https://doi.org/10.5194/gmd-2020-425
Preprint. Discussion started: 16 February 2021

Geoscientific Model Development Discussions

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

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

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

Abstract. Atmospheric composition plays a crucial role in determining the evolution of the atmosphere, but the high computational cost has been the major barrier to include atmospheric chemistry into Earth system models. Here we present an adaptive and efficient algorithm that can remove this barrier. Our approach is inspired by unsupervised machine learning clustering techniques and traditional asymptotic analysis ideas. We first partition species into 13 blocks, using a novel machine learning approach that analyzes the species network structures and their 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 on the fly which submechanism to use based on local chemical conditions. In each submechanism, we isolate slow species and unimportant reactions from the coupled system. Application to a global 3-D model shows that we can cut the computational costs of the chemical integration by 50% with accuracy losses smaller than 1% that do not propagate in time. Tests show that this algorithm is highly chemically coherent making it easily portable to new models without compromising its performance. Our algorithm will significantly ease the computational bottleneck and will facilitate the development of next generation of earth system models.

https://doi.org/10.5194/gmd-2020-425
Preprint. Discussion started: 16 February 2021


Geoscientific Model Development Discussions

1 Introduction

There is a strong motivation to couple atmospheric chemistry with meteorology and surface processes in Earth system models (ESMs) because it exerts 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 a 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 incurs very expensive computational overhead (Santillana et al., 2010). The overhead can be avoided by compiling a library of pre-defined mechanism subsets, but a challenge is to select these subsets in a manner that is chemically coherent and portable across mechanisms (Shen et al., 2020). Another direction involves the use of machine learning to replace the traditional chemical solver, but this method is subject to error growth over time and can only be applied to short-term simulations (Keller and Evans, 2019).

In this work, we use machine learning clustering ideas to guide us build an ensemble of chemically coherent subsets (chemical regimes) of a full chemical mechanism, with the appropriate regimes to be selected in the atmospheric model locally and on the fly. The success of this approach requires the chemical mechanism to have high generalizability and efficiency, which have not yet been achieved in previous studies. To address the generalizability issue, we enforce chemical coherence with a new clustering approach that includes the chemical distance between species in the full mechanism (obtained by analyzing the network structure of reactants and products) as part of the optimization. Species clusters are identified using a machine learning optimization approach that not only considers whether a species is fast or slow (a species is slow if both its production and loss rates are smaller than a threshold), but also minimizes the chemical distance between

https://doi.org/10.5194/gmd-2020-425
Preprint. Discussion started: 16 February 2021


Geoscientific Model Development Discussions

species to make the resulting adaptive mechanism chemically coherent. To improve the efficiency, we explore the potential to simplify the chemical mechanism by separately removing both slow species and unimportant reactions from the coupled system dynamically. Compared to previous studies (Shen et al., 2020), our algorithm is chemically coherent, accurate, and can halve the computational cost of the atmospheric chemistry integration. It can be ported to other chemical mechanisms or models with very little effort, which will significantly facilitate the development of the next generation of ESMs.

2 Model description

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) with a horizontal resolution of 4°×5° and 72 pressure levels extending from surface to 0.01 hPa, driven by MERRA2 assimilated meteorology. The full chemistry mechanism 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 model includes 20% gridboxes below 2km and 50% gridboxes below 12 km. The chemical reactions are integrated using a 4th-order Rosenbrock solver designed for high performance with the Kinetic Pre-Processor (Sandu et al., 1997; Damian et al., 2002).

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 model has 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 Definition of slow, long-lived species and unimportant reactions

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 the concentrations of all species). The hydroxyl radical (OH) has a daytime concentration of the order of 10^6 molecules cm^{-3} and a lifetime of 1s. Thus, species with production and loss rates smaller than 10^2-10^3 molecules cm^{-3} s^{-1} are unlikely to influence other species in the coupled scheme (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. Similarly, species with a lifetime longer than 10 days are considered as long-lived.

We pre-select a limited number (M) of submechanisms for which we pre-define the Jacobian matrix. In each submechanism,

https://doi.org/10.5194/gmd-2020-425
Preprint. Discussion started: 16 February 2021


Geoscientific Model Development Discussions

if a reaction is slower than 10 molecules cm^{-3} s^{-1} over all gridboxes that select this mechanism, this reaction is considered as unimportant and will be removed from the submechanism. We also have tested using other thresholds to remove the slow reactions, and we find any threshold higher that 10 molecules cm^{-3} s^{-1} will significantly compromise the accuracy.

2.3 Explicit analytical solution for slow or long-lived species

We solve the fast coupled species using the Rosenbrock solver to ensure high accuracy. 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

$$\frac{dn_i}{dt} = P_i - L_i = P_i - k_i n_i \quad (1)$$

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

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. When compared to the chemical solver, solving for Eq.2 requires very little computational cost.

2.4 Definition of species distances based on their network structures

We define the species distances using graph theory. From the full mechanism of 724 reactions, we find 3400 species pairs of reactants-products and map them to an undirected graph that has 228 vertices and 1422 edges. In this graph, if species i and j share the same edge, we define their distance as

$$D_{i,j} = \frac{T_{i,j}}{\sqrt{T_i T_j}} \quad (3)$$

Where $T_{i,j}$ is the number of reactions that include both species i and j, 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 S1, GLYX is glyoxal). Similarly, we can also define the distance between two species blocks using Eq.3.

Equation 3 can well 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 will decrease its distance with

https://doi.org/10.5194/gmd-2020-425
Preprint. Discussion started: 16 February 2021


Geoscientific Model Development Discussions

the 5 species that have highest similarity with it by 50%. We store the distances of all pairs in a 228x228 matrix.

2.5 Partition species into *N* blocks and construct *M* chemical regimes

In order to reduce the complexity of the optimization, we solve for the species blocks by minimizing the fraction of fast species in the coupled system. Species that are both long-lived and fast only account for ~1% of the gridboxes (Figure 1a), which are not excluded in the cost function here.

For the 228-species mechanism in GEOS-Chem, there are in 2^{228}-1 possible combinations of important species and we need to pre-select M of them to form submechanisms, which can encompass the majority of local atmospheric environments. 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. But we need to limit the number of submechanisms to a much smaller 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. The cost function Z can be written as

$$Z = f(M,N) + \gamma Dist \quad (4)$$

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 function to calculate the fraction of species that needs to be treated as fast over the testing domain based on M and N (a detailed description of f can be found in Text S1). We adjust the regulation parameter γ so that the second term on the right part of Eq.4 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 (Figure S2). In order to make the code manageable, we choose to use M = 20 and an optimal value N = 13 at which only 30-31% of the species need to be treated as fast in the global tropospheric and stratospheric domain.

To partition the species into N blocks, we use a training dataset from a GEOS-Chem simulation (version 12.0.0) for 2013, consisting of the global ensemble of tropospheric and stratospheric gridboxes for the first 10 days of February, May, August, and November sampled every 6 hours (160 time steps in total).

2.6 Error analysis

https://doi.org/10.5194/gmd-2020-425
Preprint. Discussion started: 16 February 2021


Geoscientific Model Development Discussions

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

$$RRMS_i = \sqrt{\frac{1}{Q_i}\sum_{j=1}^{Q_i}\left(\frac{n_{i,j}^{\text{reduced}} - n_{i,j}^{\text{full}}}{n_{i,j}^{\text{full}}}\right)^2} \quad (5)$$

Where $n_{i,j}^{\text{reduced}}$ and $n_{i,j}^{\text{full}}$ are the concentrations for species *i* and gridbox *j* in the reduced and full chemical mechanisms, and the sum is over the the gridboxes where $n_{i,j}^{\text{full}}$ is greater than a threshold *a*, and Q_i is the number of such gridboxes. Here we use $a = 1\times10^6$ molecules cm^{-3} as in Eller et al. (2009) and Santillana et al. (2010).

3 The adaptive algorithm for the chemical operator

3.1 Potential for local simplifications of atmospheric chemistry mechanisms

We use the standard global 4°x5° GEOS-Chem simulation for the troposphere and stratosphere including full chemistry (228 species, 724 reactions) as the reference chemical mechanism (Wang et al., 2019). 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. But coupling between species as represented in this 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. For species with slow production and loss rates (slow species), an explicit analytical solution (Eq.1-2) can be used at negligible computational cost (Santillana et al., 2010; Shen et al., 2020) without a significant loss in accuracy.

Figure 1 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 cm^{-3}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. Yong and Boris, 1977), removing slow ones is more effective because it can exclude a large majority of unimportant species. As seen from Figure 1a, long-lived but fast species are only present in the lower troposphere and their percentage is below 1% when averaged globally. Figure 1b 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

https://doi.org/10.5194/gmd-2020-425
Preprint. Discussion started: 16 February 2021


Geoscientific Model Development Discussions

90% in the stratosphere (Figure 1b). A slow reaction does not necessarily mean 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, which will be implemented in our adaptive mechanism. 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. So, in this study we will simplify the mechanisms through removing slow species and slow reactions. 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 1).

3.2 Performance of our adaptive algorithm

Even though there is great potential in simplifying the chemical mechanism (Figure 1), achieving this full potential would require tailoring reduced mechanisms to individual grid cells. This would involve repeatedly allocating and deallocating memory for storing the related Jacobian matrix, which is very expensive to re-calculate and can offset the time gains of applying the simplification (Santillana et al., 2010). In order to avoid this computational overhead, we follow Shen et al. (2020) to pre-assemble a small number of subsets of the full chemical mechanism representing the range of conditions in the troposphere and stratosphere, and selecting the most appropriate submechanism locally and on the fly. More specifically, we first split the full mechanism's atmospheric species into N (N=13) different blocks based on similarity of chemical behaviours using a machine learning clustering method. We classify a block as fast if at least one species in the block is fast. We then define the submechanisms as the assemblages of fast blocks (2^N), and we only need to select M (M=20) of them that can encompass the majority of chemical conditions in the atmosphere (see Text S1).

Two problems exist in the Shen et al. (2020) approach. First, the blocks identified by their machine learning approach based solely on minimizing computational time were not chemically coherent. Some species known to be chemically coupled by simple inspection of the mechanism were separated in different blocks. To address this shortcoming, we introduce here a regularization term to characterize species linkages through chemical reactions using graph (network) theory (Figure 2a). In the currently proposed definition (Eq.3), species that are along the same reaction chains and produce similar products have shorter distances with each other, and they are more likely to stay in the same block during the optimization process after imposing the regularization term in the cost function (Eq.4). When compared to Shen et al. (2020), the N species blocks in this work are chemically coherent. To our knowledge, the proposed approach is the first quantitative model that is able to incorporate the species distances to simplify the atmospheric chemistry mechanism. Second, Shen et al. (2020) only achieved 30-40% time-savings. Here we further improve the efficiency of our solution algorithm by not only isolating slow species but also by removing slow reactions from the coupled system yielding further time-savings while not compromising accuracy.

Figure S2 shows the fraction of fast species that needs to be solved using the chemical solver in the global domain as a

https://doi.org/10.5194/gmd-2020-425
Preprint. Discussion started: 16 February 2021


Geoscientific Model Development Discussions

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 they have to use mechanisms of higher complexity (see Section 2.5). 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. With this limited number of 20 chemical regimes, we can exclude ~70% species from the coupled scheme after optimization, compared to the theoretical maximum of 84% derived above (Figure 1).

Figure 2b shows the partitioning of species into the 13 (N=13) blocks (the detailed list of species can be found in Table 1). Oxidants and methane oxidation products are important under all circumstances so blocks 1 and 2 are important in 50-80% gridboxes. Aside from the oxidants, bromine and chlorine radicals (block 3) also play a pervasive role in tropospheric and stratospheric chemistry, and we find they are important in 39% of gridboxes. Iodine reservoirs are inert and they are found to be fast only in 2% of the gridboxes. Our algorithm can also largely separate anthropogenic VOCs from biogenic ones, although a few such species may overlap because many of them 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 S3). 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 S4). 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. Halocarbons are relatively inert in the atmosphere and they are found to be important in <0.1% gridboxes.

Figure 2c shows the network of these 13 species blocks in the atmospheric 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 3 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 over the ocean (Figure S5). The other 9 submechanisms have higher complexity and are mainly used in the lower troposphere over the continents (Figure S5). Only 0.05% of gridboxes need to use the full chemical mechanism, as defined by the 21th submechanism.

https://doi.org/10.5194/gmd-2020-425
Preprint. Discussion started: 16 February 2021


Geoscientific Model Development Discussions

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 4b-c shows the median error (see the definition in Eq.5) 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 12 months to examine the sensitivity to different δ. For each δ, we test the effects of using two strategies, including isolating slow species and removing slow reactions. By isolating slow species, 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, we can reduce the CPU time by 44-49% and the error remains at 0.53-1.0%. When using the δ of 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 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 S6), 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 S7). The error is below 0.5% everywhere for key species like O_3, OH and sulfate, and is 1-2% for NO_2 (Figure S8). Using a higher threshold of δ (> 1500) only leads to marginal improvement in computer time but the error quickly increases.

3.3 Porting our algorithm to a new chemical mechanism

Our algorithm can easily accommodate changes in the chemical mechanism. Figure S9 shows the diagram for adding new species into the mechanism. The location of the new species can be easily determined by its chemical family and the percentage of gridboxes that treat this species as important when averaged globally. In order not to compromise the computational efficiency, the basic rule is to not mix faster species with slower ones. Our algorithm is robust to misplacements of the new species, which may offset the time gains but will not enlarge the error.

To test this approach, 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 fundamental 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 S9. After running the new version of the model for 12 months, our reduced algorithm shows consistent and even better 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 4d).

4. Conclusions

The high computational cost of chemical integration has been a barrier for the inclusion of atmospheric chemistry in Earth system models. Previous research has proposed a variety of ways to speed up the chemical operator, all involving some loss

https://doi.org/10.5194/gmd-2020-425
Preprint. Discussion started: 16 February 2021


Geoscientific Model Development Discussions

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%.

In our algorithm, we first partition atmospheric species into 13 blocks using a novel machine learning approach that analyzes the species network structures and their production and loss rates. We develop a method to quantify the chemical distances of all species in the atmosphere, which plays a critical role in making our algorithm chemically coherent. We then pre-select 20 submechanisms, as defined by unique assemblages of the species blocks, to encompass the vast majority of chemical environments in the atmosphere and then pick locally on the fly which submechanism to use based on species' production and loss rates. Besides isolating slow species and solving them using a first-order explicit method, we also remove the very slow reactions in each submechanism. Our method can reduce the chemical integration time by 50% and produce error of less than 1%, with no error growth. We have also demonstrated that we can easily port our algorithm to a new chemical mechanism without compromising its performance.

Our method has many advantages over previously proposed approaches to reduce chemical mechanism: (1) it is highly chemically coherent and can be ported to different atmospheric models easily; (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 will significantly ease the computational bottleneck and will facilitate the inclusion of comprehensive atmospheric composition into 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).

https://doi.org/10.5194/gmd-2020-425
Preprint. Discussion started: 16 February 2021


Geoscientific Model Development Discussions

290 **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
295 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,
300 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,
305 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,
310 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,
315 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.
320 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.

https://doi.org/10.5194/gmd-2020-425
Preprint. Discussion started: 16 February 2021

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.

https://doi.org/10.5194/gmd-2020-425
Preprint. Discussion started: 16 February 2021

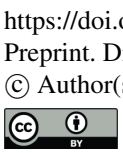
Geoscientific Model Development Discussions
Open Access
EGU

340 **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%
Biogenic VOCs	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%
	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%

https://doi.org/10.5194/gmd-2020-425
Preprint. Discussion started: 16 February 2021



Geoscientific Model Development Discussions

[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 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.

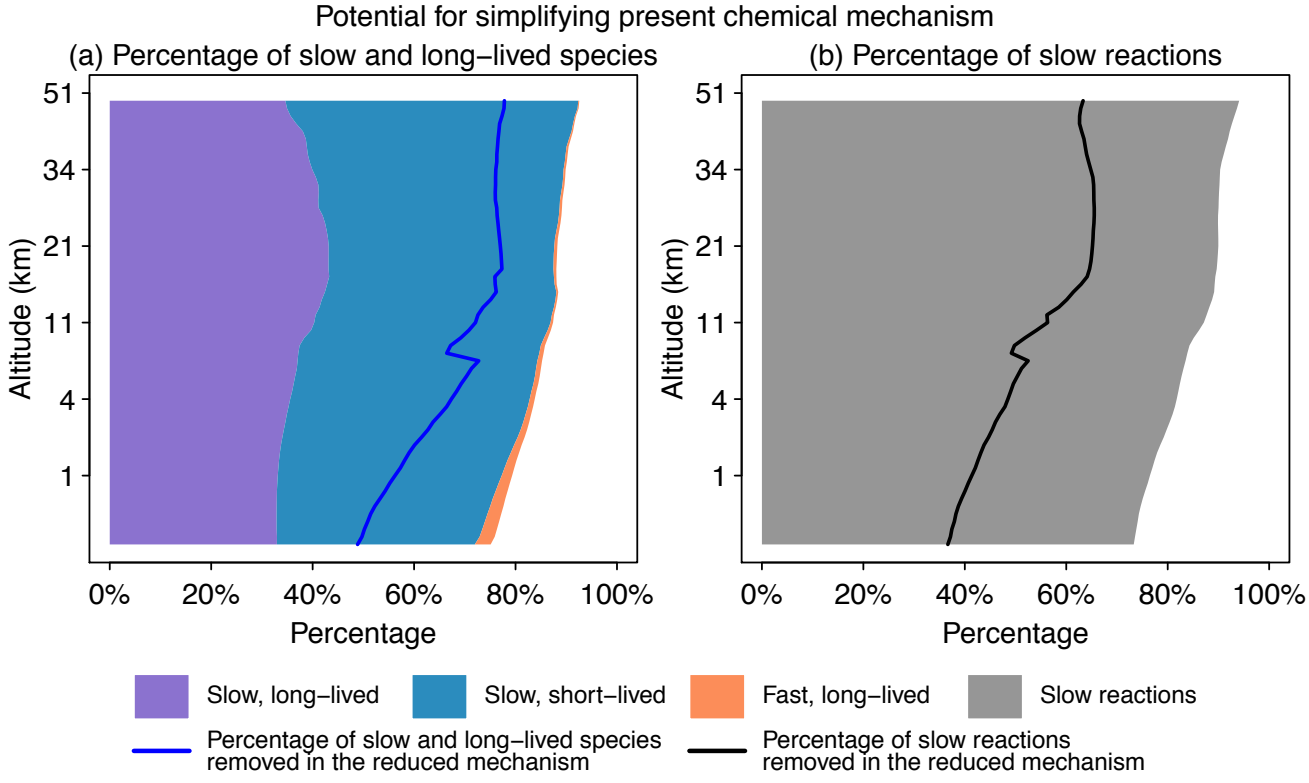
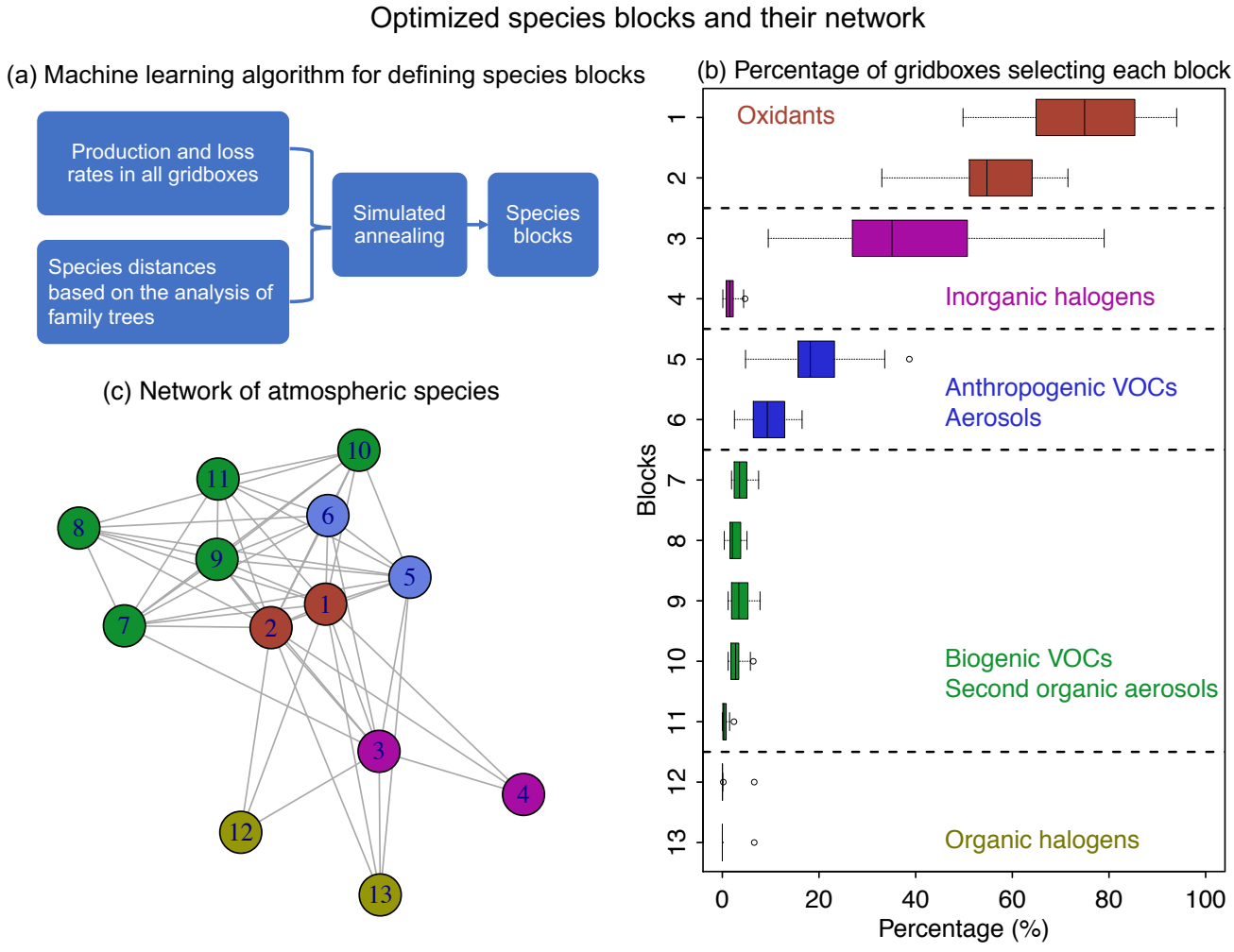


Figure 1. **The potential of simplifying the full chemical mechanism at different altitudes.** Panel (a) shows the percentage of slow and long-lived species by altitude when averaged globally on Aug 1st 2013 at 0 GMT. We use a threshold of 500 molecules cm^{-3} s^{-1} to partition fast and slow species, and a lifetime of 10 days to define long-lived and short-lived species. The blue line denotes for the percentage of slow and long-lived species 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 removed in the reduced mechanism.

https://doi.org/10.5194/gmd-2020-425
Preprint. Discussion started: 16 February 2021

Geoscientific Model Development Discussions



355

Figure 2. **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 all species in each block as fast and short-lived. A more detailed list of species in each block can be found in Table 1. We use a threshold of 500 molecules cm^{-3} s^{-1} to partition fast and slow species. Panel (c) is the
360 network of species blocks. A connection means that at least two species from these two blocks have appeared in the same reaction. Block 7 is mixed with both anthropogenic and biogenic species, and here we treat it as a biogenic block. The distance between the two blocks is proportional to the block distance as defined by Eq. 3.

https://doi.org/10.5194/gmd-2020-425
Preprint. Discussion started: 16 February 2021

Geoscientific Model Development
Discussions
Open Access
EGU

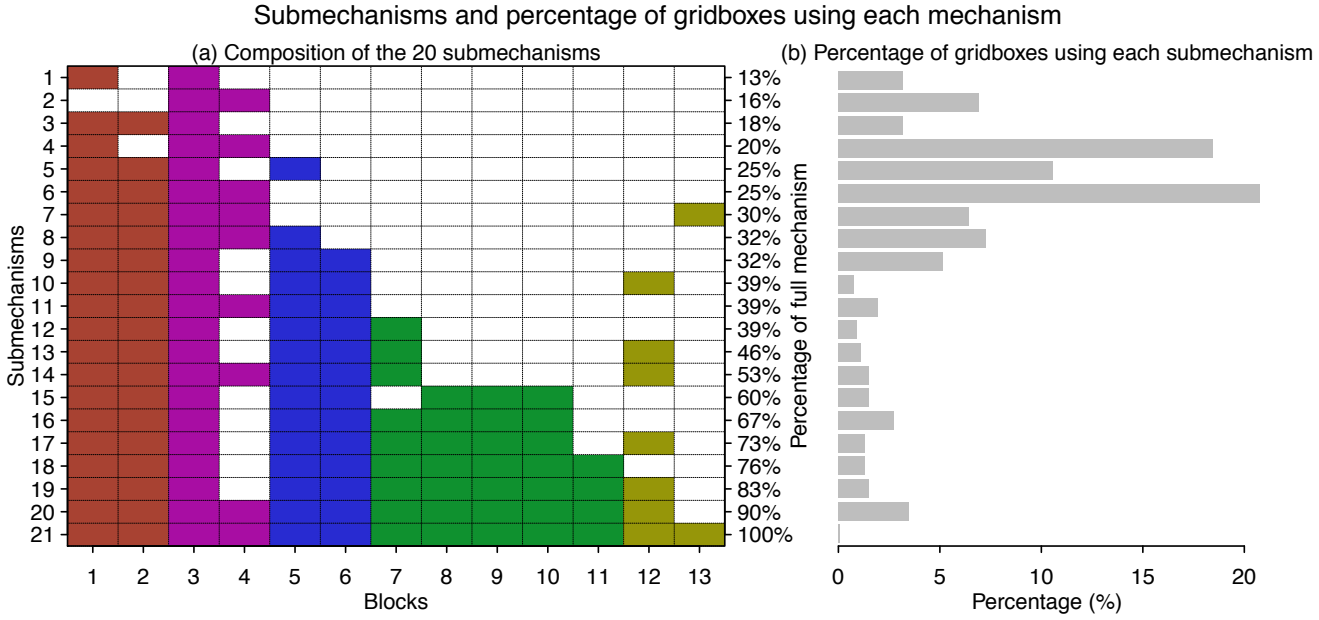


Figure 3. **Submechanisms and percentage of gridboxes using each mechanism**. Panel (a) shows the composition of the 20 submechanisms and the percentage of species from the full mechanism that are treated as fast in each of them. The 21st mechanism is the full mechanism. Colors denote species block types as defined in Figure 2. Panel (b) shows the percentage of gridboxes using each submechanism.

https://doi.org/10.5194/gmd-2020-425
Preprint. Discussion started: 16 February 2021

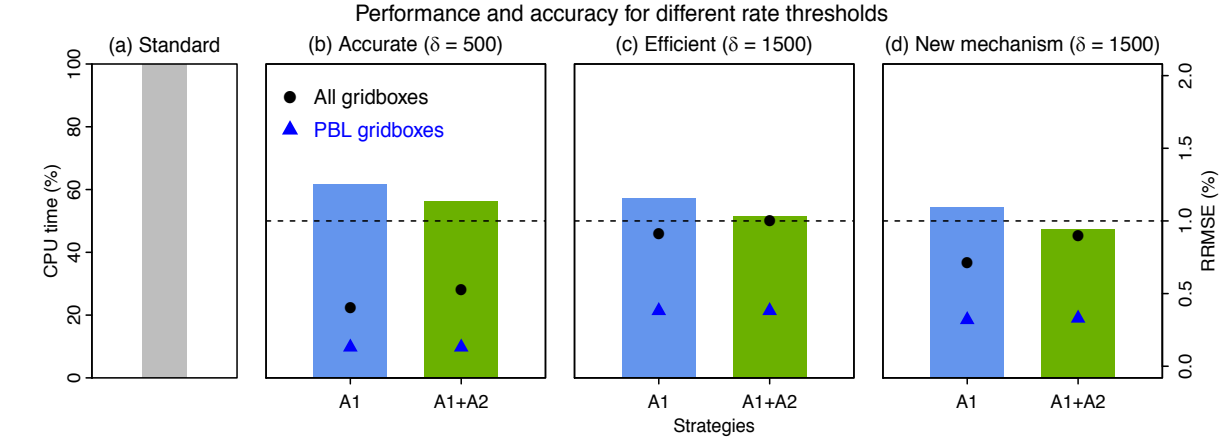


Figure 4. **Performance and accuracy of the adaptive chemical mechanism**. We test the performance of this adaptive method by (A1) removing slow species (P_i or $L_i > \delta$) and (A2) removing unimportant reactions (reaction rate < 10 molecules $cm^{-3}s^{-1}$). The unit of δ is molecules cm^{-3} s^{-1}. 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, as demonstrated by Panel (a). For (b) and (c), we use the δ as 500 and 1500 molecules cm^{-3} s^{-1} in GEOS-Chem 12.0.0 that has 228 species and 724 reactions. For (d), 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 (M) is 20.