}10^{-11}\,\text{cm}^3/\text{molec/s}$ vs. $k_{\text{toluene + OH,298}} = 5.6 \times 10^{-12}\,\text{cm}^3/\text{molec/s}$, Mellouki et al., 2021).

As the seed concentrations were selected over a wide range (see Tab. 1), the limiting effect of low pre-existing seeds can only be seen in the lowest chosen concentrations, below $40\,\mu\text{g}\,\text{m}^{-3}$. In the rest of the range of seed concentrations, SOA yield has identical distribution probability. In our simulations, as the nature of this seed is not accounted for, the fact that seed is present is enough to trigger SOA condensation and it is rarely limiting.

The NO_x mixing ratios control the formation of ozone and OH through the photolysis of NO_2 and the reaction of ozone with NO. However the relationship of NO_x with oxidants levels is not trivial and also depends on the concentrations of organic compounds. Here, the simulated higher SOA yields at lower NO_x levels could be explained by the role of NO_x on the oxidation of organic compounds. After the initial oxidation step forming a peroxy radical ($RH + OH \overset{+O_2}{\rightarrow} RO_2 + H_2O$), the peroxy radical can react with NO to form an alkoxy radical that can fragment, leading to more volatile compounds that are less likely to form SOA. If NO concentration is low enough, peroxy radicals are more likely to react with HO_2 and other peroxy radicals to form more oxidized species that are more likely to form SOA. This can explain the higher SOA yields at lower NO_x shown on Fig. 2. This effect is better seen after defining the RO_2 regime R as:

$$R = \frac{k_{\text{RO}_2\text{ + NO}} \times \text{NO}}{k_{\text{RO}_2\text{ + NO}} \times \text{NO} + k_{\text{RO}_2\text{ + HO}_2} \times \text{HO}_2} \tag{2}$$

where $k_{\text{RO}_2\text{ + NO}} = 7.7{\times}10^{-12}\,\text{cm}^3\,\text{molec}^{-1}\,\text{s}^{-1}$, $k_{\text{RO}_2\text{ + HO}_2} = 5.1{\times}10^{-12}\,\text{cm}^3\,\text{molec}^{-1}\,\text{s}^{-1}$. This ratio indicates which pathway is favored for RO_2 radicals: when $R = 1$, they only react with NO and when $R = 0$ they never react with NO. For dodecane, Fig. 2 shows that from $R = 0.4$ to $R = 1$, the median value of the SOA yield decreases from 60% to less than 1%. Below $R = 0.4$, the SOA yield decreases down to less than 1% at $R = 0.15$. This low yield for low R values can be explained by low levels of oxidants limiting SOA production in very low NO_x situations. For toluene, this peak in SOA yield happens around $R = 0.5$ with a median value of 19 %. It is less marked than for dodecane because the impact of the RO_2 + NO reaction pathway on fragmentation is lower on cyclic and shorter molecules like toluene and its oxidation products (Aumont et al., 2013).

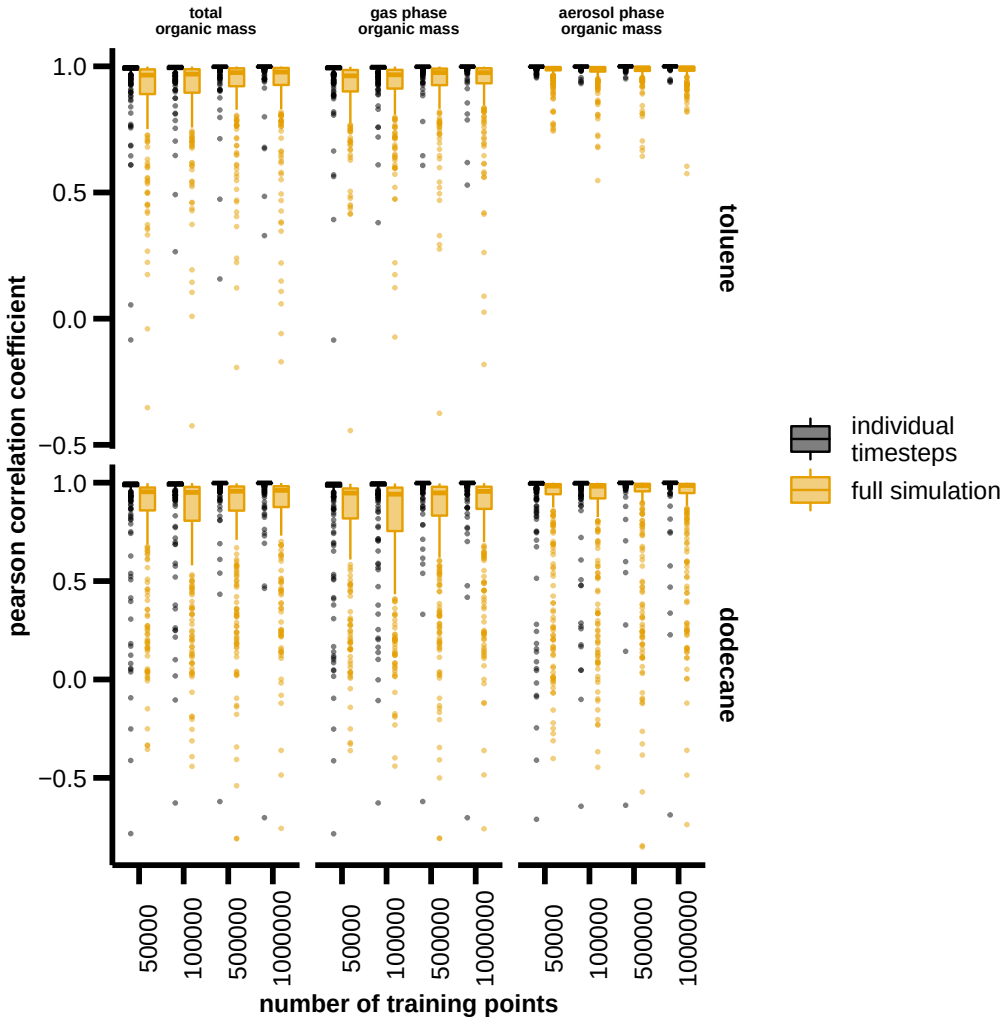

Figure 3. Boxplot distribution of Pearson correlation coefficients for testing individual time-step predictions (black) and full simulation runs (yellow) for toluene and dodecane simulations. The middle lines of the boxplots are the median, the top and bottom of the boxes denote the 1st and 3rd quartiles and the whiskers extend to the 5th and 95th percentiles of the distribution.

For both precursors, the SOA yield increases with ozone mixing ratios. Higher ozone mixing ratios are associated with higher OH concentrations, which can explain higher SOA yields.

3.2 Training Dataset Size

Two kinds of tests are presented for each validation simulation. First, the random forest was tested on each time-step individually by the reference predictors as inputs and comparing the random forest output with the reference outcomes. The resulting Pearson correlation coefficients (r) distributions are shown on Fig. 3 (black boxplots), as a function of the number of points used for training the random forest. With the exception of a few outliers, the random forest is very accurate to predict outcomes on a single time-step, even with only 50 000 training points. The Pearson correlation coefficient displays median values around $r = 0.99$ for both toluene and dodecane.

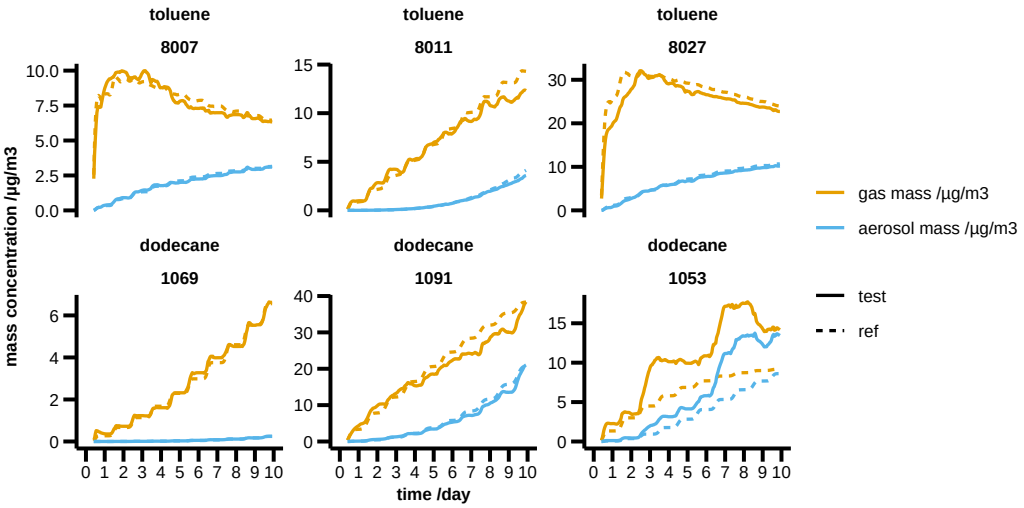

Figure 4. Time-series of three toluene and three dodecane sample experiments, comparing the reference (dashed lines) and the random forest recurrent simulations (continuous lines) for the predicted organic gas (black) and aerosol (orange) mass.

However, for practical application the random forest needs to be run recurrently. This means that for the second test in Fig. 3 (orange boxplots), as well as the rest of this paper, the random forest is constrained with the initial conditions and the environmental conditions from the reference simulation. At the end of each time-step, the predicted outcomes are used in the input predictors for the next time-step. In these validation tests, the recurrent random forest model accumulates errors with time. This is reflected in the median r values decreasing compared to individual time-steps tests to approximately $r = 0.98$ for toluene and $r = 0.96$ for dodecane. The r values are also more spread with the interquartile range for 100 000 points increasing from 0.002 to 0.05 for toluene and from 0.003 to 0.09 for dodecane. Increasing the number of points used to train the random forest slightly improve the r scores. For toluene, using 500 000 or 1 000 000 points provides similar performances while for dodecane, using 1 000 000 points still increases r and reduces the interquartile range. We could not test higher numbers of training points due to the limited size of the created dataset, but it seems that above 500 000 training points, the gains are marginal at best for the considered precursors.

3.3 Sample Simulations Tests

To illustrate the behavior of the recurrent random forest model, Fig. 4 displays the recurrent random forest results on the same sample simulations that were shown on Fig. 1. The associated relative errors on predicted aerosol mass are shown on Fig. 5. The random forest is able to reproduce the timeseries of gas and aerosol mass in all but one of the examples (simulation 1053). For these simulations, the random forest can reproduce the typical stepwise daytime growth of organic mass of the slower oxidation simulations (simulation 1069, 1091 and 8011), as well as reproducing the peaking growth of organic mass for faster oxidation simulations (8007 and 8027). For all simulations (except 1053), the relative error tends to be the highest for the first five days of the simulations, converging towards errors lower than 10% for the last five days. Finally, for the worst random forest simulation in this sample (1053), the model exhibits errors around ±100% after 2 days and cannot recover from the accumulated errors. The relative error remains between 50 % and 100 % but not producing any unrealistic mass concentrations.

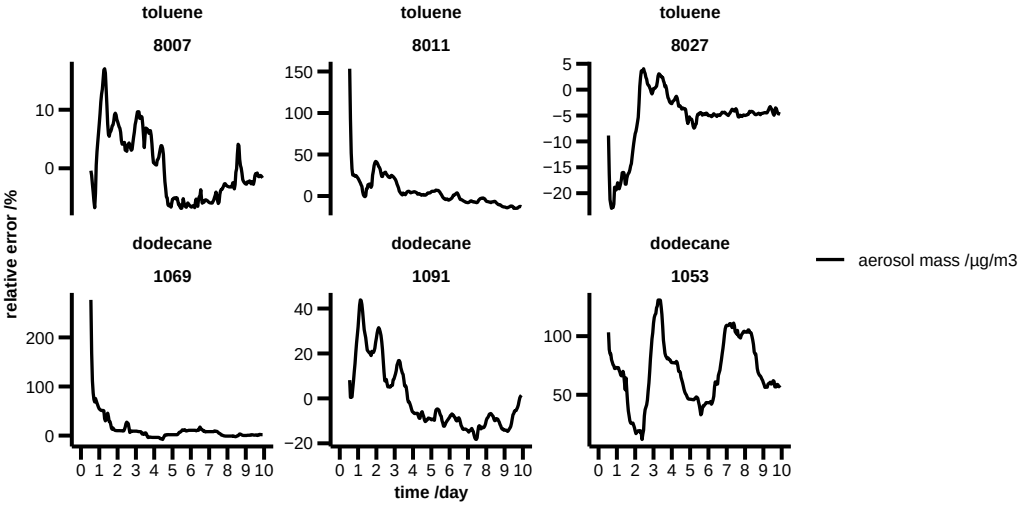

Figure 5. Time-series of the relative error on predicted aerosol mass for three toluene and three dodecane sample experiments shown on Fig. 4.

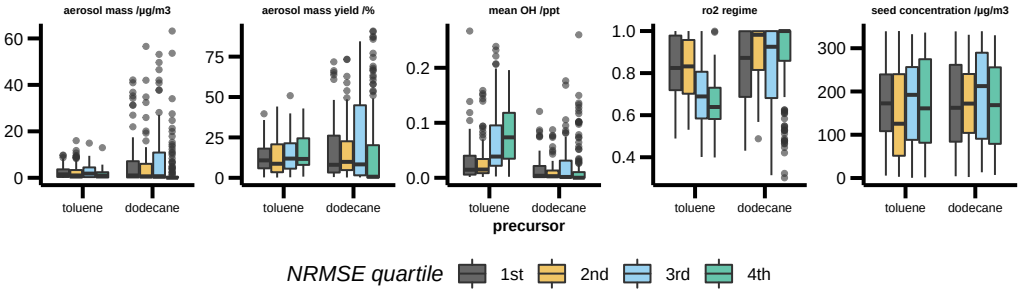

Figure 6. Boxplot distribution of simulations properties, according to their NRMSE quartile.

3.4 Errors Distribution

In order to identify the type of situations where the random forest is not able to reproduce the explicit model behaviour, we examine the distribution of the normalized root mean square error (NRMSE) defined below:

$$\mathrm{NRMSE} = \frac{\sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}}}{Q_3 - Q_1} \tag{3}$$

y_i and $\hat{y}_i$ are the i^{th} reference and predicted aerosol mass respectively, N is the number of time-steps and $Q_3 - Q_1$ is the difference between the first and the third quartiles. The validations random forest simulations were split in four categories, depending on their NRMSE. The distributions of environmental conditions according to this split are displayed on Fig. 6.

First the existing seed concentration distribution does not vary with the quality of the random forest simulations. As was shown above on Fig. 2, the aerosol mass yields dependence on seed concentration is low, which explains why the random forest perfor-

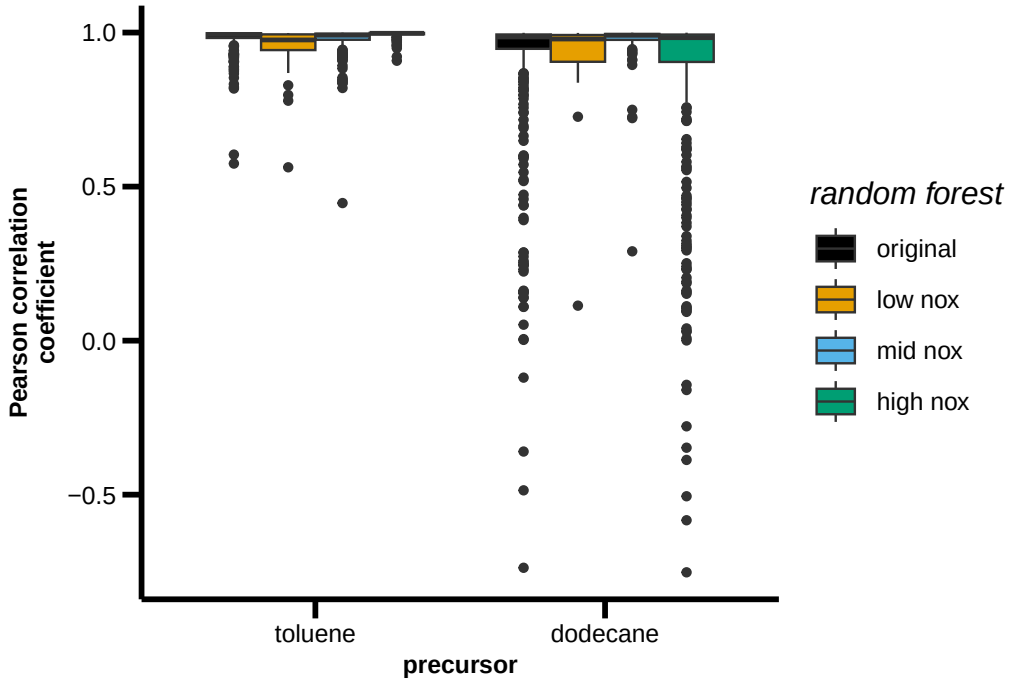

Figure 7. Boxplots of the Pearson correlation coefficients distribution for the toluene and dodecane validation simulations for the original random forests (black) and the three specialized random forests: low NO_x (orange), mid NO_x (blue) and high NO_x (green).

mances are not sensitive to this variable. As long as some seed aerosol is available, the system is not sensitive to its mass concentrations.

For toluene, the lower quality random forest simulations (third and fourth NRMSE quartiles) are typically described with lower NO_x regimes: on average $R \approx 0.82$ for the first two NRMSE quartiles, compared to $R = 0.7$ and $R = 0.65$ for the last two NRMSE quartiles. The third and fourth NRMSE quartiles simulations also exhibit higher OH mixing ratios: $OH_{median}^{3^{rd}quartile} = 0.04$ ppt and $OH_{median}^{4^{th}quartile} = 0.07$ ppt compared to $OH_{median}^{1^{st}quartile} = 0.01$ ppt for the first quartile. The aerosol mass yields ($\approx$ 12 − 13 %) are similar for all quartiles. The non-linear dependence of the SOA yield on the RO_2 regime (Fig. 2) seems to be the determining factor for toluene simulations. Under-representing lower RO_2 regimes in the training dataset therefore has a strong impact on the random forest performances. In our case, this under-representation is the likely the consequence of the simple random selection of the training simulations environmental conditions.

Dodecane lower quality simulations are heavily skewed towards simulations with high NO_x regimes ($R_{median}^{4^{th}quartile} = 1$ vs $R_{median}^{1^{st}quartile} = 0.87$) and lower OH mixing ratio ($OH_{median}^{4^{th}quartile} = 8.8 \times 10^{-5}$ ppt vs $OH_{median}^{1^{st}quartile} = 3.7 \times 10^{-3}$ ppt) and lower aerosol mass yields ($Y_{median}^{4^{th}quartile} = 0.68$ % vs $Y_{median}^{1^{st}quartile} = 8.1$ %). This behaviour difference between toluene and dodecane may be explained by the number of dodecane training simulations available for higher NO_x regimes, which include many outliers (1236 outliers out of 2568 point for the highest RO_2 regime bin on Fig. 2) in terms of aerosol yield. The random forest is therefore not able to reproduce the complex behaviour of dodecane SOA formation in this specific regime. In this case, the complexity of the dodecane oxidation for very high NO_x situations cannot be properly reproduced by the random forest with the given training dataset.

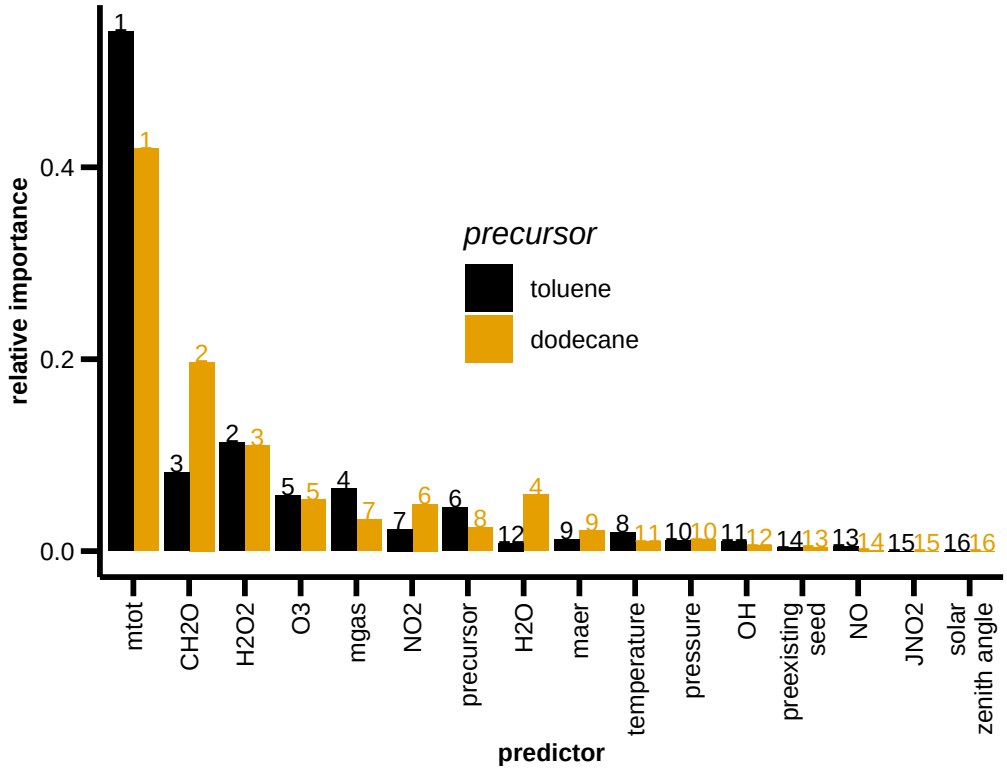

Figure 8. Predictors relative importance for the toluene (black) and dodecane (orange) random forests. The rank of each predictor is indicated at the top of the bars.

A possible way to improve the ability of the system to reproduce the complex relationship between R and SOA formation is to create independent random forests specialized for specific RO_2 regimes. To test this hypothesis, the training dataset was split in three separate sets according to the initial RO_2 regimes: a low NO_x set ($R < 0.3$, 100 280 training points for toluene, 72 200 points for dodecane), a mid NO_x set ($0.3 < R < 0.7$, 594 780 points for toluene,268 640 points for dodecane) and a high NO_x set ($R > 0.7$, 363 170 points for toluene, 650 900 points for dodecane). Three specialized random forests were therefore trained on these three datasets for each precursor.

Figure 7 displays the distribution of Pearson correlation coefficients for the toluene and dodecane validation simulations for each specialized random forest compared to the original random forest trained with 1 000 000 points. For toluene simulations, all the specialized random forests display similar performances to the original random forest: for low NO_x, $r_{\text{median}} = 0.98$, for mid NO_x, $r_{\text{median}} = 0.99$ and for high NO_x,$r_{\text{median}} = 1.0$. Similarly for dodecane simulations, all specialized random forests exhibit performances similar to the original one, with median r values ranging frm 0.98 to 0.99.

It is likely that the potential improvement caused by specializating the random forests over different RO_2 regimes is negatively compensated by performance reduction caused by a lower number number of training points. Furthermore, in the dodecane case, specializing a random forest for high NO_x does not have a significant impact on the number of validation outliers in this regime.

3.5 Predictors Importance

After training the random forest, it is possible to estimate the relative importance of the chosen predictors by counting their occurrences as threshold criteria in individual nodes (Fig. 8). For both toluene and dodecane, the most important predictor for organic mass is the total organic mass in the previous time-step. This finding is consistent with the fact that total organic mass is one of the predicted outcomes by the random forests. CH_2O and H_2O_2 are the second and third (resp. third and second) most important predictors for dodecane (resp. toluene). Since higher H_2O_2 concentrations are indicative of low NO_x situations, H_2O_2 can be considered as a proxy for the RO_2 regime. CH_2O is the only predictor related to secondary organic gaseous species and could be interpreted as a proxy for organic gases formation.

The water vapor concentration and organic gaseous fraction (m_g) are the fourth most important predictors for dodecane and toluene respectively. Since the random forests are predicting the trend of m_g, it is logical that its previous step value is a significant predictor. The importance of water vapor is likely related to its role in OH production.

O_3 is the fifth most important predictor for both precursors. The information brought by the ozone predictor is related to general oxidants levels, the diurnal cycle as well as the RO_2 regime. NO_2 is the sixth most important predictor for dodecane while it is the precursor's concentration for toluene. These two predictors provide information about the diurnal cycle, the oxidants levels as well as the potential for secondary organic matter production. The seventh is the organic gaseous mass fraction for dodecane and NO_2 for toluene. The precursor's concentration and temperature are the eighth most important predictors for dodecane and toluene respectively. For both precursors, the organic aerosol fraction (m_a) is the ninth most important predictor. Because the gaseous mass fraction is a more important predictor for both precursors, m_a only provides complementary information. The remaining predictors only have negligible contributions to the random forests. Since we've shown that SOA formation is not sensitive to pre-existing particle seed (see Fig. 2) it is logical that this predictor is not important. Predictors directly related to the diurnal cycle (J_{NO_2} and solar zenith angle) are unimportant here, meaning that there is enough information provided by the time evolution of ozone and precursor concentrations to control for daytime vs. nighttime organic matter production. Similarly, the precursor decay as well as ozone concentrations give enough information related to oxidant concentrations and RO_2 regime, explaining the weak importance of OH and NO predictors.

Since the contributions of the various predictors are dominated by only a few of them, we trained new random forests only using the 8 most important predictors for each precursor: m_t, CH_2O, H_2O_2, H_2O, O_3, NO_2 , m_g and the precursor concentration for dodecane and m_t, H_2O_2, CH_2O, m_g, O_3, the precursor concentration, NO_2 and temperature for toluene. Figure 9 compares the Pearson correlation coefficients (r) and NRMSE scores calculated for each validation simulation (359 for each precursor) of the random forest trained with the 16 original predictors with the random forest trained with the eight most important identified predictors (see Fig. 8).

Reducing the number of predictors for dodecane improves r for 239 (67 %) validation simulations, with an average r increase of 272 %. The NRMSE decreased for 212 (59 %) dodecane validation simulations, with an average NRMSE reduction of 59 %. For 160 (45 %) of the dodecane validation simulations, both r and NRMSE are improved. Reducing the number of predictors leads to an improvement of r for 212 (62 %) toluene validation simulations, with an average r increase of 2 %. The NRMSE decreased for 190 (53 %) toluene validation simulations, with an average NRMSE reduction of 40 %. For 148 (41 %) of the toluene validation simulations, both r and NRMSE are improved.

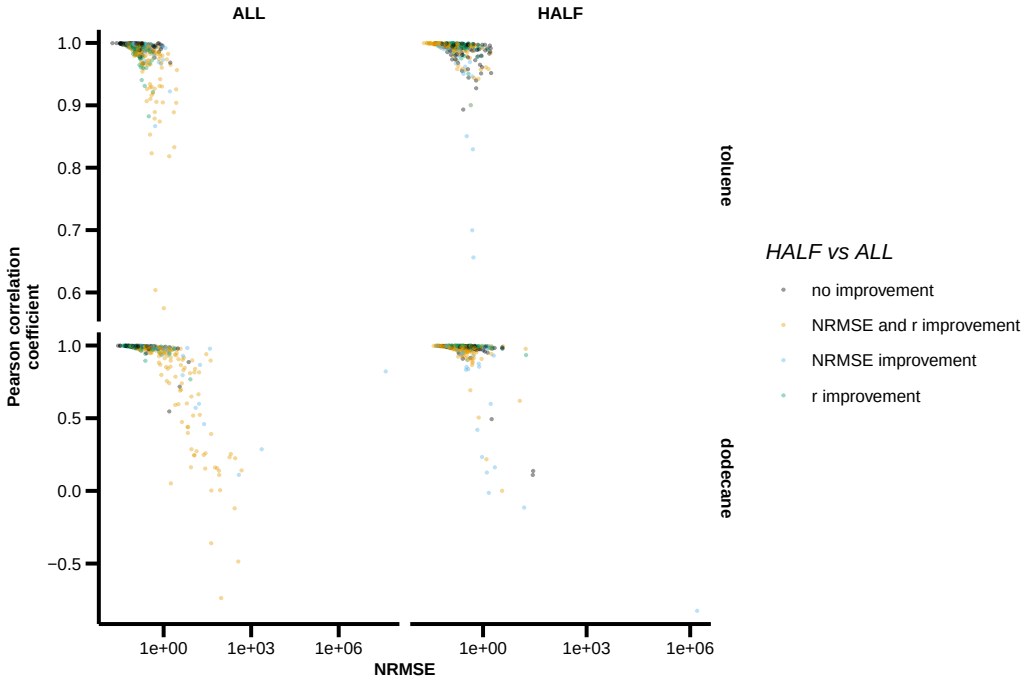

Figure 9. Scatterplot of the validation simulations' Pearson correlation coefficients (r) as a function of their NRMSE for toluene (top row) and dodecane (bottom row), for the random forests trained with the 16 originally selected predictors (ALL, left column) and the random forests trained with only the eight most important predictors (HALF, left column). The colors depict whether r, NRMSE or both scores are improved when reducing the number of predictors.

As shown in Fig. 9, reducing the number of predictors is beneficial for the worst performing simulations, especially for dodecane. For both toluene and dodecane, the majority of validation simulations are improved when halving the number of predictors. However, 7 out of the 8 selected predictors are shared by both random forests. The relative importance and ranks of the predictors differ between both random forests. There is therefore no guarantee that different precursors have the same optimal number of predictors and that they share the same predictors.

4 Conclusions

In this work we trained two recurrent random forests to predict organic mass production in both gas and aerosol phases resulting from toluene and dodecane oxidation. The random forests were trained on a dataset created with the GECKO-A explicit organic chemistry box model. The dataset contains a series of single box-model simulations covering a wide range of environmental conditions to ensure that the resulting random forests are able to reproduce the complex relations between organic aerosol production and the chemical environment. The resulting random forests show very good performances in predicting organic mass evolution in varied conditions when tested on a similar random set of box-model simulations. The distribution of errors in testing simulations highlights however the importance of carefully preparing the training dataset. Our results suggest that random sampling over a range of possible environmental conditions is insufficient to build a robust training dataset, and that is is more important to properly sample a range of more complex chemical parameters such as the RO_2 regimes (R).

We have shown that the range of R that needs focus depends on the precursors. For instance, the dodecane random forest has weaker performance for high R whereas the toluene random forest has lower performance for medium R values. However, creating multiple random forests each trained over a smaller range of R does not lead to more robust results than a single random forest. It seems more efficient to add additional training data points for the poorly performing RO_2 regimes.

The selection of predictors is also a crucial step. We have shown that it is possible to increase the random forest performance by reducing the number of predictors to the most important ones. However, there is no reason to think that these predictors have to be the same for different precursors, highlighting the care that must be taken in their selection. In this work, we selected the most important predictors by first training a random forests with a wide selection of predictors, and then training a new random forest with only the most important predictors identified in the first random forest.

In this work, we have therefore shown the feasibility of building random forests that behave like a detailed chemical mechanism for predicting secondary organic mass and its partitioning between gas and particle phases. There are still some limitations to overcome before the implementation of the random forest SOA emulator within a chemistry-climate model. First, even if the random forests are performing well, there are still some critical outliers at the validation stage (*e.g.,* high NO_x dodecane). More work needs to be focused at removing these outliers because falling in one of these bad cases in a 3D model run would likely make the full simulation diverge. Second, the random forests were each trained to reproduce the oxidation of a single precursor. Additional studies are required to quantify whether it is important to represent the interactions of multiple primary hydrocarbons, their competition for oxidants, and the impact on the resulting SOA formation.

Open Research Section

The dataset that was constructed to train and test the random forests and the Python code implementing the random forest training and testing have been uploaded on Zenodo (Mouchel-Vallon & Hodzic, 2022).

Acknowledgments

The authors thank John S. Schreck, David John Gagne and Prasad Kasibhatla for their helpful advice on training random forests.

References

Atkinson, R. (1997). Gas-Phase Tropospheric Chemistry of Volatile Organic Compounds: 1. Alkanes and Alkenes. *Journal of Physical and Chemical Reference Data*, *26*(2), 215. doi: 10.1063/1.556012

Aumont, B., Camredon, M., Mouchel-Vallon, C., La, S., Ouzebidour, F., Valorso, R., ... Madronich, S. (2013, October). Modeling the influence of alkane molecular structure on secondary organic aerosol formation. *Faraday Discussions*, *165*(0), 105–122. doi: 10.1039/c3fd00029j

Aumont, B., Szopa, S., & Madronich, S. (2005, September). Modelling the evolution of organic carbon during its gas-phase tropospheric oxidation: Development of an explicit model based on a self generating approach. *Atmospheric Chemistry and Physics*, *5*(9), 2497–2517. doi: 10.5194/acp-5-2497-2005

Aumont, B., Valorso, R., Mouchel-Vallon, C., Camredon, M., Lee-Taylor, J., & Madronich, S. (2012, August). Modeling SOA formation from the oxidation of intermediate volatility n-alkanes. *Atmospheric Chemistry and Physics*, *12*(16), 7577–7589. doi: 10.5194/acp-12-7577-2012

Camredon, M., Aumont, B., Lee-Taylor, J., & Madronich, S. (2007, November). The SOA/VOC/NOx system: An explicit model of secondary organic aerosol formation. *Atmospheric Chemistry and Physics*, *7*(21), 5599–5610. doi: 10.5194/acp-7-5599-2007

Cappa, C. D., & Wilson, K. R. (2012, October). Multi-generation gas-phase oxidation, equilibrium partitioning, and the formation and evolution of secondary organic aerosol. *Atmospheric Chemistry and Physics*, *12*(20), 9505–9528. doi: 10.5194/acp-12-9505-2012

Computational and Information Systems Laboratory. (2017). *Cheyenne: HPE/SGI ICE XA System (NCAR Community Computing)*. Boulder, CO: National Center for Atmospheric Research. doi: 10.5065/D6RX99HX

Couvidat, F., Debry, É., Sartelet, K., & Seigneur, C. (2012). A hydrophilic/hydrophobic organic (H2O) aerosol model: Development, evaluation and sensitivity analysis. *Journal of Geophysical Research: Atmospheres*, *117*(D10). doi: 10.1029/2011JD017214

Dask Development Team. (2016). *Dask: Library for dynamic task scheduling.*

Donahue, N. M., Epstein, S. A., Pandis, S. N., & Robinson, A. L. (2011). A two-dimensional volatility basis set: 1. organic-aerosol mixing thermodynamics. *Atmospheric Chemistry and Physics*, *11*(7), 3303–3318. doi: 10.5194/acp-11-3303-2011

Donahue, N. M., Robinson, A. L., Stanier, C. O., & Pandis, S. N. (2006, April). Coupled Partitioning, Dilution, and Chemical Aging of Semivolatile Organics. *Environmental Science & Technology*, *40*(8), 2635–2643. doi: 10.1021/es052297c

Foley, K. M., Roselle, S. J., Appel, K. W., Bhave, P. V., Pleim, J. E., Otte, T. L., ... Bash, J. O. (2010, March). Incremental testing of the Community Multiscale Air Quality (CMAQ) modeling system version 4.7. *Geosci. Model Dev.*, *3*(1), 205–226. doi: 10.5194/gmd-3-205-2010

Hodzic, A., Madronich, S., Kasibhatla, P. S., Tyndall, G., Aumont, B., Jimenez, J. L., ... Orlando, J. (2015, August). Organic photolysis reactions in tropospheric aerosols: Effect on secondary organic aerosol formation and lifetime. *Atmospheric Chemistry and Physics*, *15*(16), 9253–9269. doi: 10.5194/acp-15-9253-2015

Jenkin, M. E., Saunders, S. M., Wagner, V., & Pilling, M. J. (2003, February). Protocol for the development of the Master Chemical Mechanism, MCM v3 (Part B): Tropospheric degradation of aromatic volatile organic compounds. *Atmospheric Chemistry and Physics*, *3*(1), 181–193. doi: 10.5194/acp-3-181-2003

Keller, C. A., & Evans, M. J. (2019). Application of random forest regression to the calculation of gas-phase chemistry within the GEOS-Chem chemistry model v10. *Geoscientific Model Development*, *12*, 1209–1225. doi: 10.5194/gmd-12-1209-2019

Kelp, M. M., Jacob, D. J., Kutz, J. N., Marshall, J. D., & Tessum, C. W. (2020). Toward Stable, General Machine-Learned Models of the Atmospheric Chemical System. *Journal of Geophysical Research: Atmospheres*, *125*(23), e2020JD032759. doi: 10.1029/2020JD032759

Kelp, M. M., Jacob, D. J., Lin, H., & Sulprizio, M. P. (2022). An Online-Learned Neural Network Chemical Solver for Stable Long-Term Global Simulations of Atmospheric Chemistry. *Journal of Advances in Modeling Earth Systems*, *14*(6), e2021MS002926. doi: 10.1029/2021MS002926

Kourtchev, I., Godoi, R. H. M., Connors, S., Levine, J. G., Archibald, A. T., Godoi, A. F. L., ... Kalberer, M. (2016, September). Molecular composition of organic aerosols in central Amazonia: An ultra-high-resolution mass spectrometry study. *Atmospheric Chemistry and Physics*, *16*(18), 11899–11913. doi: 10.5194/acp-16-11899-2016

La, Y. S., Camredon, M., Ziemann, P. J., Valorso, R., Matsunaga, A., Lannuque,

V., ... Aumont, B. (2016, February). Impact of chamber wall loss of gaseous organic compounds on secondary organic aerosol formation: Explicit modeling of SOA formation from alkane and alkene oxidation. *Atmospheric Chemistry and Physics*, *16*(3), 1417–1431. doi: 10.5194/acp-16-1417-2016

Lannuque, V., Camredon, M., Couvidat, F., Hodzic, A., Valorso, R., Madronich, S., ... Aumont, B. (2018, September). Exploration of the influence of environmental conditions on secondary organic aerosol formation and organic species properties using explicit simulations: Development of the VBS-GECKO parameterization. *Atmospheric Chemistry and Physics*, *18*(18), 13411–13428. doi: 10.5194/acp-18-13411-2018

Lannuque, V., Couvidat, F., Camredon, M., Aumont, B., & Bessagnet, B. (2020). Modeling organic aerosol over Europe in summer conditions with the VBS-GECKO parameterization: Sensitivity to secondary organic compound properties and IVOC (intermediate-volatility organic compound) emissions. *Atmospheric Chemistry and Physics*, *20*(8), 4905–4931. doi: 10.5194/acp-20-4905-2020

Lee-Taylor, J., Hodzic, A., Madronich, S., Aumont, B., Camredon, M., & Valorso, R. (2015). Multiday production of condensing organic aerosol mass in urban and forest outflow. *Atmospheric Chemistry and Physics*, *15*(2), 595–615. doi: 10.5194/acp-15-595-2015

Li, J., Cleveland, M., Ziemba, L. D., Griffin, R. J., Barsanti, K. C., Pankow, J. F., & Ying, Q. (2015). Modeling regional secondary organic aerosol using the Master Chemical Mechanism. *Atmospheric Environment*, *102*(February), 52–61. doi: 10.1016/j.atmosenv.2014.11.054

Liao, Q., Zhu, M., Wu, L., Pan, X., Tang, X., & Wang, Z. (2020, September). Deep Learning for Air Quality Forecasts: A Review. *Current Pollution Reports*, 1–11. doi: 10.1007/s40726-020-00159-z

Mellouki, A., Ammann, M., Cox, R. A., Crowley, J. N., Herrmann, H., Jenkin, M. E., ... Wallington, T. J. (2021, March). Evaluated kinetic and photochemical data for atmospheric chemistry: Volume VIII – gas-phase reactions of organic species with four, or more, carbon atoms ($\geq$ C₄). *Atmos. Chem. Phys.*, *21*(6), 4797–4808. doi: 10.5194/acp-21-4797-2021

Menut, L., Bessagnet, B., Khvorostyanov, D., Beekmann, M., Blond, N., Colette, A., ... Vivanco, M. G. (2013, July). CHIMERE 2013: A model for regional atmospheric composition modelling. *Geosci. Model Dev.*, *6*(4), 981–1028. doi: 10.5194/gmd-6-981-2013

Mouchel-Vallon, C., & Hodzic, A. (2022, November). *Towards emulating an explicit organic chemistry mechanism with a random forest model: Dataset and training code.* Zenodo. doi: 10.5281/zenodo.7327053

Mouchel-Vallon, C., Lee-Taylor, J., Hodzic, A., Artaxo, P., Aumont, B., Camredon, M., ... Madronich, S. (2020). Exploration of oxidative chemistry and secondary organic aerosol formation in the Amazon during the wet season: Explicit modeling of the Manaus urban plume with GECKO-A. *Atmospheric Chemistry and Physics*, *20*(10), 5995–6014. doi: 10.5194/acp-20-5995-2020

Pai, S. J., Heald, C. L., Pierce, J. R., Farina, S. C., Marais, E. A., Jimenez, J. L., ... Vu, K. (2020, March). An evaluation of global organic aerosol schemes using airborne observations. *Atmospheric Chemistry and Physics*, *20*(5), 2637–2665. doi: 10.5194/acp-20-2637-2020

Pankow, J. F., Marks, M. C., Barsanti, K. C., Mahmud, A., Asher, W. E., Li, J., ... Kleeman, M. J. (2015). Molecular view modeling of atmospheric organic particulate matter: Incorporating molecular structure and co-condensation of water. *Atmospheric Environment*, *122*, 400–408. doi: 10.1016/j.atmosenv.2015.10.001

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ...

Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Raventos-Duran, T., Camredon, M., Valorso, R., Mouchel-Vallon, C., & Aumont, B. (2010, August). Structure-activity relationships to estimate the effective Henry's law constants of organics of atmospheric interest. *Atmospheric Chemistry and Physics*, *10*(16), 7643–7654. doi: 10.5194/acp-10-7643-2010

Saunders, S. M., Jenkin, M. E., Derwent, R. G., & Pilling, M. J. (2003, February). Protocol for the development of the Master Chemical Mechanism, MCM v3 (Part A): Tropospheric degradation of non-aromatic volatile organic compounds. *Atmospheric Chemistry and Physics*, *3*(1), 161–180. doi: 10.5194/acp-3-161-2003

Schreck, J. S., Becker, C., Gagne, D. J., Lawrence, K., Wang, S., Mouchel-Vallon, C., ... Hodzic, A. (2022, August). Neural network emulation of the formation of organic aerosols based on the explicit GECKO-A chemistry model. *J Adv Model Earth Syst*. doi: 10.1029/2021MS002974

Valorso, R., Aumont, B., Camredon, M., Raventos-Duran, T., Mouchel-Vallon, C., Ng, N. L., ... Madronich, S. (2011, July). Explicit modelling of SOA formation from α-pinene photooxidation: Sensitivity to vapour pressure estimation. *Atmospheric Chemistry and Physics*, *11*(14), 6895–6910. doi: 10.5194/acp-11-6895-2011