Supplementary materials

Text S1. Solving for species blocks using machine learning

We build the $f(M, N)$ of Eq. 4 following a previous study (Shen et al., 2020). For each gridbox j, we diagnose species i as fast or slow following the definition of Section 2.5. Each $y_{i,j}$ will be assigned to one of the N blocks. 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 is slow. Thus, the fraction Z_1 of all species that needs to treated as fast can be written as

$$Z_1 = \frac{1}{\Omega} \sum_j \sum_i y_{i,j} \qquad (S1)$$

Where Ω is the number of $y_{i,j}$. As mentioned in Section 2.5, gridboxes that do not correspond to any of the M submechanisms need to be matched to one of the M by moving some blocks from slow to fast. 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 Z_2 of species that needs to be treated as fast over the global domain is given by:

$$f(M,N) = \frac{1}{\Omega} \left(\sum_{D_1} \sum_i y_{i,j} + \sum_{D_2} \sum_i y^*_{i,j} \right) \qquad (S2)$$

where D_1 are the gridboxes that can be represented by the M chemical regimes, and D_2 are the gridboxes that are represented by other regimes and must be matched to the M regimes.

We use the simulated annealing to solve for the species block that has a minimum of cost function. We start from a randomly generated partitional 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 decreases gradually as the optimization proceeds.

The explicit solution by Eq. 3 does not strictly conserve mass (Shen et al., 2020), which can result in slowly growing mass of reactive halogen species in the stratosphere due to the long lifetime of halogen family there. To avoid this effect, 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.

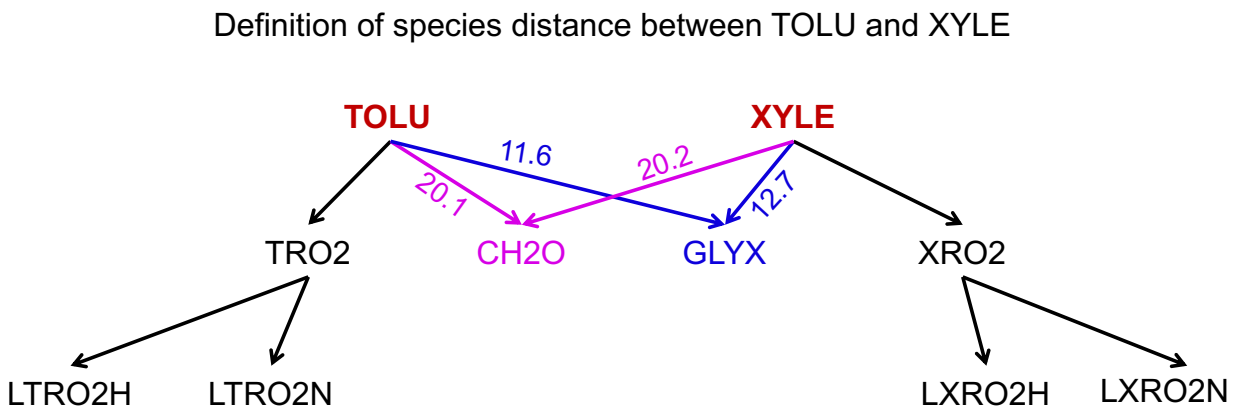

Fig. S1. Definition of species distances for TOLU (toluene) and XYLE (xylene) using the analysis of family trees in the graph theory. The number denotes for the distance between species as calculated by Eq. 3. The shortest path from TOLU to XYLE is TOLU-GLYX-XYLE in this graph, where GLYX is glyoxal.

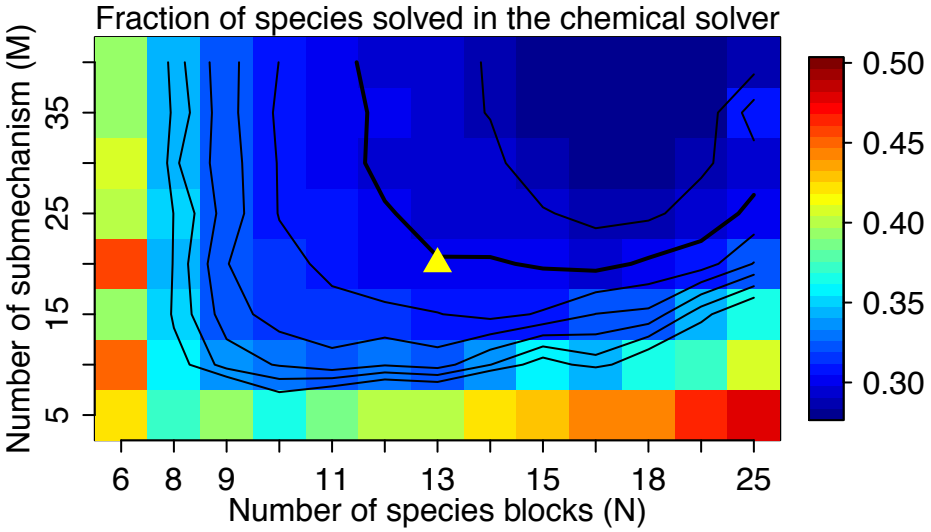

Fig. S2. The fraction of fast species solved in the chemical solver as a function of *M* and *N*. We use *M*=20 and *N*=13 in our work, as shown by the triangle in the figure. We use a threshold δ of 500 molecules cm^{-3} s^{-1} to partition the fast and slow species. The contour lines are spaced by 0.01 with the bold line for 0.30.

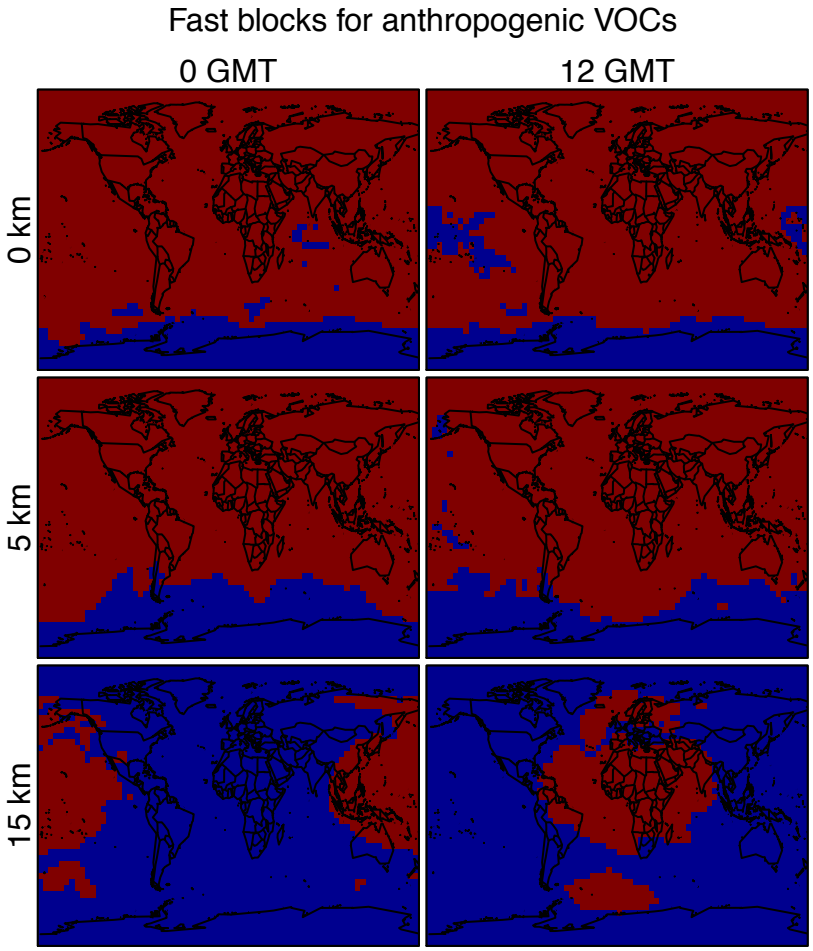

Fig. S3. Locations of anthropogenic blocks (block 5 and 6) that need to be solved as fast (red color). Results are shown on August 1 2013 at 0 and 12 GMT using a threshold δ of 500 molecules cm^{-3} s^{-1} to define the fast and slow species, and a lifetime of 10 days to define the short-lived and long-lived species.

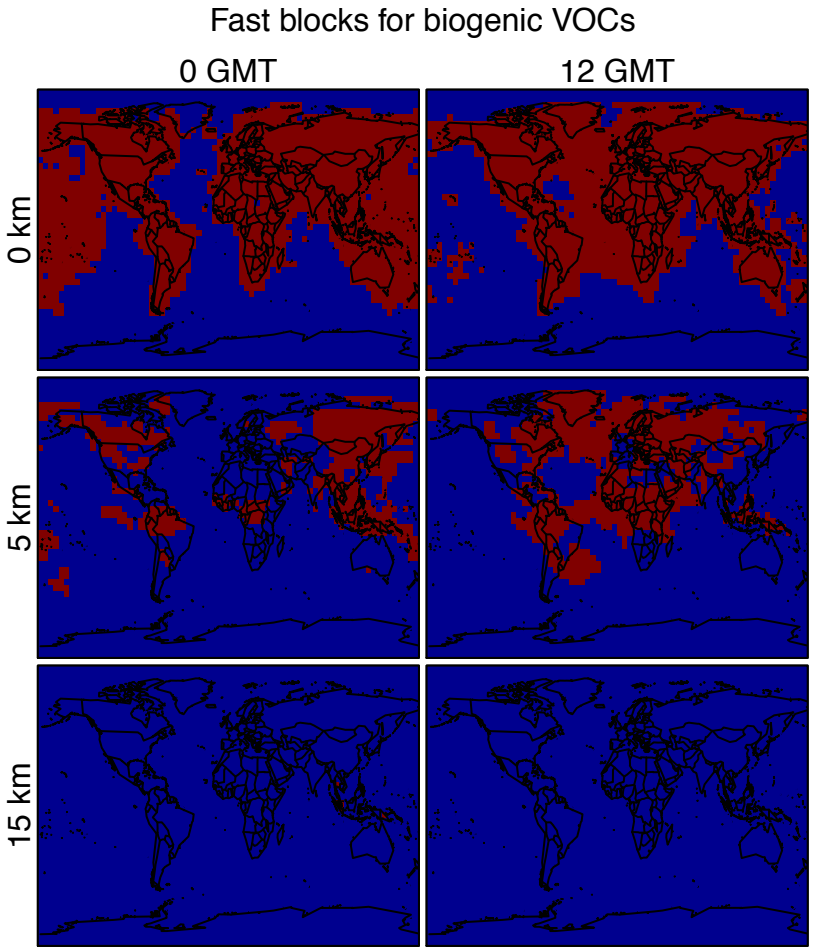

Fig. S4. Locations of anthropogenic blocks (block 7-11) that needs to be solved as fast (red). Results are shown on August 1 2013 at 0 and 12 GMT using a threshold δ of 500 molecules cm^{-3} s^{-1} to define the fast and slow species, and a lifetime of 10 days to define the short-lived and long-lived species.

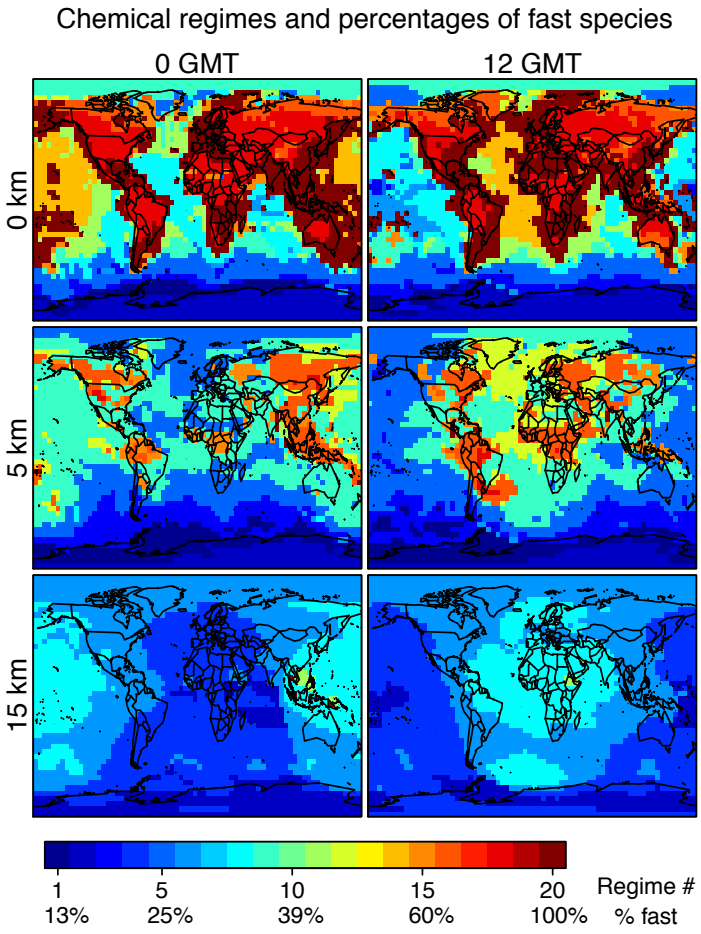

Fig. S5. Chemical mechanism complexity needed in different regions of the atmosphere. The Figure identifies the submechanism from Fig. 3 needed to simulate a given GEOS-Chem gridbox on August 1 2013 at 0 and 12 GMT. The percentage of species treated as fast in that chemical regime is shown on the colorbar and more details are in Table 1. We use a threshold δ of 500 molecules cm^{-3} s^{-1} to partition the fast and slow species.

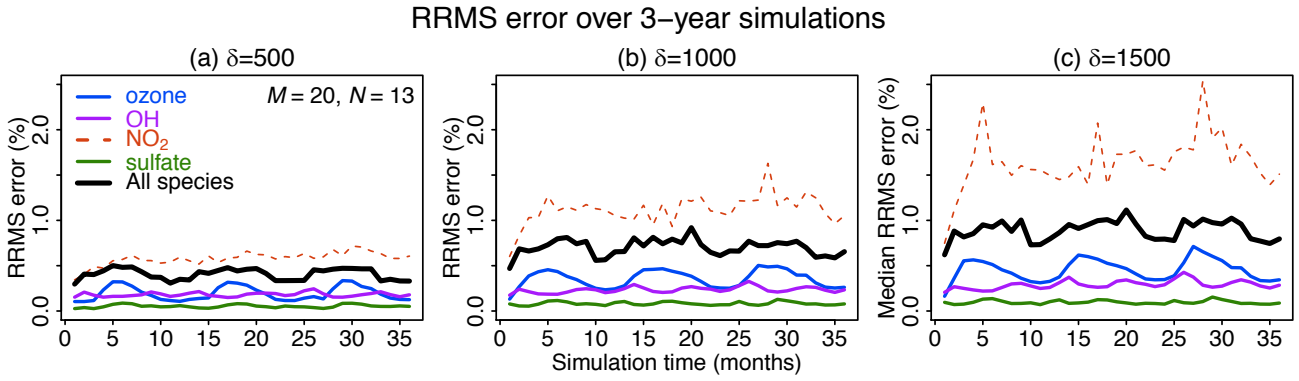

Fig. S6. Accuracy of the adaptive reduced chemistry mechanism algorithm over a three-year GEOS-Chem simulation (see text). The accuracy is measured by the Relative Root Mean Square (RRMS, see Eq. 5) error relative to a simulation including the full chemical mechanism. Results are shown for the median RRMS error across all species in the mechanism and more specifically the RRMS error for ozone, OH, NO_2, and sulfate. For NO_2 (dashed red line), we disregard the errors in regions where its concentration is below 10^7 molecules cm^{-3}. The calculation removes slow (P_i and $L_i < \delta$ molecules cm^{-3} s^{-1}) and slow reactions (rate < 10 molecules cm^{-3} s^{-1}).

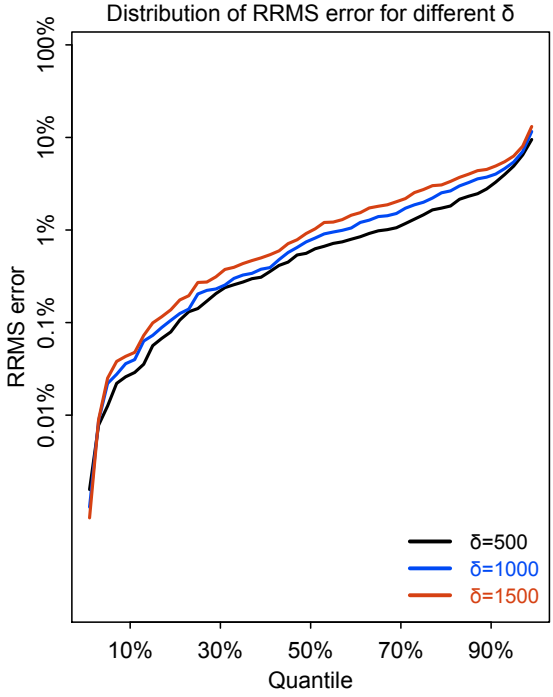

Fig. S7. Distribution of RRMS error for different thresholds δ (N=13 and M=20). The y axis is on the logarithmic scale.

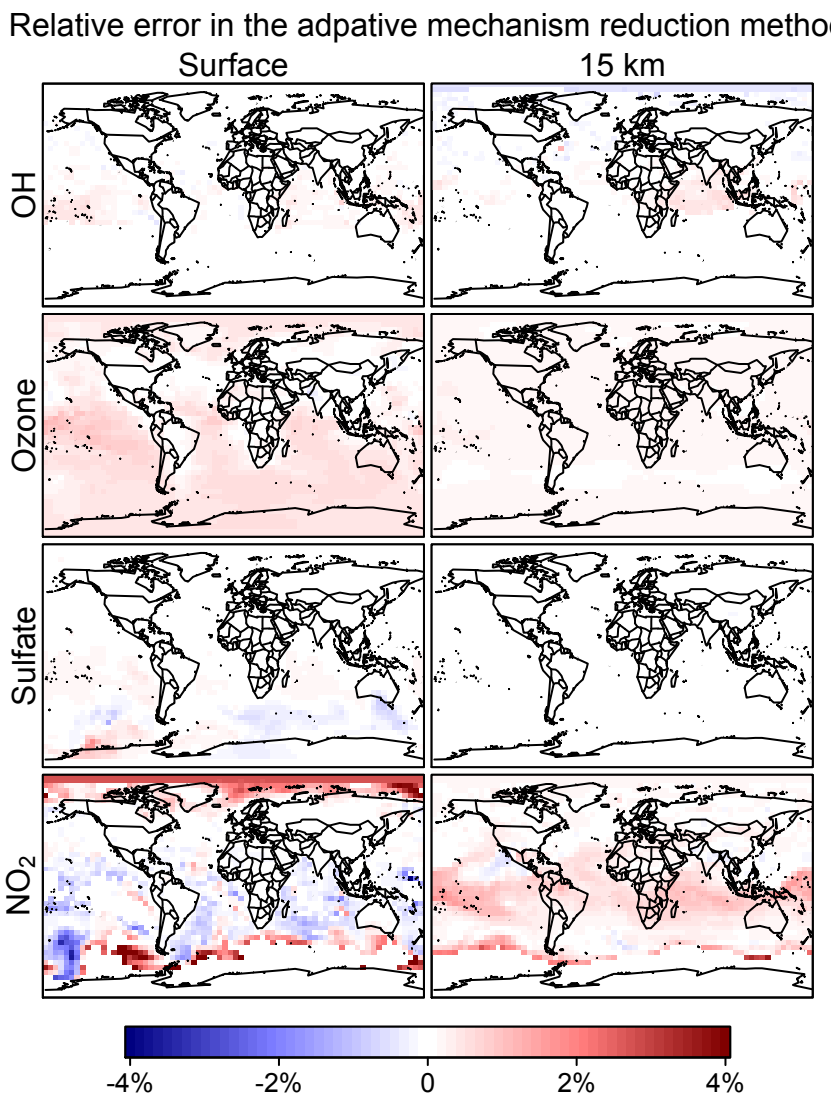

Fig. S8. Relative error from the adaptive mechanism reduction method after three years of simulation in the GEOS-Chem global 3-D model for tropospheric-stratospheric chemistry. The figure shows relative differences of 24-h average OH, ozone, sulfate and NO_2 concentrations relative to the full-chemistry simulation on the last day of the three-year simulation. The calculation removes slow (P and L < 1500 molecules cm^{-3} s^{-1}) and slow reactions (rate < 10 molecules cm^{-3} s^{-1}). The number of blocks (*N*) is 13 and the number of chemical regimes (*M*) is 20.

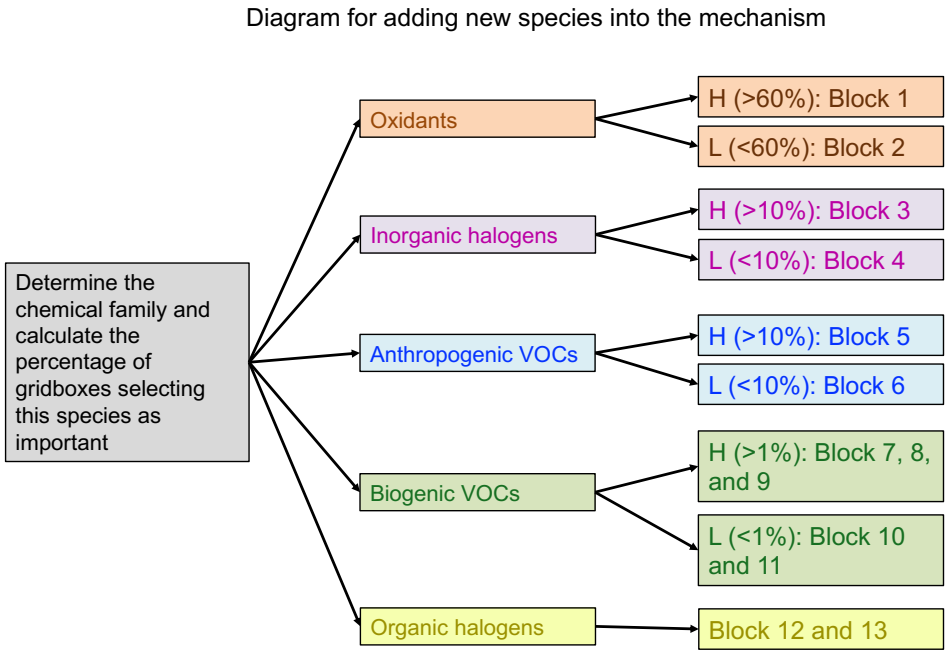

Fig. S9. Diagram of adding a new species to the mechanism. The location of the species is determined by its chemical family and the percentage of gridboxes (in parentheses in the last column) that select this species as fast. We use a threshold of 500 molecules cm^{-3} s^{-1} to partition fast and slow species.