A machine learning-guided adaptive algorithm to reduce the computational cost of atmospheric chemistry in Earth System models: application to GEOS-Chem versions 12.0.0 and 12.9.1

Lu Shen[1,2], Daniel J. Jacob[2], Mauricio Santillana[3,4], Kelvin Bates[2], Jiawei Zhuang[2], Wei Chen[5]

[1]Department of Atmospheric and Oceanic Sciences, School of Physics, Peking University, Beijing, China
[2]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA
[3]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA
[4]Department of Pediatrics, Harvard Medical School, Boston, MA, USA
[5]Center for Functional Nanomaterials, Brookhaven National Laboratory, Upton, NY 11973, USA

Correspondence to: Lu Shen (lshen@fas.harvard.edu)

Abstract. Inclusion of comprehensive atmospheric chemistry in Earth system models has been limited by the computational cost of integrating chemical mechanisms with typically over 100 coupled species. Here we present an adaptive algorithm to ease this computational bottleneck with no significant loss in accuracy, and apply it to the GEOS-Chem global 3-D model for tropospheric and stratospheric chemistry (228 species, 724 reactions). Our approach is inspired by unsupervised machine learning clustering techniques and traditional asymptotic analysis ideas. We first partition the species from the full mechanism into 13 blocks, using a machine learning approach that analyzes the chemical linkages between species and the production and loss rates. Building on these blocks, we pre-select 20 submechanisms, as defined by unique assemblages of the species blocks, and then pick locally and on the fly which submechanism to use in the model based on local chemical conditions. In each submechanism, we isolate slow species and slow reactions from the coupled system of fast species to be solved. Because many species in the full mechanism are important only in source regions, we find that we can reduce the effective size of the mechanism by 70% globally without sacrificing complexity where/when it is needed. The computational cost of the chemical integration decreases by 50% with accuracy losses smaller than 1% over multi-year simulations. The chemical coherence of the algorithm allows it to accommodate updates to the original chemical mechanism without having to reconstruct the suite of submechanisms.

1 Introduction

There is a strong motivation to couple atmospheric chemistry with meteorology and surface processes in Earth system models (ESMs) because chemical and aerosol species exert strong forcing and feedbacks on the radiative budget of the Earth both directly and indirectly (CLIMA et al., 2016), but this is challenging because of the high computational cost (National Research Council, 2012). Global atmospheric chemistry mechanisms typically include over one hundred chemical species coupled through kinetics, and integrating the chemical evolution of that system requires solving a large and stiff system of differential equations (Brasseur and Jacob, 2017). However, characterizing the chemical composition in most regions of the atmosphere does not in fact require solving for the full chemical complexity of the mechanism. Here we present an adaptive, stable and chemically coherent algorithm for solving atmospheric chemistry in ESMs that reduces the computational cost in half, with losses in accuracy less than 1% that do not propagate forward in time. Our algorithm is based on general principles that can be easily applied to a wide range of mechanisms.

Previous approaches of simplifying atmospheric chemistry mechanisms all involve some loss of accuracy or generality (Brasseur and Jacob, 2017). Reducing the dimension of the coupled system can be obtained by decreasing the number of species (Sportisse and Djouad, 2000), isolating long-lived species (Young and Boris, 1977), and removing unimportant reactions (Brown-Steiner et al., 2018). However, the importance of a species or a reaction varies in different atmospheric conditions, so these schemes are not well adapted to global models. Some studies (Jacobson 1995; Rastigeyev et al., 2007) use different subsets of the full chemical mechanism for different regions with specified or locally determined boundaries, but this has limited success because the atmosphere has a continuum of chemical regimes, and geographic boundaries between regimes should be dynamic rather than pre-defined. An adaptive method to define mechanism subsets locally and on the fly has been proposed by Santillana et al. (2010) but the computational overhead of customizing the mechanism on the fly offsets computational gains. The overhead can be avoided by compiling a library of pre-defined mechanism subsets (Shen et al., 2020), but a challenge is to select these subsets in a manner that is chemically coherent and portable across mechanisms.

In this work, we continue developing the adaptive method described by Shen et al. (2020). This method pre-assembles a small number of subsets of the full chemical mechanism representing the range of conditions in the troposphere and stratosphere, and selects the most appropriate submechanism to use in the model locally and on the fly. The submechanisms are constructed by first splitting the full mechanism's atmospheric species into *N* different blocks based on similarity of chemical behaviors, using a machine learning clustering method. We then define the submechanisms as different assemblages of blocks, select *M* of these assemblages to encompass the majority of chemical conditions in the atmosphere, and build them into the model. The choice of submechanism in the model is then made locally by computing chemical production and loss rates of the mechanism species and deciding which need to be part of the coupled chemical computation ('fast' species) and which can be tracked independently ('slow' species). A major development here is to define chemically coherent blocks that allow the method to easily accommodate changes in the chemical mechanism and to be readily applied to different mechanisms. We further improve the performance of the method by reducing the number of reactions as well as the number of species in the submechanisms.

2 Method description

Here we describe the adaptive method as applied in the GEOS-Chem global model, although it is applicable to any model. We begin with a brief description of the model as relevant to the presentation.

2.1 GEOS-Chem model

We use the GEOS-Chem version 12.0.0 global 3-D model for tropospheric and stratospheric chemistry (https://doi.org/10.5281/zenodo.1343547). For development and testing purposes, we choose a horizontal resolution of 4°×5° and 72 pressure levels extending from surface to 0.01 hPa, and drive the model with MERRA2 assimilated meteorological data. The full mechanism for oxidant-aerosol chemistry in the model has 228 species and 724 reactions, including coupled gas-phase and aerosol chemistry for the troposphere and stratosphere (Sherwen et al., 2016; Eastham et al., 2014). The chemical operator uses a 4th-order Rosenbrock implicit method, implemented through the Kinetic Pre-Processor (KPP) (Sander and Sandu, 1996), to solve for the chemical evolution of species concentrations, involving iterative calculations and inversion of the Jacobian matrix that stores the sensitivity of species reaction rates to concentrations.

In part of this study, we test the performance robustness of our reduced algorithm by porting it to GEOS-Chem version 12.9.1 (https://doi.org/10.5281/zenodo.3950473). This new version of has a thoroughly updated mechanism of 262 species and 850 reactions, including improved organic nitrate chemistry (Fisher et al., 2018), isoprene chemistry (Bates and Jacob, 2019), and halogen chemistry (Wang et al., 2019). From version 12.0.0 to 12.9.1, we need to remove 49 old species and add 83 new species. We use 12 CPUs in a shared-memory Open Message Passing (Open-MP) parallel environment to test the performance of our algorithm throughout this study.

2.2 Separation of fast and slow species and reactions

Coupling between species as represented in the chemical solver is needed only for species with sufficiently fast production or loss rates (fast species), and similarly reactions need to be considered only if they are sufficiently fast. We separate the atmospheric species as fast or slow based on their production and loss rates relative to a threshold δ: fast if either $P_i(\boldsymbol{n}) \geq \delta$ or $L_i(\boldsymbol{n}) \geq \delta$, slow if $P_i(\boldsymbol{n}) < \delta$ and $L_i(\boldsymbol{n}) < \delta$ (P_i and L_i refer to the production and loss rates of the i^{th} species, δ is a threshold, and $\boldsymbol{n}$ is a vector of concentrations of all species). To get a sense of a relevant threshold, consider the hydroxyl radical (OH) which is central to driving oxidant-aerosol chemistry. OH has a daytime concentration of the order of 10^6 molecules cm^{-3} and a lifetime of 1s, so its production and loss rates are of the order of 10^6 molecules cm^{-3} s^{-1}. Species with production and loss rates smaller than 10^2-10^3 molecules cm^{-3} s^{-1} are unlikely to have fast influence other species in the mechanism (Santillana et al., 2010; Shen et al., 2020). In this study, we use δ from 500 to 1500 molecules cm^{-3} s^{-1} to partition the fast and slow species. We also define species with a chemical lifetime longer than 10 days as long-lived.

We pre-select a limited number (M) of submechanisms for which we pre-code the Jacobian matrix. In each submechanism, if a reaction is slower than 10 molecules cm^{-3} s^{-1} over all gridboxes that select this submechanism, this reaction is considered as unimportant in contributing to the threshold δ and is removed from the submechanism; but this reaction will be kept if it is faster than 10 molecules cm^{-3} s^{-1} in any of these gridboxes selecting this submechanism. The threshold we used to separate fast and slow reactions is slightly larger than 0 molecules cm^{-3} s^{-1} because of numerical precisions (unimportant reactions may still have a reaction rate > 0 molecules cm^{-3} s^{-1} in the numerical chemical solver in some timesteps). About 40-60% reactions can be removed using this strategy. For example, reactions of short-lived volatile organic compounds (VOCs) are removed in stratospheric gridboxes, and daytime photochemical reactions are removed in nightime gridboxes. Here we remove these slow reactions in each submechanism based on present-day atmospheric chemistry environment and it should be re-evaluated if this method is applied in other periods (e.g. pre-industrial times) when the atmospheric conditions could be very different from our present-day one.

We solve the fast species in their submechanism using the standard Rosenbrock solver. For the slow or long-lived species, we approximate the evolution of concentrations using an explicit analytical solution that assumes first-order loss (Santillana et al., 2010), written as

$$\frac{dn_i}{dt} = P_i - L_i = P_i - k_i n_i \tag{1}$$

$$n_i(t+\Delta t) = \frac{P_i(t)}{k_i(t)} + (n_i(t) - \frac{P_i(t)}{k_i(t)})e^{-k_i(t)\Delta t} \tag{2}$$

where n_i is the concentration of species i, P_i and L_i are the production and loss rates, k_i is the rate coefficient of the first-order loss, and Δt is the time step. Solving for Eq.2 entails negligible computational cost. As such, we still update the concentrations of all species but in a more efficient way.

2.3 Defining the distance between species in the mechanism

We construct coherent subsets ('blocks') of the species in the mechanism species based on their linkages through the mechanism reactions. This is done objectively by defining the species distances in the mechanism using graph theory. In general, two species should have shorter distances if they appear in the same reaction more times and have similar products in the mechanism. From the full mechanism of 228 species and 724 reactions, we find 3400 species pairs of reactants-products and map them to an undirected graph that has 228 vertices and 1422 edges. For example, in the reaction A+B-→C, there are 2 pairs (A-C and B-C) of reactants-products, 3 vertices (A, B, and C) and 2 edges (A-C and B-C). If species i and j

share the same edge, we define their distance as

$$D_{i,\,j} = \frac{T_{i,j}}{\sqrt{T_i T_j}} \tag{3}$$

Where $T_{i,\,j}$ is the number of reactions that include both species i and j (one is the reactant and the other is the product), and T_i (or T_j) is the number of species that appear in the same reactions with species i (or j). If species i and j never appear in the same reaction so they do not share the same edge in the graph, their distance is calculated as the length of the shortest path from species i to j. For example, the distance of toluene (TOLU) and xylene (XYLE) can be defined as the length of path TOLU-GLYX-XYLE (Figure 1, GLYX is glyoxal). Similarly, we can also define the distance between two blocks using Eq.3, in which we define $T_{i,\,j}$ as the number of reactions that include species in block i and j (one is the reactant and the other is the product) and T_i (or T_j) as the number of blocks that have reactions with block i (or j).

Equation 3 can define the distance of species along reaction chains, but it may overestimate the distance of species that do not react with each other but have similar products (e.g. XYLE and TOLU). These species usually come from the same chemical family and should be close to each other in terms of distances. In our work, we address this shortcoming as follows. First, we denote each species i by a vector (D_i) that contains its distance with all other species. The similarity of two species i and j can be thus defined as their Euclidean distance $||D_i - D_j||$. Second, for each species i, we decrease its distance with the 5 species that have highest similarity with it by 50% and this scaling is applied only once for each species pair. Using 10 highest-similarity species instead of 5 and decreasing distances by 30% or 70% does not change the results. We store these modified distances of all species pairs in a 228x228 matrix. We also tried weighting the species distances using the logarithms of their global mean reactions rates but this does not have significant effects on our final results.

2.4 Selection of species blocks and submechanisms

We construct submechanisms by assemblage of chemically coherent blocks in order to minimize the fraction of fast species to be tracked in the model. To partition the species into N blocks, we use a training dataset from a GEOS-Chem simulation for 2013 consisting of the first 10 days of February, May, August, and November sampled every 6 hours (160 time steps in total).

For the 228-species mechanism in GEOS-Chem, there are in 2^{228}-1 possible combinations of species and we need to pre-select M of them to form submechanisms that can encompass the range of atmospheric conditions. To reduce the dimensionality of this problem, we start by splitting the 228 species into N different blocks. A block is considered as fast if at least one species in that block is fast (P or $L > \delta$). Building on the N blocks, we define the submechanism as different assemblages of fast blocks, which yields 2^N - 1 possible submechanisms. Each gridbox in the model domain may correspond to one of these 2^N - 1 submechanisms. More specifically, for each gridbox *j*, we diagnose species *i* as fast or slow following the definition of Section 2.2. We define $y_{i,j}$ = 1 if any species in the block is fast or $y_{i,j}$ = 0 if all species in the block are slow. Thus, the fraction Z_1 of all species that needs to treated as fast can be written as

$$Z_1 = \frac{1}{\Omega} \sum_j \sum_i y_{i,j} \tag{4}$$

where Ω is the number of species × gridboxes.

We need to limit the number of submechanisms to a small number *M* in order to keep the compilation of the code manageable. Gridboxes that do not correspond to any of the *M* submechanisms need to be matched to one of the *M* submechanisms by moving some blocks from slow to fast, and we select the submechanism has a minimum number of moves. As such, the values of some $y_{i,j}$ need to be changed from 0 to 1 and we refer to $y*_{i,j}$ as the indicators adjusted by these changes. The fraction of species *f*(*M*, *N*) that need to be treated as fast over the global domain is given by:

$$f(M,N) = \frac{1}{\Omega} (\sum_{V_1} \sum_i y_{i,j} + \sum_{V_2} \sum_i y^*_{i,j}) \tag{5}$$

where V_1 are the gridboxes that can be represented directly by the *M* chemical submechanisms, and V_2 are the gridboxes that must be matched to the *M* submechanisms.

The cost function *Z* to be minimized in the selection of submechanisms can be written as

$$Z = f(M,N) + \gamma Dist \tag{6}$$

Where *Dist* is the sum of distances for all pairs of species if they are in the same block, γ is a regularization factor; *f* is the fraction of species that needs to be treated as fast over the testing domain based on *M* and *N* (Eq. 5). We adjust γ so that the second term on the right part of Eq.6 contributes to 20% of the total cost function. We seek the partitioning of species into blocks that will minimize *Z,* and we use for that purpose the simulated annealing algorithm (Kirkpatrick et al., 1983). We treat all 37 reactive inorganic halogen species as fast in the stratosphere to conserve the total mass of halogen species, same as Shen et al. (2020). We tested a range of values from 5 to 20 for *N* and from 10 to 40 for *M.* In the simulated annealing algorithm, we start from a randomly generated partition of the *N* blocks. In each iteration, we randomly move one species from one block to another. If the cost function decreases, this transition is accepted; otherwise, it is accepted with a probability controlled by a parameter named temperature. The temperature parameter decreases gradually as the optimization proceeds (Kirkpatrick, 1983).

The explicit solution by Eq. 3 does not strictly conserve mass (Shen et al., 2020), and we find that this is a problem for halogen species in the stratosphere due to the long lifetime of the collective halogen families and the alternance of the component species as fast and slow over day and night. To avoid this problem, we treat all 37 reactive inorganic halogen species as fast in the stratosphere. Thus, among the *N* blocks, 2 are allocated to the reactive inorganic halogen species, and *N*-2 are allocated to the other species. The transition of species between the 2 inorganic halogens blocks and other *N*-2 blocks are not accepted in the optimization process.

2.5 Error analysis

We use the Relative Root Mean Square (RRMS) metric as given by Sandu et al. (1997) to characterize the error:

$$\boldsymbol{RRMS_i} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{Q_i}}\sum_{\boldsymbol{j=1}}^{\boldsymbol{Q_i}}\left(\frac{\boldsymbol{n_{i,j}^{\mathrm{reduced}}} - \boldsymbol{n_{i,j}^{\mathrm{full}}}}{\boldsymbol{n_{i,j}^{\mathrm{full}}}}\right)^{\mathbf{2}}} \tag{7}$$

Where $n_{i,j}^{\mathrm{reduced}}$ and $n_{i,j}^{\mathrm{full}}$ are the concentrations for species *i* and gridbox *j* in the reduced and full chemical mechanisms, the sum is over the gridboxes where $n_{i,j}^{\mathrm{full}}$ is greater than a threshold *a*, and Q_i is the number of such gridboxes. Here we use a = 1×10^6 molecules cm^{-3} as in Eller et al. (2009) and Santillana et al. (2010), and will also show results with a = 1×10^5 molecules cm^{-3}

A second metric to evaluate our adaptive chemical mechanism is the relative difference of atmospheric abundances for all species compared to the standard simulation. This tests for accumulating bias over long simulation periods.

3 The adaptive algorithm for the chemical operator

3.1 Potential for local simplifications of atmospheric chemistry mechanisms

Figure 2 displays the potential for local simplification of the full mechanism over the global domain, based on local chemical production and loss rates for the 228 species simulated by GEOS-Chem. Using a threshold δ of 500 molecules $\mathrm{cm}^{-3}\mathrm{s}^{-1}$ for production and loss rates to define the fast and slow species (see Section 2.2 for the selection of this threshold), a given percentage of species can be excluded from the coupled chemical mechanism. That percentage is 75% for surface grid cells and reaches 90% in the stratosphere. When compared with removing long-lived species (lifetime > 10 days), a strategy that is most commonly used in simplifying the chemical mechanism (e.g. Young and Boris, 1977), removing slow ones is more effective because it can exclude a large majority of unimportant species. As seen from Figure 2a, long-lived but fast species are only present in the lower troposphere and their percentage is below 1% when averaged globally. Figure 2b shows the percentage of slow reactions (<10 molecules $cm^{-3}s^{-1}$) in the atmosphere, which is found to be 75-85% in the troposphere and 90% in the stratosphere (Figure 2b). A slow reaction does not necessarily mean that it is not important, but if it is slow in all gridboxes of a subdomain of the atmosphere then we can safely remove it in this subdomain. These results show that most of the atmosphere does not in fact require solving for the full complexity of the mechanism, so considerable simplification is possible if we can recognize the spatial and temporal patterns of chemical complexity in different atmospheric subdomains. As we will show later, we are able to exclude 50-80% species and 40-60% reactions at different altitudes of the atmosphere from the coupled system in our adaptive algorithm (Figure 2).

3.2 Performance of our adaptive algorithm

Our work addresses two problems in the original Shen et al. (2020) approach. First, the blocks identified by their machine learning approach based solely on minimizing computational time (Equation 6 with no regularization term) were not chemically coherent. Some species known to be chemically coupled by simple inspection of the mechanism were separated in different blocks. The regularization term addresses this shortcoming by penalizing the separation of species that are linked in the mechanism by direct and indirect reactant-product relationships. Second, Shen et al. (2020) only achieved 30-40% time-savings. Here we improve the performance of the algorithm by not only isolating slow species but also removing slow reactions from the submechanisms, thus speeding up the computation of the Jacobian. The slow reactions removed in each submechanism are pre-defined (see Section 2.2 for more details).

Figure 3 shows the fraction of fast species that needs to be solved using the chemical solver in the global domain as a function of *M* (submechanisms) and *N* (blocks). If *N* is low so each block is large, the mixing of slow species with fast ones will increase the likelihood of treating all species in this block as fast. If *N* is too high relative to *M*, more gridboxes cannot be represented by the *M* submechanisms and hence have to use submechanisms of higher complexity than needed. For each *N*, there exists a threshold for *M* above which the cost function remains almost unchanged. In order to make the code manageable, we choose to use $M = 20$ resulting in an optimal value $N = 13$ at which only 30% of the species need to be treated as fast in the global tropospheric and stratospheric domain (Figure 3). As shown in Figure 3, this performance is relatively insensitive to the choice of *M*.

Figure 4a-b shows the method and the results of partitioning of species into the 13 (N=13) blocks (the detailed list of species is in Table 1). Oxidants and methane oxidation products are important everywhere so blocks 1 and 2 are part of the submechanism in 50-80% of gridboxes (Figure 4b). Aside from the oxidants, bromine and chlorine radicals (block 3) also play a pervasive role in tropospheric and stratospheric chemistry, and are part of the submechanism in 39% of gridboxes (Figure 4b). Our algorithm can also largely separate anthropogenic VOCs from biogenic ones, although a few such species may overlap because they have similar products (e.g. block 7 contains both anthropogenic and biogenic precursors of glyoxal; see Table 1). Anthropogenic VOC species are important in 10-20% gridboxes, which are mainly found in the lower troposphere (Figure S1). Biogenic VOC species generally have shorter lifetimes, so they are found to be important only in 0.5-4% gridboxes in the terrestrial lower troposphere near their sources (Figure S2). Most of the secondary organic aerosols can be found in Block 8 and 11, which are found to be fast in 0.5-3% gridboxes (Figure 4b). Halocarbons are relatively inert in the atmosphere and they are found to be important in <2 % of gridboxes (Figure 4b).

Figure 4c shows the network of these 13 blocks in the full mechanism. A connection between two blocks means that species from these two blocks are reactants or products in the same reactions. If more species from two blocks are found in the same reactions and have similar products, the distance between these two blocks is shorter (Eq.3), as represented by the length of edges in the graph. As seen from the figure, atmospheric oxidants play a central role in the mechanism; thus they connect with all other blocks. Anthropogenic and biogenic VOCs have similar products (e.g. acetone and formaldehyde) and they are found to be interconnected with each other. Halogen species interact with the system mainly through the atmospheric oxidants. This network also shows that the optimized blocks by our algorithm are chemically coherent.

Figure 5 shows the composition of the 20 submechanisms as defined by the 13 blocks. The first 11 submechanisms do not need to solve any biogenic VOC species and include <40% of the full mechanism's species. More than 70% of gridboxes select these non-biogenic submechanisms, which are mainly distributed in the stratosphere and free troposphere (Figure 5b and S3). The other 9 submechanisms have higher complexity and are mainly used in the lower troposphere over the continents (Figure 5b and S3). Only 0.05% of gridboxes need to use the full chemical mechanism.

Based on different choices of the rate thresholds δ separating fast and slow species, we can adjust the complexity and accuracy in the adaptive mechanism. Increasing the threshold can speed up the computation but at the expense of accuracy. Figure 6b-c shows the median RRMS error (see the definition in Eq.7) of all species and the CPU time used by chemical integration for threshold rates of 500 and 1500 molecules cm^{-3} s^{-1}, compared to the full chemical mechanism. This comparison is conducted by running the simulation for 3 years to examine the sensitivity to different δ. For each δ, we test the effects of using two strategies, including isolating slow species (A1) and removing slow reactions (A2) (see Figure 6). By isolating slow species (A1), we can reduce the chemical integration time by 38-43% with errors of 0.4-0.9%. By further removing the slow reactions in each submechanism (A1+A2), we can reduce the CPU time by 44-49% and the median RRMSE error remains at 0.53-1.0%. When using a higher threshold δ = 1500 molecules cm^{-3} s^{-1} to isolate slow species and removing the slow reactions, we can reduce the chemical integration time by 50%, and the median RRMSE error maintains at the level of 1% for all gridboxes in the atmosphere and less than 0.5% in the boundary layer. Three-year simulation tests show that the errors of our method are stable over time (Figure 7), so it can be safely used for long-term simulations. The distribution of errors shows that >97.5% species have an error lower than 10% (Figure S4). The relative error on concentrations compared to the standard simulation is below 0.5% everywhere for key species like O_3, OH and sulfate, and is 1-6% for NO_2 (Figure S5-S7). Using a higher threshold of δ (> 1500) only leads to marginal improvement in computer time but the RRMSE error quickly increases.

The median relative difference in atmospheric abundances among all species remains at 0% over this 3-year period; the relative differences for key species like ozone, OH, sulfate and NO_2 also remain at 0% and are within ±10% for >99% of the other species (Figure S7). Computing the RRMSE for all species with concentrations higher than $a=1\times10^5$ molecules cm^{-3} (instead of $1x10^6$ molecules cm^{-3}) shows similar results except that the magnitude of the error is higher because the relative difference is expected to be higher at low species concentrations (Fig. S6, S8, S9).

3.3 Adapting to mechanism updates

Chemical mechanisms in models are frequently updated, including addition and removal of species. Because the species blocks are chemically coherent, our algorithm can accommodate mechanism updates without requiring reconstruction of the submechanisms. New species simply need to be added to the appropriate blocks. Figure S10 shows the diagram for adding new species into the mechanism. Attribution of a species to a given block can be easily determined by its chemical family and the percentage of gridboxes that treat this species as fast when averaged globally. In order not to compromise the computational efficiency, the basic rule is to not mix faster species with slower ones. For example, biogenic VOC species and their products could go to Block 8-9 if the percentage of gridboxes that treat them as fast is >1% or Block 10-11 if the percentage is <1%. Our algorithm is robust to misplacements of new species, which may affect computational performance but will not enlarge the error.

To demonstrated this procedure, we ported our method originally developed with the GEOS-Chem 12.0.0 chemical mechanism (228 species and 724 reactions) to the latest GEOS-Chem 12.9.1 version (262 species and 850 reactions). This involved major changes to the mechanism including for organic nitrate chemistry (Fisher et al., 2018), isoprene chemistry (Bates and Jacob, 2019), and halogen chemistry (Wang et al., 2019), with removal of 49 species and addition of 83 new ones. We add these new species following the diagram in Figure S10. After running the new version of the model for 12 months, our reduced algorithm shows consistent improvement in performance, reducing the chemical integration time by 53% and maintaining error of 0.8% in the atmosphere and <0.4% in the boundary layer (Figure 6d).

4. Conclusions

The high computational cost of chemical integration has been a barrier for the inclusion of atmospheric chemistry in Earth system models. Typical chemical mechanisms include over 100 species coupled on short time scales. Previous research has proposed a variety of ways to speed up the chemical operator, all involving some loss of accuracy or generality. In this study, we have presented a machine learning-guided adaptive method that can reduce the chemical integration time by 50% when compared to the full chemical mechanism while maintaining error at the level of 1% and retaining full diagnostic capability.

In our algorithm, we first partition the mechanism species in into chemically coherent blocks using a machine learning approach that analyzes production/loss rates and chemical linkages between species. . We then assemble these blocks into an ensemble of submechanismsto encompass the range of chemical environments in the atmosphere. The model picks locally on the fly which submechanism to use based on species' production and loss rates. The original mechanism can thus be greatly reduced in most environments while maintaining complexity where needed. Our method can reduce the chemical integration time by 50% while incurring errors of less than 1%, with no error growth over multi-year global simulations. Updates to the original mechanism can be accommodated by assigning new species to the existing chemically coherent blocks without having to reconstruct the suite of submechanisms.

Our method has many advantages over previously proposed approaches to reduce chemical mechanism: (1) it is chemically coherent; (2) it can save 50% computer time in chemical integration and maintain the error better than 1%; (3) it is stable (no error growth over time) and can be used for long-term integrations; (4) it retains full diagnostic information of concentration and rates; and (5) it is scale-independent. Our algorithm can significantly ease the computational bottleneck for inclusion of comprehensive atmospheric chemistry in the next generation of earth system models.

Code availability. The standard GEOS-Chem code is available through https://doi.org/10.5281/zenodo.1343547 (version 12.0.0) and https://doi.org/10.5281/zenodo.3950473 (version 12.9.1). The updates for the adaptive mechanism can be found at https://doi.org/10.7910/DVN/KASQOC.

Data availability. All datasets used in this study are publically accessible at https://doi.org/10.7910/DVN/KASQOC.

Author contribution. L. Shen and D. Jacob designed the experiments and L. Shen carried them out. L. Shen and D. Jacob prepared the manuscript with contributions from all co-authors.

Competing Interests. The authors declare that they have no conflict of interest.

Acknowledgments. This research has been supported by the NASA Modeling and Analysis Program (NASA-80NSSC17K0134) and by the US EPA Science to Achieve Results (STAR) Program (EPA- G2019-STAR-C1).

References

Bates, K.H., and D.J. Jacob, A new model mechanism for atmospheric oxidation of isoprene: global effects on oxidants, nitrogen oxides, organic products, and secondary organic aerosol, Atmos. Chem. Phys., 19, 9613-9640, 2019

Brown-Steiner, B., Selin, N. E., Prinn, R., Tilmes, S., Emmons, L., Lamarque, J.-F., and Cameron-Smith, P.: Evaluating simplified chemical mechanisms within present-day simulations of the Community Earth System Model version 1.2

with CAM4 (CESM1.2 CAM-chem): MOZART-4 vs. Reduced Hydrocarbon vs. Super-Fast chemistry, Geosci. Model Dev., 11, 4155–4174, http://sci-hub.tw/10.5194/gmd-11-4155-2018, 2018.

Brasseur, G.P. and Jacob, D.J.: Modeling of atmospheric chemistry, Cambridge University Press, 2017

CLIMA, T.: Coupled Chemistry-Meteorology/Climate Modelling (CCMM): Status and relevance for numerical weather prediction, atmospheric pollution and climate research (Symposium materials). WMO GAW Report, Geneva, Switzerland, 2016.

Damian, V., Sandu, A., Damian, M., Potra, F., and Carmichael, G. R.: The kinetic preprocessor KPP – a software environment for solving chemical kinetics, Comput. Chem. Eng., 26, 1567– 1579, 2002.

Eastham, S. D., Weisenstein, D. K., and Barrett, S. R. H.: Development and evaluation of the unified tropospheric–stratospheric chemistry extension (UCX) for the global chemistry-transport model GEOS-Chem, Atmos. Environ., 89, 52–63, doi:10.1016/j.atmosenv.2014.02.001, 2014.

Eller, P., Singh, K., Sandu, A., Bowman, K., Henze, D. K., and Lee, M.: Implementation and evaluation of an array of chemical solvers in the Global Chemical Transport Model GEOS-Chem, Geosci. Model Dev., 2, 89–96, http://sci-hub.tw/10.5194/gmd-2-89-2009, 2009.

Fisher, J.A., Atlas, E.L., Barletta, B., Meinardi, S., Blake, D.R., Thompson, C.R., Ryerson, T.B., Peischl, J., Tzompa-Sosa, Z.A. and Murray, L.T.: Methyl, ethyl, and propyl nitrates: global distribution and impacts on reactive nitrogen in remote marine environments. Journal of Geophysical Research: Atmospheres, 123(21), 12-429, 2018.

Jacobson, M. Z.: Computation of global photochemistry with SMVGEAR II, Atmos. Environ., 29, 2541–2546, 1995.

Keller, C. A. and Evans, M. J.: Application of random forest regression to the calculation of gas-phase chemistry within the GEOS-Chem chemistry model v10, Geosci. Model Dev., 12, 1209–1225, https://doi.org/10.5194/gmd-12-1209-2019, 2019.

Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P.: Optimization by simulated annealing, Science, 220 (4598), 671-680, 1983.

National Resarch Council: A National Strategy for Advancing Climate Modeling, National Academies Press, Washington DC, 2012.

Rastigeyev, Y., Brenner, M.P., Jacob, D.J.: Spatial reduction algorithm for atmospheric chemical transport models. Proc. Natl. Acad. Sci. USA, 104, 13875-13880, 2007.

Sportisse, B., Djouad, R.: Reduction of chemical kinetics in air pollution modeling, J. Comput. Phys., 164, 354-376, 2000.

Sandu, A., Verwer, J. G., Blom, J. G., Spee, E. J., Carmichael, G. R., and Potra, F. A.: Benchmarking stiff ode solvers for atmospheric chemistry problems II: Rosenbrock solvers, Atmos. Environ., 31, 3459–3472, 1997

Santillana, M., Le Sager, P., Jacob, D.J., and Brenner, M.P.: An adaptive reduction algorithm for efficient chemical calculations in global atmospheric chemistry models, Atmos. Environ., 44(35), 4426-4431, 2010.

Shen, L., Jacob, D. J., Santillana, M., Wang, X., and Chen, W.: An adaptive method for speeding up the numerical integration of chemical mechanisms in atmospheric chemistry models: application to GEOS-Chem version 12.0.0, Geosci. Model Dev., 13, 2475–2486, https://doi.org/10.5194/gmd-13-2475-2020, 2020.

Sherwen, T., Schmidt, J. A., Evans, M. J., Carpenter, L. J., Großmann, K., Eastham, S. D., Jacob, D. J., Dix, B., Koenig, T. K., Sinreich, R., Ortega, I., Volkamer, R., Saiz-Lopez, A., Prados-Roman, C., Mahajan, A. S., and Ordóñez, C.: Global impacts of tropospheric halogens (Cl, Br, I) on oxidants and composition in GEOS-Chem, Atmos. Chem. Phys., 16, 12239–12271, http://sci-hub.tw/10.5194/acp-16-12239-2016, 2016.

Wang, X., Jacob, D. J., Eastham, S. D., Sulprizio, M. P., Zhu, L., Chen, Q., Alexander, B., Sherwen, T., Evans, M. J., Lee, B. H., Haskins, J. D., Lopez-Hilfiker, F. D., Thornton, J. A., Huey, G. L., and Liao, H.: The role of chlorine in global tropospheric chemistry, Atmos. Chem. Phys., 19, 3981–4003, https://doi.org/10.5194/acp-19-3981-2019, 2019.

Young T. R. and Boris J. P.: A numerical technique for solving stiff ordinary differential equations associated with the chemical kinetics of reactive flow problems, J. Phys. Chem., 81, 2424–2427, 1977.

Figures and Tables

Table 1. Partitioning of GEOS-Chem chemical species into *N* = 13 blocks[a].

Categories	Blocks	Major components	Species	%gridbox[b]
Oxidants and methane products	1	Oxidants	MPN, N2O5, HNO3, O3, NO2, MO2, H2O, NO3	74.3±14.5%
	2	Oxidants, methane	HNO4, HNO2, H, CH4, H2O2, CH2O, HO2, NO, O, CO, O1D, OH	55.3±11.6%
Inorganic halogens	3	Bromine and chlorine radicals	BrNO2, IONO, OIO, ClOO, OClO, BrCl, HOI, Br2, IONO2, BrNO3, I, IO, HOBr, HOCl, ClNO3, BrO, HCl, HBr, Cl, Br, ClO	39.4±18.1
	4	Iodine reservoirs	AERI, ISALA, ISALC, I2O4, I2O2, I2O3, IBr, INO, HI, ICl, Cl2O2, ClNO2, BrSALC, BrSALA, I2, Cl2	1.7±1.4%
Anthropogenic VOCs and sulfate	5	Alkanes, alkenes, acetone, sulfur compounds	MSA, MAP, ETP, DMS, PAN, SO4, ATOOH, MP, C2H6, ATO2, ACET, ETO2, ALD2, MCO3, SO2	20.0+9.1%
	6	Higher alkanes and oxidized organics	PPN, RA3P, RB3P, RP, ALK4, R4P, C3H8, EOH, A3O2, B3O2, RCO3, KO2, ACTA, MGLY, R4O2, R4N2, RCHO, MEK	9.5±4.1%
	7	Aromatics, glyoxal, and related OVOCs	SOAGX, IMAE, DHDC, BENZ, TOLU, TRO2, BRO2, XRO2, XYLE, HPALD, DHPCARP, HPC52O2, GLYX, HCOOH, GLYC, HAC	3.9±1.7%
Biogenic VOCs	8	Isoprene products (low NOx), secondary organic aerosols	LVOCOA, LVOC, SOAIE, SOAME, IEPOXD, IEPOXA, IEPOXB, HC187, IAP, VRP, MOBA, DHMOB, RIPB, RIPA, RIPD, IEPOXOO, HC5OO	2.5±1.4%
	9	Isoprene, isoprene nitrates	IMAO3, PP, MRP, DIBOO, IPMN, INPN, ISOPNB, MVKOO, CH2OO, PO2, ISOPNDO2, MACROO, ISOP, LIMO2, ISOPNBO2, ISOPND, VRO2, ISN1, HC5, RIO2, INO2, MRO2, PRPE, MACR, MVK	3.8±2.0%

	10	Terpenes	INDIOL, MONITA, IONITA, PIP, HONIT, ISNP, MTPA, MTPO, MOBAOO, LIMO, ROH, MONITS, CH3CHOO, MVKN, MONITU, MGLOO, R4N1, OLND, OLNN, PIO2	3.0±1.5%
	11	Isoprene products (high NOx), secondary organic aerosols	ISN1OA, ISN1OG, PYAC, SOAMG, DHDN, PMNN, PRPN, MAOP, ETHLN, ISNOHOO, NPMN, ISNOOB, MACRNO2, GAOO, MGLYOO, PRN1, PROPNN, MAN2, ISNOOA, MACRN, MAOPO2, NMAO3	0.5±0.6%
Organic halogens and other long-lived species	12	Halocarbons	CH2I2, CH2ICl, CH2IBr, CH3CCl3, CH3I, CHBr3, CH2Cl2, CHCl3, CH2Br2, HCFC123, HCFC141b, HCFC142b, HCFC22, CH3Br, CH3Cl	0.47±1.70%
	13	Chlorofluorocarbons	H1301, H2402, CCl4, CFC11, CFC12, CFC113, CFC114, CFC115, H1211, N2O, N, OCS	0.55±1.91%

[a]The full GEOS-Chem mechanism has 228 species. The full names of these acronyms can be found at http://wiki.seas.harvard.edu/geos-chem/index.php/Species_in_GEOS-Chem.

[b]Percentage of gridboxes in the global tropospheric+stratospheric domain that treat this species block as fast. We use a threshold δ of 500 molecules cm^{-3} s^{-1} to partition the fast and slow species.

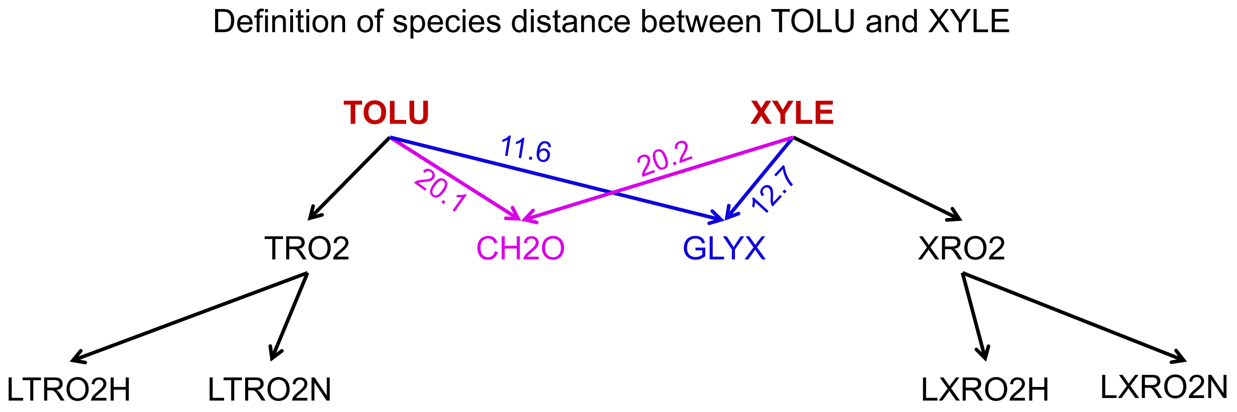


Figure 1. Definition of species distances for TOLU (toluene) and XYLE (xylene) using the analysis of family trees in graph theory. The number denotes for the distance between species as calculated by Eq. 3. The shortest path from TOLU to XYLE is TOLU-GLYX-XYLE in this graph, where GLYX is glyoxal.

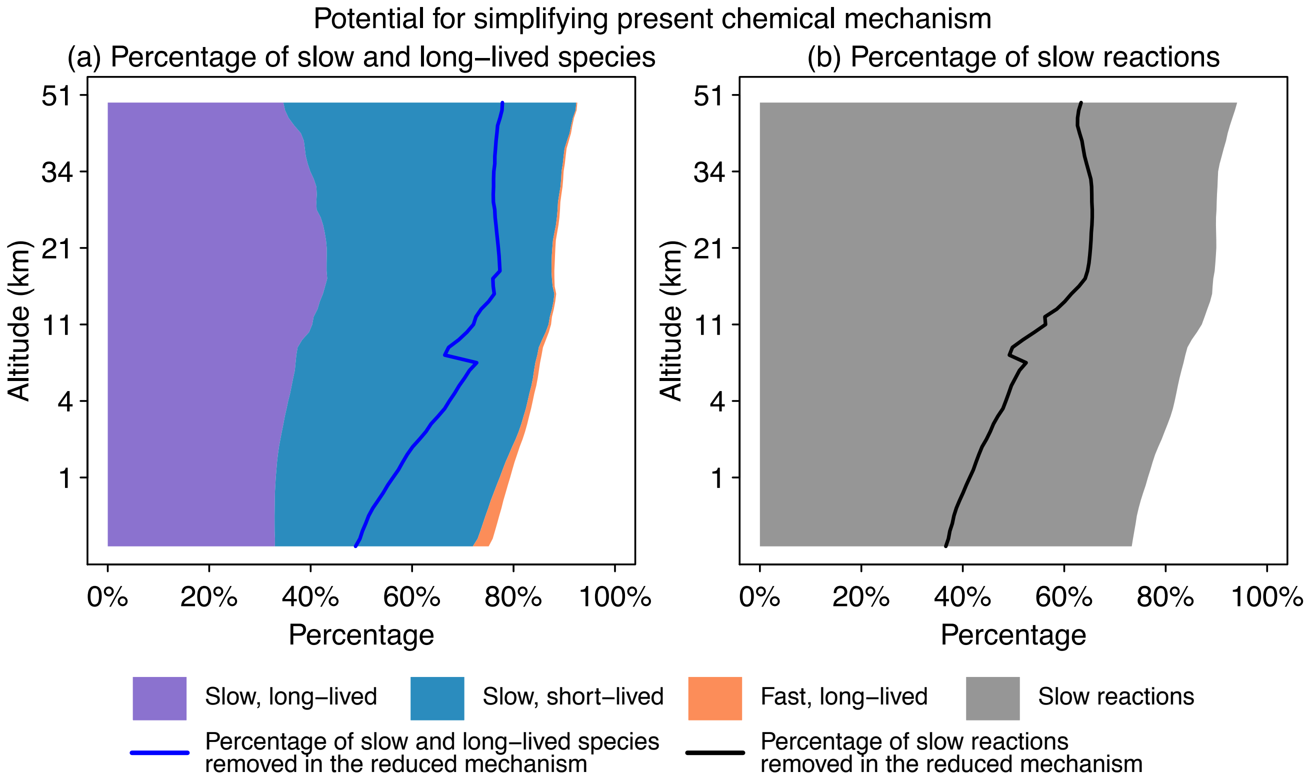


Figure 2. **Potential for simplifying the full chemical mechanism in a global GEOS-Chem model simulation.** Panel (a) shows the percentage of slow and long-lived species by altitude when averaged globally on Aug 1st 2013 at 0 GMT. We use a threshold of 500 molecules cm^{-3} s^{-1} to partition fast (*P* or *L* is > 500 molecules cm^{-3}s^{-1}) and slow species (*P* and *L* are both < 500 molecules cm^{-3}s^{-1}), and a lifetime of 10 days to separate long-lived and short-lived species. The blue line denotes for the percentage of slow and long-lived species that are actually removed in the reduced mechanism. Panel (b) shows the percentage of slow reactions (<10 molecules cm^{-3}s^{-1}) by altitude. The black line is the percentage of slow reactions actually removed in the reduced mechanism.

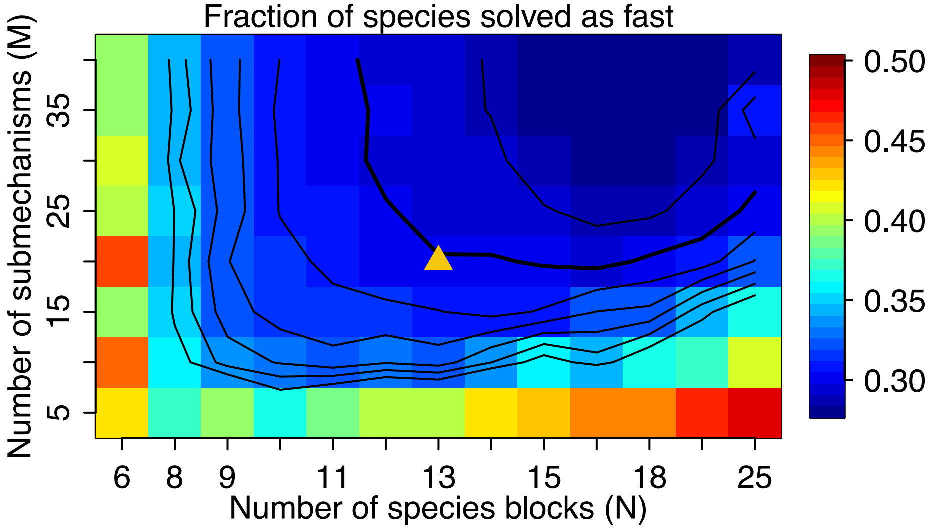
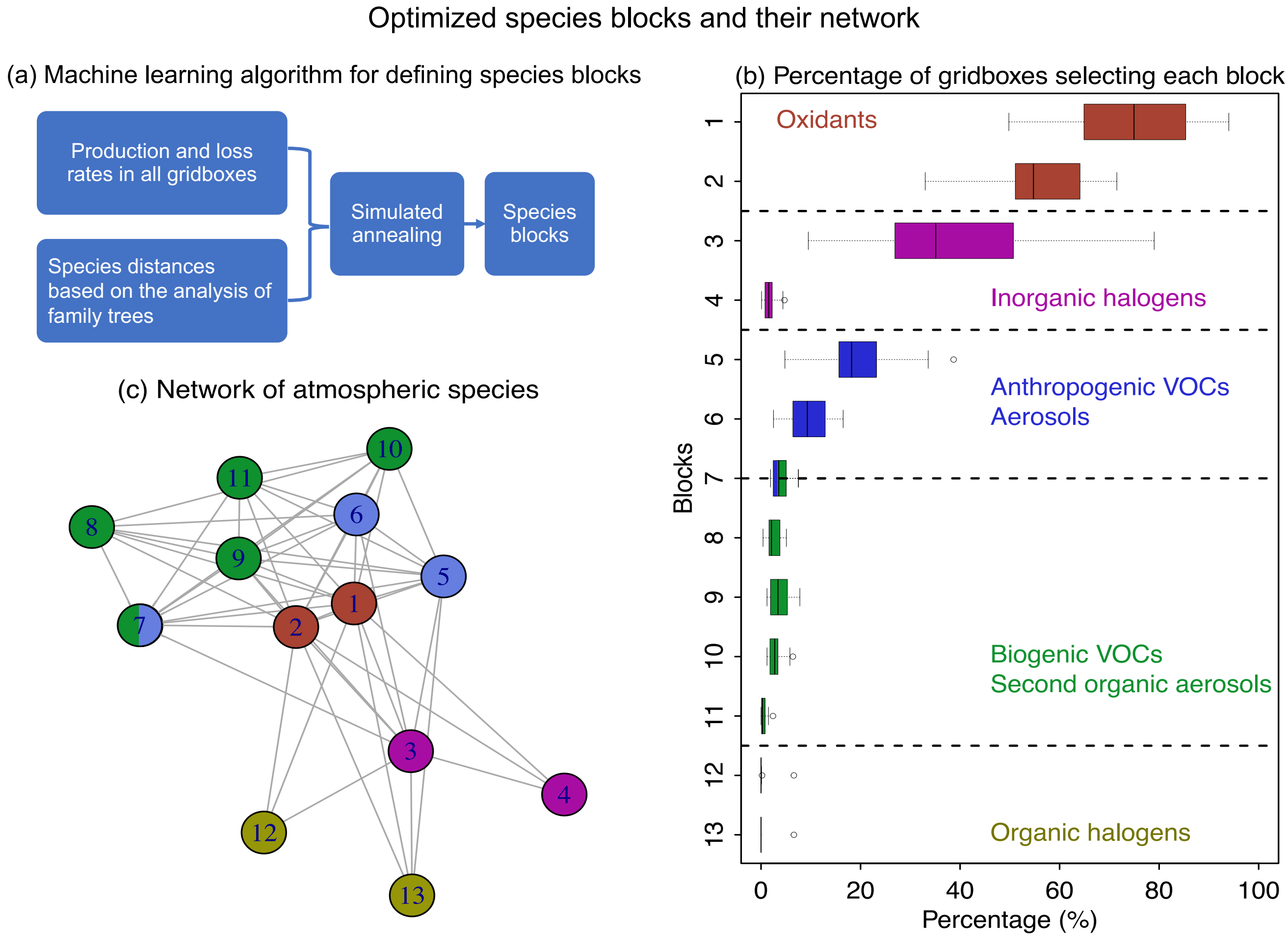


Figure 3. The fraction of species solved as fast as a function of *M* and *N*. We use *M*=20 and *N*=13 in our work, as shown by the triangle in the figure, with a threshold δ of 500 molecules cm^{-3} s^{-1} to partition the fast and slow species. The contour lines are spaced by 0.01 with the bold line for 0.30.



Optimized species blocks and their network

(a) Machine learning algorithm for defining species blocks

(b) Percentage of gridboxes selecting each block

(c) Network of atmospheric species

Figure 4. **Optimized species blocks and their network in the full chemical mechanism.** Panel (a) describes the machine learning method to solve for the species blocks. See more details in Section 2. Panel (b) shows the 13 species blocks and the percentage of gridboxes that treat the blocks in their submechanisms. The list of species in each block is given in Table 1. Block 7 includes both anthropogenic and biogenic VOCs. The left and right of each box are the 25th and 75th percentile, and the centerline is the 50th percentile. We use a threshold of 500 molecules cm^{-3} s^{-1} to partition fast and slow species. Panel (c) is the network of species blocks. A connection means that at least two species from these two blocks appear in the same reaction. The distance between the two blocks is proportional to the block distance as defined by Eq. 3.

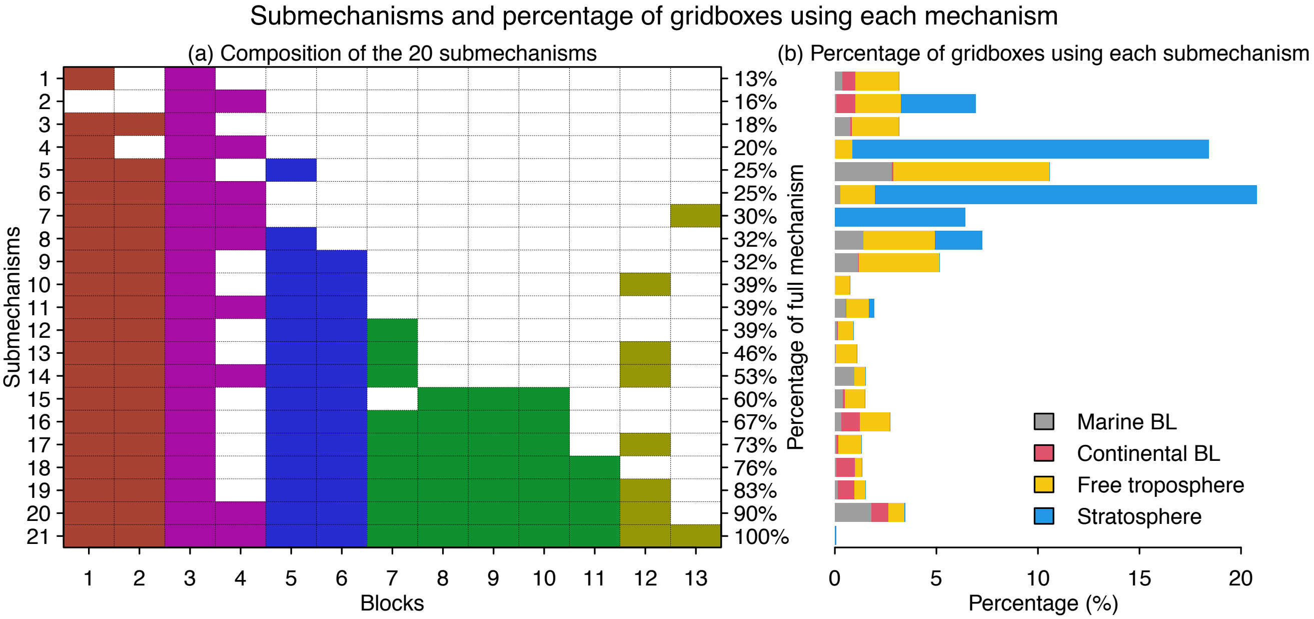
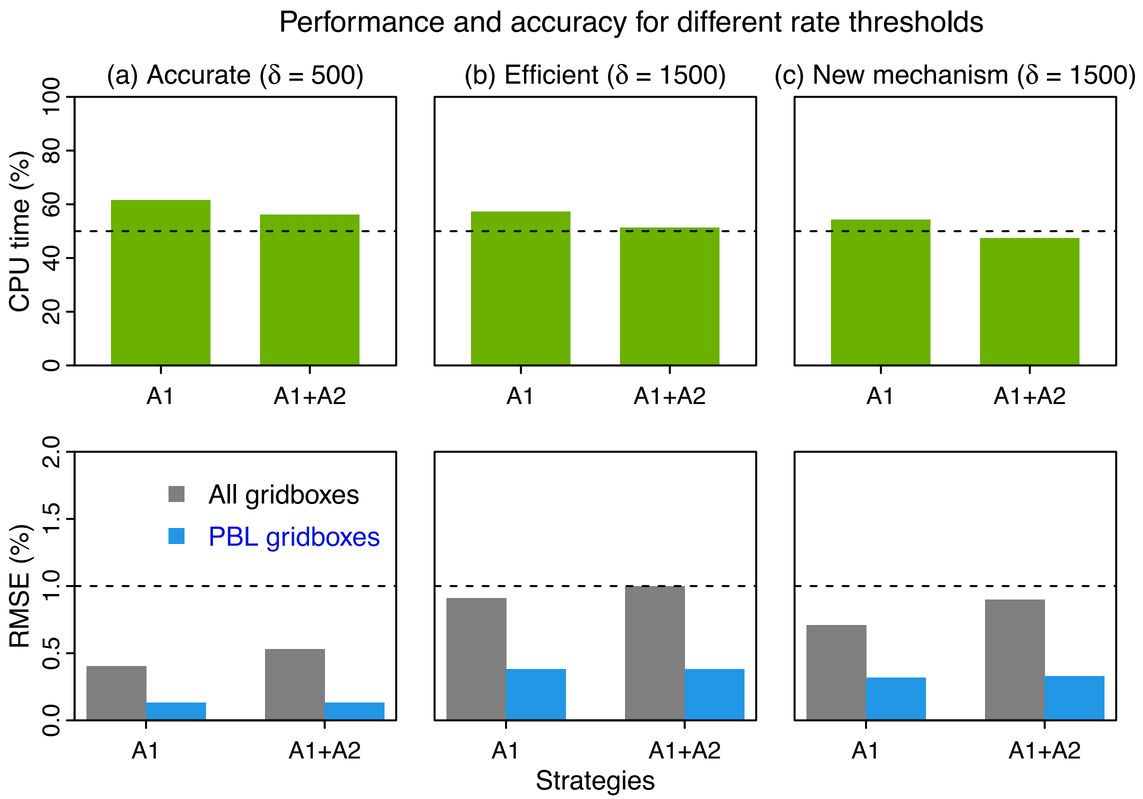


Figure 5. **Submechanisms and percentage of gridboxes using each mechanism**. Panel (a) shows the composition of the 20 submechanisms and full mechanism (the 21st one) as well as the percentage of species from the full mechanism that are treated as fast in each of them. Colors denote species block types as defined in Figure 4. Panel (b) shows the percentage of gridboxes using each submechanism in the marine boundary layer (BL), continental BL, free troposphere, and stratosphere.



410 **Figure 6**. **Performance and accuracy of the adaptive chemical mechanism**. We test the performance of this adaptive method by (A1) removing slow species (P_i or $L_i > \delta$) and (A2) removing slow reactions (reaction rate < 10 molecules cm^{-3}s^{-1}) on the last day of 3-year simulations. The unit of δ is molecules cm^{-3} s^{-1}. The performance is measured by the computing processor unit (CPU) time used by the chemical operator, and the accuracy is measured by the median relative root mean square (RRMS) error for species concentrations using the full chemical mechanism. For (a) and (b), we use the δ as 500 and 415 1500 molecules cm^{-3} s^{-1} in GEOS-Chem 12.0.0 that has 228 species and 724 reactions. For (c), we port the algorithm to GEOS-Chem 12.9.1 that has 262 species and 850 reactions. The number of blocks (*N*) is 13 and the number of chemical regimes is 21 (20 submechanisms (M=20) and one full mechanism).

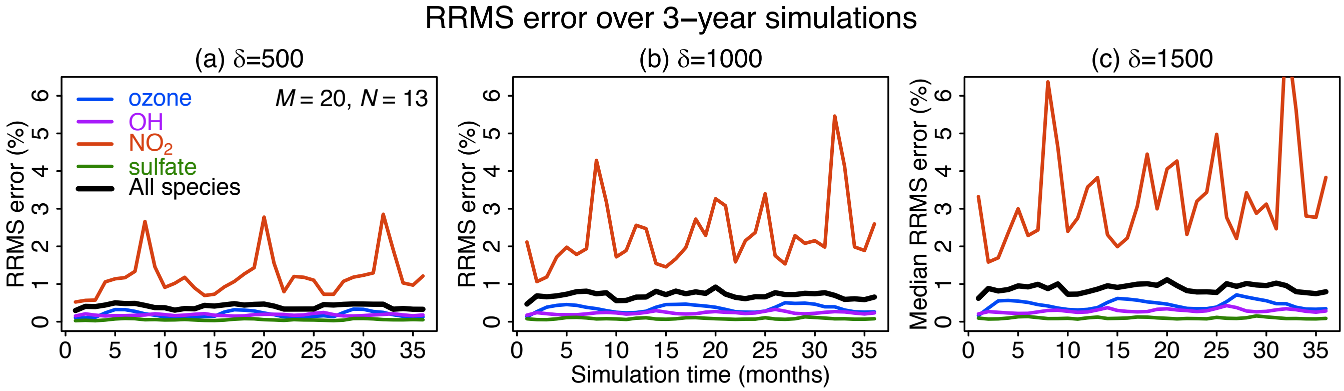


Figure 7. Accuracy of the adaptive reduced chemistry mechanism algorithm over a three-year GEOS-Chem simulation (see text). The accuracy is measured by the Relative Root Mean Square (RRMS, see Eq. 5) error on simulated concentrations relative to a simulation including the full chemical mechanism. Results are shown for the median RRMS error across all species in the mechanism and more specifically the RRMS error for ozone, OH, NO_2, and sulfate. The three panels show the effect of using different thresholds δ ranging from 500 to 1500 molecules cm^{-3} s^{-1} to separate fast and slow species.