AGU PUBLICATIONS

JGR

Journal of Geophysical Research: Atmospheres

RESEARCH ARTICLE
10.1002/2017JD026487

Key Points:
- Lagrangian predictions of ozone mixing ratios are viable diagnostic tools provided that ensemble averaging and sensitivity testing are employed
- Ozone mixing ratios from global models and reanalysis data have removable systematic errors
- Sensitivity calculations identify the most realistic region of model parameter space and reveal subtle dynamical relationships

Correspondence to:
J. W. Bergman,
j.w.bergman921@gmail.com

Citation:
Bergman, J. W., L. Pfister, D. E. Kinnison, E. J. Hintsa, and T. D. Thornberry (2017), The viability of trajectory analysis for diagnosing dynamical and chemical influences on ozone concentrations in the UTLS, *J. Geophys. Res. Atmos.*, *122*, doi:10.1002/2017JD026487.

Received 11 JAN 2017
Accepted 29 APR 2017
Accepted article online 4 MAY 2017

The viability of trajectory analysis for diagnosing dynamical and chemical influences on ozone concentrations in the UTLS

J. W. Bergman[1,2], L. Pfister[3], D. E. Kinnison[2], E. J. Hintsa[4], and T. D. Thornberry[5]

[1]Bay Area Environmental Research Institute, Petaluma, California, USA, [2]Atmospheric Chemistry Observations and Modeling Laboratory, National Center for Atmospheric Research, Boulder, Colorado, USA, [3]Earth Science Division, NASA Ames Research Center, Moffett Field, California, USA, [4]Global Monitoring Division, NOAA Earth System Research Laboratory, Boulder, Colorado, USA, [5]Chemical Sciences Division, NOAA Earth System Research Laboratory, Boulder, Colorado, USA

Abstract To evaluate the utility of trajectory analysis in the tropical upper troposphere/lower stratosphere, Lagrangian predictions of ozone mixing ratio are compared to observations from the Airborne Tropical TRopopause EXperiment. Model predictions are based on backward trajectories that are initiated along flight tracks. Ozone mixing ratios from analysis data interpolated onto "source locations" (at trajectory termini) provide initial conditions for chemical production models that are integrated forward in time along parcel trajectories. Model sensitivities are derived from ensembles of predictions using two sets of dynamical forcing fields, four sets of source ozone mixing ratios, three trajectory formulations (adiabatic, diabatic, and kinematic), and two chemical production models. Direct comparisons of analysis ozone mixing ratios to observations have large random errors that are reduced by averaging over 75 min (~800 km) long flight tracks. These averaged values have systematic errors that motivate a similarly systematic adjustment to source ozone mixing ratios. Sensitivity experiments reveal a prediction error minimum in parameter space and, thus, a consistent diagnostic picture: The best predictions utilize the source ozone adjustment and a chemical production model derived from Whole Atmosphere Community Climate Model (a chemistry-climate model) chemistry. There seems to be slight advantages to using ERA-Interim winds compared to Modern-Era Retrospective Analysis for Research and Applications and to using kinematic trajectories compared to diabatic; however, both diabatic and kinematic formulations are clearly preferable to adiabatic trajectories. For these predictions, correlations with observations typically decrease as model error is reduced and, thus, fail as a model comparison metric.

Plain Language Summary To evaluate the utility of trajectory analysis in the tropical upper troposphere/lower stratosphere, predictions of ozone mixing ratio are compared to observations from the Airborne Tropical TRopopause EXperiment. Model predictions are based on backward trajectories that are initiated along flight tracks. Ozone mixing ratios from analysis data interpolated onto "source locations" (at trajectory termini) provide initial conditions for chemical production models that are integrated forward in time along parcel trajectories. Model sensitivities are derived from ensembles of predictions using two sets of dynamical forcing fields, four sets of source ozone mixing ratios, three trajectory formulations, and two chemical production models. Direct comparisons of analysis ozone mixing ratios to observations have large random errors that are reduced by averaging over 75 min (~800 km) long flight tracks. These averaged values have systematic errors that motivate a similarly systematic adjustment to source ozone mixing ratios. Sensitivity experiments reveal a prediction error minimum in parameter space and, thus, a consistent diagnostic picture: The best predictions utilize the source ozone adjustment and chemical production derived from National Center for Atmospheric Researchs Whole Atmosphere Community Climate Model. For these predictions, correlations with observations typically decrease as model error is reduced and, thus, fail as a model comparison metric.

1. Introduction

The tropical upper troposphere/lower stratosphere (UTLS) is an important transitional region. Dynamically, the transition is from highly turbulent conditions in the troposphere, where vertical transport occurs rapidly via convection, to much more quiescent conditions in the stratosphere, where vertical transport is a slow process balanced primarily by radiative heating. In terms of modeling, the transition is one from a heavy

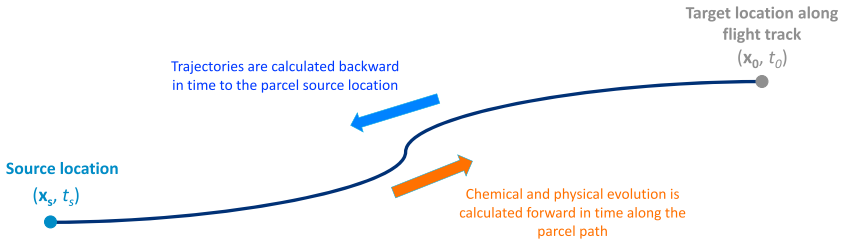

Figure 1. A schematic representation of the Lagrangian prediction model. A backward trajectory calculation is initiated at each target location ($\mathbf{x_0}$, t_0), along the flight track of an airborne observing station. Trajectory calculations determine air parcel paths and terminate at source locations ($\mathbf{x_s}$, t_s). Chemical and physical process models are initialized at the source location and integrated forward in time along parcel paths to the target location, where the model prediction is compared to observed values.

dependence on parameterizations of unresolved processes to one for which numerical models can resolve the dominant circulation features. Observationally, the UTLS encompasses the highest altitudes attainable for aircraft observations and the lowest viable altitudes for many satellite observations. In that regard, it is important that the tropical UTLS lies above most convective activity and clouds, making it safe for aircraft travel and providing a cloud-free line of sight for satellite-based measurements. The overlap of aircraft-based and satellite-based observations makes the UTLS an important region for validating satellite data. As an instrument-accessible region that is directly affected by convective outflow, the tropical UTLS is important for both studying the impacts of convection on the chemical makeup of the atmosphere and for defining the lower boundary conditions for stratospheric studies. The latter role has prompted its identification as the "gateway to the stratosphere" and has motivated many scientific endeavors (see reviews by *Fueglistaler et al.* [2009] and *Randel and Jensen* [2013]).

Lagrangian calculations that track air parcel trajectories backward in time form the foundation of a potentially powerful diagnostic technique for determining dynamical influences on chemical concentrations, particularly when combined with aircraft observations. This technique, schematically depicted in Figure 1, initializes trajectory calculations at "target" locations ($\mathbf{x_0}$, t_0), where observational data are available along flight tracks. Parcel paths are traced backward in time using wind field data to "source" locations ($\mathbf{x_s}$, t_s). Note that the word *source* is a convenient definition for the final (in a backward sense) location of each trajectory and does not necessarily indicate a physically meaningful chemical source. Process models that "predict" (in a statistical modeling sense, not in a forecast sense) chemical concentrations at the target locations are initialized at the source locations and integrated forward in time along parcel paths. For our study, the process models determine net chemical production of ozone via photolysis and chemical reactions among other atmospheric constituents. However, process models can also include phase changes, which are important for water vapor or small-scale mixing between the idealized air parcel and the environment through which it is transported. In principle, this technique provides a comprehensive analysis of the interactions among dynamical and chemical processes that determine the chemical makeup of the atmosphere. As a result, these calculations have been at the core of studies that examine distributions of water vapor, ozone, carbon monoxide, bromocarbons, and clouds in the tropical UTLS [e.g., *Folkins et al.*, 1997; *Pfister et al.*, 2001; *Pisso et al.*, 2009; *Law et al.*, 2010; *Ploeger et al.*, 2011, 2012; *Ashfold et al.*, 2012; *Ueyama et al.*, 2014].

There are, however, complications that obscure the interpretation of Lagrangian model predictions. In the context of the experimental design illustrated in Figure 1, we divide these complications into those that affect the air parcel trajectories, those that affect source concentrations, and those that affect chemical production along the parcel paths. These distinctions are blurred by coupling, but they nevertheless form a useful framework for diagnosing prediction error. Complications with the trajectories are related primarily to the wind fields that drive them. These fields are typically supplied by "analysis data sets" (operational analyses and reanalyses) that are derived with state-of-the-art fusions of observations with numerical modeling. While analysis data are available at horizontal resolutions of ~1°, vertical resolutions of ~1 km near the tropical tropopause, and temporal resolutions of 3–6 h, they are ultimately limited by the space-time sampling of observational networks and are subject to substantial uncertainties in regions, such as the tropical UTLS, where in situ observations are sparse. These uncertainties direct us toward a probabilistic

Table 1. A List of Different Prediction Model Configurations for the Sensitivity Experiments[a]

Category	Number of Versions	Versions
Wind/heating rate data source	2	MERRAERA-Interim
Trajectory formulation	3	Adiabatic (isentropic)Diabatic (potential temperature coordinates)Kinematic (pressure coordinates)
Ozone data	4	WACCMAdjusted WACCMERA-InterimAdjusted ERA-Interim
Ozone production model	2	WACCM ChemistryNone
Trajectory Length	4	5 days10 days20 days30 days

[a]Calculations are performed with model alterations in five categories: The wind and heating rate data for the trajectory calculations (two versions), how the trajectory calculations are formulated (three versions), the ozone source data (four versions), the ozone production model (two versions), and the length of the trajectories (four versions). All combinations of these versions are utilized yielding a total of $2 \times 3 \times 4 \times 2 \times 4 = 192$ model configurations.

view of parcel trajectories in which the parcel location is specified via a probability distribution that widens with time. This widening is a form of dispersion that renders the information content of a single-parcel trajectory negligible and necessitates the use of trajectory ensembles. *Bergman et al.* [2015] and *Bergman et al.* [2016] investigated dispersion of parcel location probability in the UTLS and found that dispersion rates are ~3°/d horizontally and ~2–3 hPa/d vertically. These rates imply that parcels, initially confined within an analysis grid box, disperse to fill the tropical troposphere in approximately 1 month. However, even widely dispersed parcels have nonuniform probability densities characterized by large-scale features (on the order of 30°) that provide valuable information. This characteristic of parcel dispersion implies that extended trajectory calculations (trajectory lengths greater than a few days) are potentially, but not necessarily, useful; their utility depends on the scientific question being addressed and requires thoughtful experimental design.

We address these issues using Lagrangian predictions of ozone (O_3) that we compare to observed values from ATTREX (Airborne Tropical TRopopause EXperiment; see *Jensen et al.* [2017] for details) near the tropical tropopause over the Pacific during the boreal winters of 2013 and 2014. Using 165 h of ozone observations (one measurement every 10 s) between the altitudes of 14 and 19 km and 192 different model configurations, sensitivity tests compare prediction errors from different configurations. These tests help us determine appropriate ensemble sizes for model-observational comparisons, determine the model parameter subspace that includes the best model configuration, and determine which of the model components (e.g., ozone source data, trajectory calculations, and chemical production) lead to the most model uncertainty. Most important, though, these tests help us determine which scientific questions can be effectively addressed with Lagrangian calculations and which cannot.

2. Method

2.1. Experimental Design

Our analysis is based on a battery of modeled ozone calculations using Lagrangian parcel paths as schematically represented in Figure 1. Calculations of these paths are initiated at target locations along ATTREX flight tracks, where ozone mixing ratios have been measured and run backward in time to determine source locations. A chemical production model is then initiated with ozone mixing ratios at the source locations (obtained from analysis data) and run forward in time along the parcel paths to the target locations. The modeled mixing ratios at the target locations are then compared to observational values to quantify model error. We divide sources of error into three components: errors related to the trajectory calculations, which focus on the determination of source locations, errors in the specified ozone source mixing ratios, and errors related to the chemical production model. It is important to recognize that this division, while convenient, is arbitrary and coupling exists among the three components that prevents unambiguous error attribution. The error diagnosis is performed via evaluation of model sensitivities to changes of these components. To obtain these sensitivities, we perform trajectory calculations using three model formulations: "adiabatic," for which the parcel paths are isentropic (i.e., vertical motion is neglected), "diabatic," for which vertical motion is determined by total diabatic heating rates, and "kinematic" trajectories for which vertical motion is determined by pressure velocity. We also use two sets of "forcing fields" (wind and heating rate data) for the trajectories, four different sets of source ozone data, four different trajectory lengths to determine the source locations, and two different chemical production models (one of which is a null model that assumes no chemical production). As detailed in Table 1, all combinations of these model alterations are utilized for

Table 2. A List of Flights Including the Date of the Flight, the Primary Target Location, and Duration of the Continuous Observations Used in the Analysis

Date	Target Region	Duration of Continuous Observations (Hours)
5 Feb 2013	Eastern Pacific	18.75
21 Feb 2013	Central Pacific	21.25
26 Feb 2013	Eastern Pacific	18.75
1 Mar 2013	Central Pacific	18.75
16 Jan 2014	Southern California to Guam	13.75
12 Feb 2014	Western Pacific	8.75
16 Feb 2014	Western Pacific	8.75
4 Mar 2014	Western Pacific	8.75
6 Mar 2014	Western Pacific	12.5
9 Mar 2014	Western Pacific	11.25
11 Mar 2014	Western Pacific	11.25
13 Mar 2014	Guam to Southern California	12.5

the sensitivity tests, yielding a total of 192 different model configurations each predicting ozone mixing ratios at 59,400 target locations.

2.2. Trajectory Calculations

The trajectory model used is the Lagrangian particle model of *Bergman et al.* [2012], which is a slightly modified version of the *Schoeberl and Sparling* [1995] model. The model uses linearly interpolated winds from analysis data to advance the parcel trajectories back in time with the fourth-order Runge-Kutta method at a time step of 30 min. Parcel pathways are calculated for 30 day back trajectories initiated at 10 s (roughly every 1.8 km) intervals along ATTREX flight tracks. The 30 day length of the trajectories is roughly consistent with the typical transport time from convective detrainment to the tropical tropopause during boreal winter [e.g., *Bergman et al.*, 2012]. The analysis uses a total of 132 75 min (~800 km) continuous flight segments. Flight segments are obtained from both the 2013 deployment (four flights; two targeting the east Pacific and two targeting the central Pacific; Table 2 and Figure 2a) and the 2014 deployment that sampled air primarily over the western Pacific (six flights) but also included two transit flights between Southern California and Guam.

2.3. Analysis and Observational Data

We use two reanalysis data sets to obtain wind/heating rate data for the trajectories. ERA-Interim is produced at the European Centre for Medium-Range Weather Forecasts [*Dee et al.*, 2011]. These data are available 6 hourly at ~0.7° horizontal resolution, which we smooth to 1°, and vertical resolutions of 15 hPa (~1 km) near the tropical tropopause and 50 hPa (~0.6 km) in the midtroposphere. The NASA Modern-Era Retrospective Analysis for Research and Applications (MERRA) [*Rienecker et al.*, 2008; *Rienecker et al.*, 2011] provides 3-hourly data at 1.25° horizontal resolution and 50 hPa (~0.6–3 km) vertical resolution in the mid-to-upper troposphere. ERA-Interim was chosen for its relatively high vertical resolution near the tropopause; MERRA was chosen because of its relationship to ozone values from Whole Atmosphere Community Climate Model (WACCM) (discussed below).

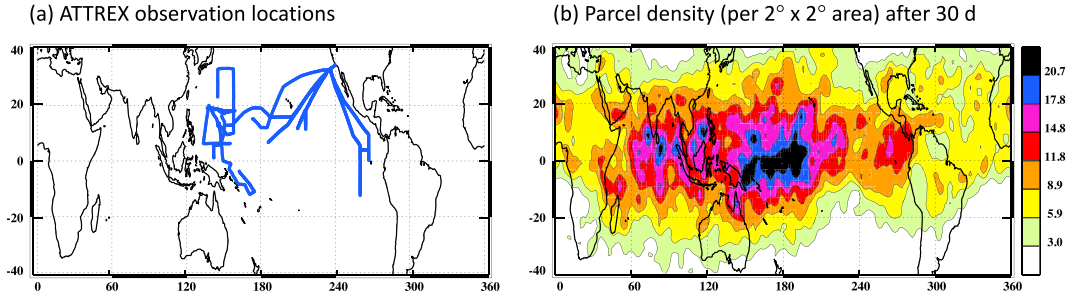

Figure 2. (a) ATTREX flight tracks (target locations) for data used in this study. (b) The source location density (parcels per 2° × 2° horizontal area) at the end of 30 day back trajectory calculations. After 30 days, air parcels that are initially confined to a limited range of target locations spread throughout low latitudes.

Ozone source concentrations are derived from two analysis data sets: daily averaged values from the National Center for Atmospheric Research (NCAR) Whole Atmosphere Community Climate Model (WACCM version 4) [*Marsh et al.*, 2013] and 6-hourly values from ERA-Interim. WACCM is a chemistry climate model (an atmospheric general circulation model containing chemical production and loss equations; details of the chemistry are provided in the supporting information for *Kinnison et al.* [2007]). We use a "specified dynamics" version that includes relaxation terms to dynamical quantities from MERRA (see *Lamarque et al.* [2012] for details). These terms allow WACCM chemistry to act in the context of realistic temperature and wind fields. The ozone fields we used have 2.5° horizontal resolution, 15 hPa (~1 km) resolution near the tropical tropopause, and 40 hPa (~0.5 km) resolution in the midtroposphere. ERA-Interim ozone mixing ratios are obtained from the same data source as the ERA-Interim wind fields discussed above. For the time periods that we are investigating, ERA-Interim assimilates radiance data from three satellites: two ozone profilers, the solar backscatter ultraviolet radiometer and the Aura Microwave Limb Sounder, and a total column ozone instrument, the Ozone Monitoring Instrument [*Fujiwara et al.*, 2017, Figure 10]. These observations are augmented with linear relaxation toward photochemical equilibrium (method developed by *Cariolle and Teyssèdre* [2007]). The pertinent differences between WACCM and ERA-Interim ozone are as follows: WACCM contains daily averages that are determined with a free-running chemistry model, while ERA-Interim contains 6-hourly values, assimilates radiance observations, and uses a crude chemistry model.

We evaluate predicted ozone mixing ratios with measurements taken during ATTREX (see details of the ATTREX mission in *Jensen et al.* [2017]). Our primary source of observational data is from the Unmanned Aircraft System (UAS) Chromatograph for Atmospheric Trace Species (UCATS) developed at the Global Monitoring Division of the National Oceanic and Atmospheric Administration (NOAA) Earth System Research Laboratory (ESRL). Ozone measurements are provided every 10 s and are based on the absorption of ultraviolet (254 nm) light. For the ATTREX flights, the accuracy and precision of ozone mixing ratios from UCATS are estimated to be 5% and ±2 ppbv, respectively. For the 2013 flights, ATTREX employed an additional ozone instrument, the NOAA dual-beam ozone photometer (NOAA-Chemical Sciences Division (CSD)) developed at the Chemical Sciences Division of ESRL. We use comparisons between the two instruments to estimate observational uncertainty. For the ATTREX flights, NOAA-CSD recorded observations every 0.5 s with an accuracy estimated to be ±3% and a precision of 1.1×10^{10} molecules per cm^3 (~4 ppbv at a pressure of 100 hPa and temperature of 200 K).

1. See http://www.esrl.noaa.gov/gmd/hats/airborne/ucats.html for UCATS instrument details.
2. See https://airbornescience.nasa.gov/instrument/UAS-O3 or *Gao et al.* [2012] for NOAA-CSD instrument details.

2.4. The WACCM Chemical Model

We use ozone production and loss rates from WACCM to create a chemical production model as described by *Wang et al.* [2014]. This model integrates an equation for ozone mixing ratio forward in time along each Lagrangian path from its source location to its target location. For this purpose, we calculate the change of ozone using the equation

$$\frac{d_\chi}{dt} = P - L{\cdot}\chi, \tag{1}$$

where χ is the ozone volume mixing ratio, P represents ozone production, and L is the ozone loss per unit mixing ratio. For the sake of simplicity, we have dropped the explicit functional dependence on space and time in (1). To determine the modeled ozone mixing ratio $\chi(\boldsymbol{x_o}, t_o)$ at a target location, we initialize χ with the corresponding source mixing ratio $\chi(\boldsymbol{x_s}, t_s)$ and integrate (1) forward in time using values P and L from WACCM that have been projected onto the parcel path. The form of the loss-rate term ($L \cdot \chi$) is motivated by the fact that chemical loss-rate equations contain explicit dependence on the local ozone mixing ratio and that this form prevents the calculated ozone mixing ratios from becoming negative. This formulation means that our calculations of net ozone production $P - L \cdot \chi$ will differ from the corresponding values in WACCM. The inconsistency between modeled ozone and WACCM chemistry is a source of undetermined uncertainty; however, this method is computationally efficient and is essentially equivalent to running a chemical box model with all important quantities except ozone (e.g., temperature, water vapor, and NO)

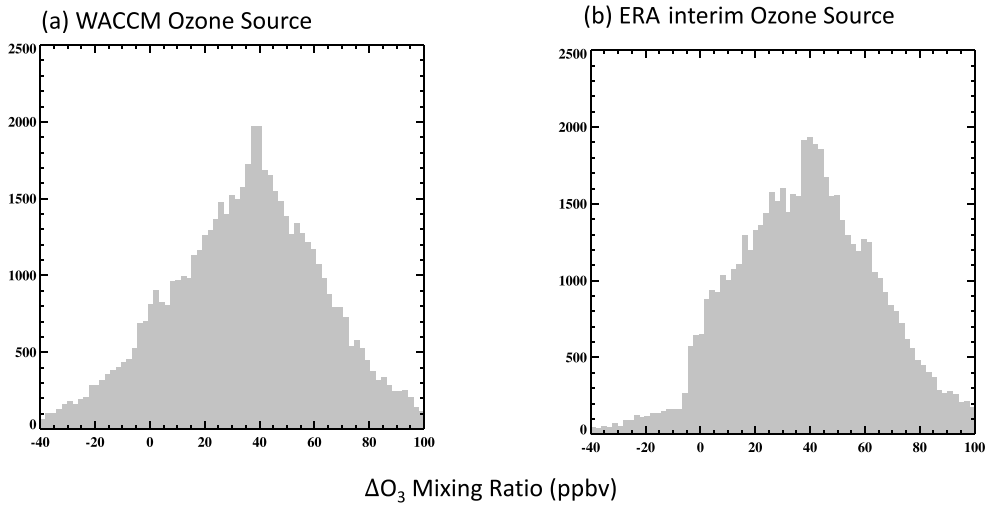

Figure 3. Histograms of ozone production along parcel paths for 30 day trajectories. For these histograms, the WACCM chemical production model is initiated at the source locations using (a) source-adjusted ozone mixing ratios from WACCM and (b) source-adjusted ozone mixing ratios from ERA-Interim. The source adjustment is explained in section 3.2. Net chemical production of ozone is positive for the vast majority of the calculations.

specified from their values in WACCM. Figure 3, which displays values of net production integrated over a 30 day trajectory for a specific model configuration initialized with WACCM (Figure 3a) and ERA-Interim (Figure 3b) ozone, demonstrates that our trajectories occur in a regime of positive net production with typical rates (1–2 ppbv/d) that are consistent with the chemical model results of *Avallone and Prather* [1996] for the lower tropical stratosphere.

2.5. Linear Error Analysis

Our prediction evaluations are based on linear error analyses that compare model predictions to observations. That is, for each observation, or average of observations, x_i, the corresponding model prediction y_i is of the form

$$y_i = a + bx_i + \zeta_i,$$

where ζ_i is a random value uncorrelated with x_i. Model error e_i is defined as the difference between model predictions and observed values

$$e_i = y_i - x_i = a + (b-1)x_i + \zeta_i.$$

It is useful to divide model error into a systematic component

$$e_{\mathrm{sys},i} = a + (b-1)x_i$$

and a random component

$$e_{\mathrm{ran},i} = \zeta_i.$$

In general, we will quantify model error using the "root-mean-square" (RMS) value. That is,

$$e_{\mathrm{RMS}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} e_i^2}$$

quantifies the total model error,

$$e_{\mathrm{sys}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (a + (b-1)x_i)^2}$$

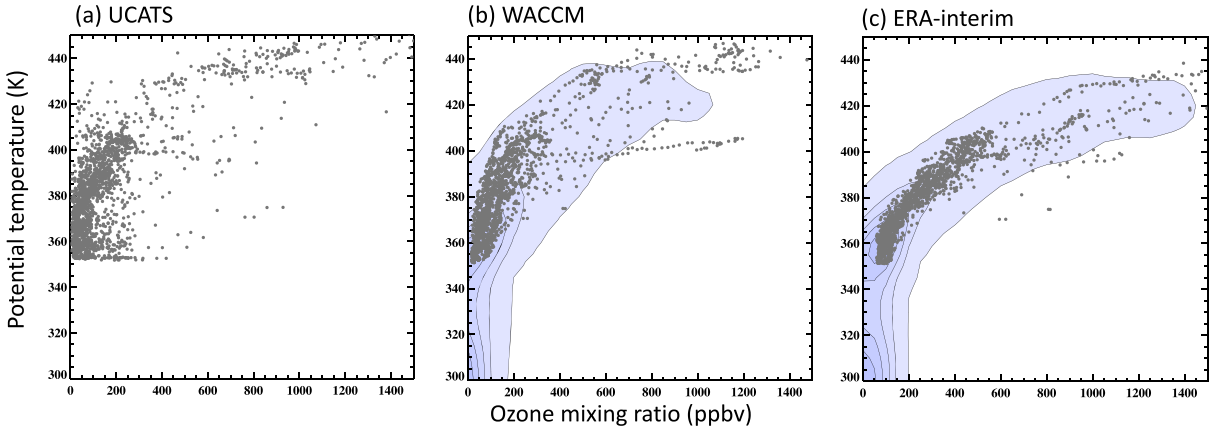

Figure 4. Vertical distributions (in terms of potential temperature) of (a) observed (UCATS) ozone mixing ratios along ATTREX flight tracks, (b) ozone mixing ratios from WACCM projected onto flight tracks (gray dots), and (c) ozone mixing ratios from ERA-Interim projected onto flight tracks (gray dots). For display purposes only, the density of gray dots has been reduced by subsampling, choosing values every 5 min instead of every 10 s. The blue shading in Figures 4b and 4c corresponds to parcel densities at the end of 30 day trajectories (i.e., at source locations) as functions of potential temperature and ozone mixing ratio. These figures demonstrate that ozone mixing ratios at low latitudes are strong functions of potential temperature, that the mixing ratios from analysis data are systematically too large, and that the distributions sampled along ATTREX flight tracks are representative of the distributions at source locations.

quantifies the systematic error, and$e_{\text{ran}} \sim \sigma_\zeta$,where σ_ζ is the standard deviation of random values ζ_i, quantifies the random error. Since the random errors are uncorrelated with the observations, RMS errors satisfy the relation

$$e^2_{\text{RMS}} \sim e^2_{\text{sys}} + e^2_{\text{ran}}.$$

The correlation coefficient r that compares modeled to observed variability is also a commonly used evaluation metric. In our linear analysis, the correlation is related to the random error variance, observational variance, and the regression slope b via

$$r^{2\sim} \frac{b^2\sigma_x^2}{b^2\sigma_x^2 + \sigma_\zeta^2} \tag{2}$$

where σ_x is the standard deviation of the observations. Note that we have, for the sake of simplifying verbiage, assumed the equivalence of RMS and standard deviation. This is only strictly valid for infinite sample sizes but leads to discrepancies of less than 1% for our sample sizes.

3. Evaluation of Analysis Ozone Data

3.1. Evaluation of Analysis Data Along Flight Tracks

Our prediction models, as formulated, are composed of three components that highlight three primary sources of uncertainty: the specified source ozone values, the chemical production model, and the trajectory calculation. These uncertainties are coupled, and so there is no definitive method for disentangling their impacts on modeled ozone errors. However, we can proceed systematically such that our diagnosis provides a consistent interpretation in a comprehensible context. We begin that process by comparing analysis ozone (i.e., from WACCM and ERA-Interim) projected along the flight tracks to observed values, thus obtaining an indication of how reliable ozone mixing ratios at source locations might be.

Figure 4 compares distributions of ozone mixing ratio as functions of potential temperature. The dark gray dots in Figure 4a represent individual 10 s observations from UCATS; the gray dots in Figures 4b and 4c show the corresponding values from WACCM and ERA-Interim (respectively) interpolated onto the flight tracks. These figures indicate (1) that ozone mixing ratios near the tropical tropopause are strong functions of potential temperature and (2) that analysis ozone data tend to overestimate ozone mixing ratios at all altitudes. While these features are determined from data along flight tracks only (i.e., at target locations), they are likely to be representative of ozone distributions at source locations as well. This is demonstrated by the

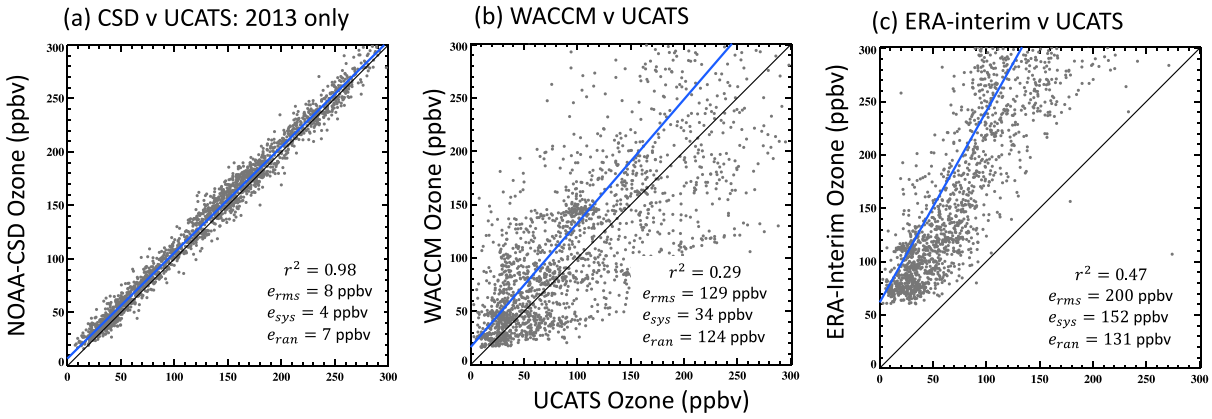

Figure 5. Point-by-point comparisons of ozone mixing ratios along flight tracks for (a) NOAA-CSD observations v UCATS observations for 2013 flights only, (b) WACCM ozone v UCATS, and (c) ERA-Interim ozone v UCATS. For display purposes only, values in Figures 5b and 5c have been subsampled, choosing values every 5 min instead of every 10 s. The blue lines in each panel represent the line of best fit (minimizing random error variance). The analysis data sets have large random errors as well as values that are systematically too large. The comparison between observational data sets demonstrates that observational uncertainty is not an important factor for our analysis.

overlapping of distributions of the gray dots with the blue shading, which represents the source distributions for 30 day trajectories, in Figures 4b and 4c. After 30 days of model integration, parcels are spread throughout the tropics (see the corresponding horizontal parcel density in Figure 2b) compared to the small spatial sampling represented by the flight tracks (Figure 2a). Yet despite the substantial difference in sampling locations, ozone distributions from the analysis data sampled at the flight tracks and after 30 days have similar vertical distributions.

Other important aspects of the analysis versus observation comparisons are illustrated in Figures 5 and 6. Figure 5 displays point-by-point comparisons among data sets: Figure 5a compares observations from the two instruments that were aboard the aircraft during the 2013 flights, Figure 5b compares WACCM ozone to UCATS observations at target locations, while Figure 5c compares ERA-Interim ozone to UCATS. For these figures, we have confined comparisons to those target locations for which UCATS mixing ratios are $\leq$300 ppbv, eliminating high values of ozone observed in the subtropical stratosphere. This helps us focus on the tropical UTLS and provides a more stringent test of the analysis ozone data. When we include the large values of ozone (not shown), we obtain very large correlations ($>$0.9) between analysis data and observations

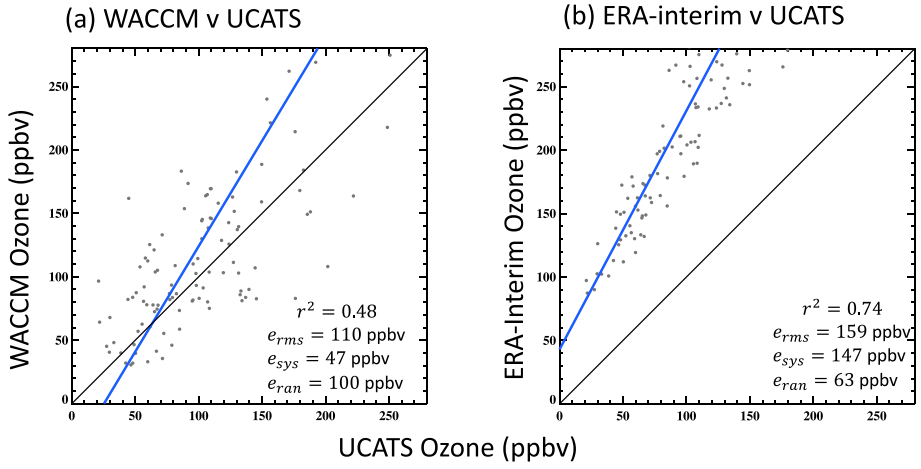

Figure 6. Comparisons of mixing ratios averaged over 75 min flight segments for (a) WACCM ozone v UCATS and (b) ERA-Interim ozone v UCATS. The segments are chosen arbitrarily with some segments containing dives from cruising altitude (~60,000 ft; ~1819 km, to ~45,000 ft; ~14 km). The blue lines in each panel represent the line of best fit. Flight segment averaging reduces random error—particularly for ERA-Interim ozone.

Table 3. Linear Regression Coefficients Used to Adjust Source Ozone Mixing Ratio[a]

	a	*b*
WACMM ozone	17.7 ppbv	1.09
ERA-Interim ozone	65.2 ppbv	1.78

[a]Values correspond to a and b in equation (3).

that reflect little more than the ability of the analysis data to distinguish the troposphere from the stratosphere and mask the difficulties of analysis data to reproduce local ozone values near the tropical tropopause. Figure 5 demonstrates that discrepancies between analysis data and observations dwarf those between observational data sets, indicating that observational uncertainty is not an important factor in our study.

Large discrepancies for point-by-point comparisons are no surprise; analysis data sets do not resolve small-scale variations that are captured by aircraft measurements, are based on imperfect physics and chemistry, and are subject to chaotic variations that cause solutions to the governing equations to diverge. However, averaging can substantially improve these comparisons. Figure 6 compares analysis and UCATS observations averaged over 75 min (~800 km) flight segments that often include vertical sampling over the altitude range of 14–19 km. Such averaging has a profound effect on random errors from ERA-Interim data, reducing RMS random errors by more than 50% and increasing explained variance r^2 from 0.47 to 0.74 (corresponding to a correlation coefficient r = 0.86). Averaging does not have a large effect on systematic errors and so only reduces the total RMS error by ~20%. Averaging has a smaller (~20%) impact on random errors for WACCM ozone; however, the correlation between WACCM and UCATS improves substantially with averaging (explained variance r^2 changes from 0.29 to 0.48). The latter is partially an artifact of the change in slope of the line of best fit (blues lines in Figures 4 and 5), a subject discussed in more detail in section 4.4. We also performed these comparisons for 15 min (~160 km) averages; while this smaller amount of averaging does reduce random errors (by ~12% for ERA-Interim, not shown), the reduction is not nearly as dramatic as for 800 km averages. To reduce the influence of large random errors on our analysis, the remainder of this study examines only 800 km averages.

3.2. Adjusted Ozone Sources

The same linear relationship that apparently characterizes systematic errors for analysis ozone data at both target and source locations motivates our implementation of a linear adjustment to source mixing ratios. This adjustment provides us with two additional source ozone data sets, "adjusted WACCM" and "adjusted ERA-Interim," and the sensitivity of predicted ozone mixing ratios to the implementation of this adjustment provides a meaningful characterization of the component of prediction uncertainty that is related to source ozone uncertainty. It is important to note that we are using prediction errors to test the validity of the linear adjustment; we *are not* using the adjustment as an a priori improvement to source mixing ratios. The adjusted ozone mixing ratios O_3^{adj} are calculated from the unadjusted values O_3^{src} using the systematic linear relationship between analysis ozone $O_3^{analysis}$ and observations O_3^{obs}:

$$O_3^{analysis} = a + bO_3^{obs} + \zeta.$$

If we assume that O_3^{adj} has the same systematic relationship with O_3^{src} that O_3^{obs} has with $O_3^{analysis}$, then

$$O_3^{adj} = \frac{O_3^{src} - a}{b}. \quad (3)$$

The values for a and b are determined by the linear regressions of WACCM ozone and ERA-Interim ozone onto UCATS observations (see Table 3). These values are based on regressions over all target locations (including those with UCATS ozone >300 ppbv) due to the robust linear relationships (i.e., large correlations) that exist when all values are included.

4. Model Evaluations

We begin the analysis of model predictions with a broad-brush examination of model errors for all 192 model configurations to identify which sets of model "parameters" (e.g., which combinations of source ozone data,

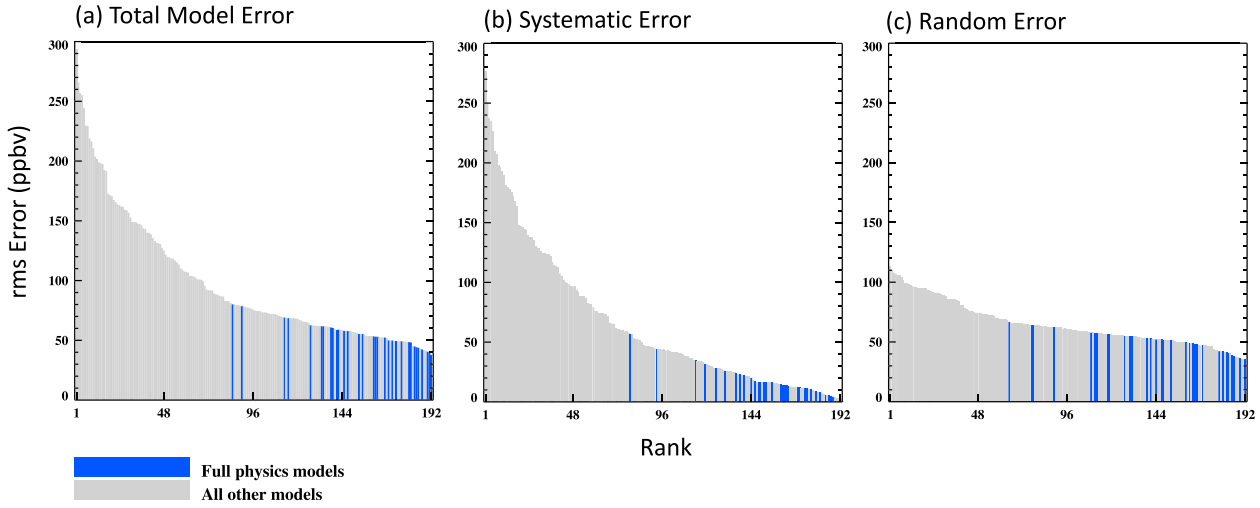

Figure 7. The bar graph representations of (a) total RMS error, (b) systematic error, and (c) random error for all 192 model configurations. The blue bars identify model configurations that use full physics, i.e., that use the ozone source adjustment, diabatic or kinematic trajectories, and WACCM-based chemical production. The models that use full physics have, in general, smaller model errors than those that do not.

trajectory formulation, and chemical production model) consistently lead to smaller prediction errors. We then use sensitivity tests to isolate a preferred region of parameter space, one for which model errors are not only small but also for which any change of an important parameter degrades the model prediction; that is, the preferred region corresponds to a model error minimum in parameter space. This method helps us remove grossly unrealistic models from further analysis and also helps to identify models that only have small prediction errors because of a fortuitous cancellation of errors. Compared to the full set of 192 models, the reduced parameter space provides a more realistic estimate of prediction error and a more realistic estimate of model sensitivity; the latter allowing us to probe subtle relationships that are lost in the larger sensitivities of the full set of models. We close section 4 with a comparison of two prediction evaluation criteria: RMS error and the correlation between predictions and observations.

The analysis in this section relies on bar graphs such as those found in Figure 7. In Figure 7a, for example, each bar (of which there are 192) quantifies the RMS error for a single model configuration. The bars are arranged (i.e., ranked) according to the magnitude of the error—from large values on the left to small on the right. This is true for all individual bar graphs shown, and as a result, the ordering of the model configurations differs in each bar graph. We typically color individual bars according to some pertinent detail. For example, the blue bars in Figure 7 represent that "full physics" models, i.e., those that utilize the source adjustment (detailed in section 3.2), are nonadiabatic trajectories (i.e., either diabatic or kinematic) and include the WACCM chemical production model; the gray bars represent all other model configurations.

4.1. A Broad Portrayal of Model Error

As a broad-brush examination of model error, Figure 7 compares total model error (e_{RMS}; Figure 7a), systematic error (e_{sys}; Figure 7b), and random error (e_{ran}; Figure 7c) for all 192 model configurations, highlighting errors from full physics models in blue. The full range of model parameters feature a wide range of total model errors (~45–290 ppbv) that is driven mostly by a wide range of systematic errors (~5–270 ppbv), while the range of random errors (~43–170 ppbv) is more compact.

While random errors are not as important as systematic errors for evaluating model performance, they are nevertheless important as a measure of the uncertainty for total RMS errors because they are uncorrelated with observed values. If we treat the total RMS error for each model configuration as the mean of a one-tailed distribution of errors, we can employ a standard significance test (i.e., a one-sided Z test) to estimate of the uncertainty δ_e (5% significance level)

$$\delta_e \sim \frac{1.645\sigma_\xi}{\sqrt{N}},$$

where $N = 116$ is the number of 800 km segments in our observational sample (i.e., that have mean observational values <300 ppbv). Thus, based on the range of random errors in Figure 7c, our estimate of the uncertainty of model prediction errors for the 192 model configurations varies approximately from 7 to 25 ppbv.

Our preliminary assessment is that models with full physics tend to have smaller errors than those without. However, this analysis is far from conclusive. For example, the four models with the smallest systematic errors and three of the four models with the smallest random errors do not use full physics. Furthermore, since there are three components that define full physics (source adjustment, vertical motion, and chemistry), Figure 7 does not reveal which of the components is most important or if a realistic model actually needs all three components. We cannot yet dismiss the possibility that one of the components of the full physics should not be part of the best model configuration.

4.2. Using Model Sensitivities to Constrain the Parameter Space

In addition to somewhat ambiguous results, the broad-brush approach contains an unnecessarily wide range of model parameters and a corresponding exaggerated range of model errors. Furthermore, sensitivity tests performed using unrealistic models can mask subtle sensitivities and lead to misleading results. Here we probe model sensitivities in an attempt to constrain the parameter space and exclude blatantly unrealistic models from the analysis. For these tests, we consider model error differences Δe_{RMS} for pairs of model configurations that differ in one component only.

Figure 8 examines the sensitivity of model error Δe_{RMS} to the source adjustment (error with the adjustment minus that without; Figure 8a), to vertical motion (errors from diabatic or kinematic calculations minus those from adiabatic calculations; Figure 8b), and sensitivity to chemistry (calculations using WACCM chemistry minus those ignoring chemical production; Figure 8c). The most salient feature of these comparisons is that the source adjustment overwhelmingly improves model performance; for all but two cases (i.e., 94 out of 96 error comparisons in Figure 8a), the models using the ozone source adjustment have smaller total RMS errors than those that do not. Furthermore, for the two exceptions, the discrepancy is smaller (less than 5 ppbv) than the smallest estimate of error uncertainty (~7 ppbv). Errors are also typically reduced if vertical motion is included (90% of the models in Figure 9b have smaller error for the kinematic or diabatic run compared to the adiabatic). However, these sensitivities also show that for most of the configurations, model error is smaller if chemistry is ignored. Does this mean that our chemical production model is so bad that models are better without it? Not necessarily.

We now take the first step toward constraining the viable parameter space. Since nearly all models are improved by the use of the source adjustment, Figures 8d and 8e revisit the sensitivities in Figures 8b and 8c using only source-adjusted models. For these figures, we have also identified models that use all three components of full physics using blue bars. Figure 8d demonstrates that vertical motion improves all source-adjusted models that use WACCM chemistry (all blue bars are negative). Figure 8e demonstrates that WACCM chemistry improves most models using vertical motion (most blue bars are negative) but degrades all adiabatic models (all gray bars are positive), despite degrading more models overall than it improves. To summarize, Figure 8a demonstrates that the source adjustment is important for reliable ozone prediction; Figure 8d implies that if chemistry is an important component then vertical motion must be as well, and Figure 8e implies that if vertical motion is important then chemistry probably is also important. Since many of these sensitivities are well below the threshold of detection at the 5% significance level, we cannot determine if the exceptions are due to sampling or due to problems with the wind data and chemistry model. Nevertheless, there are so few exceptions to the determination of the preferred parameter space (i.e., none for source adjustment sensitivity, none for vertical motion sensitivity, and only two cases for which the chemistry model degrades the prediction by more than 5 ppbv) that it seems valid.

4.3. Diagnosing Sensitivities Within the Reduced Parameter Space

Having identified a preferred region of model parameter space, we now reassess our diagnosis of model performance and perform additional sensitivity comparisons that more faithfully portray the importance of each of the model components. Figure 9 displays model errors within the reduced space as functions of trajectory length. Model errors in the reduced parameter space vary only from 40 to 80 ppbv, compared to 40 to 270 ppbv for all models (Figure 7a). These are sizable errors (roughly 40–80% of the mean observed mixing ratios) that restrict the range of science questions that can be reliably addressed with back

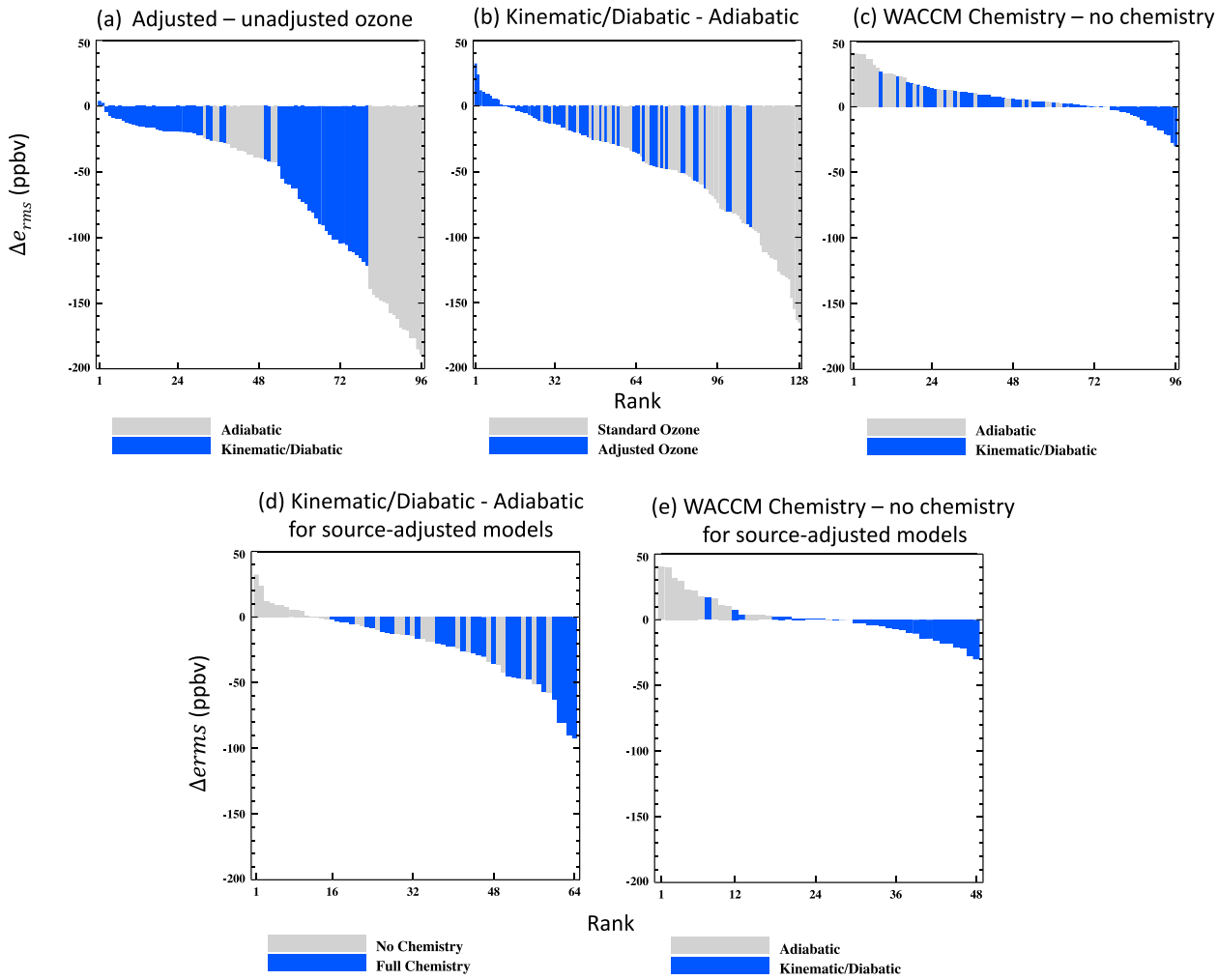

Figure 8. The bar graph representations of total RMS error sensitivities: (a) sensitivities to the ozone source adjustment (errors using the adjustment minus without the adjustment), (b) to the use of vertical motion in the trajectory calculations (errors using diabatic or kinematic trajectories minus those using adiabatic trajectories), (c) to the use of WACCM-based chemical production (errors with WACCM chemistry minus those without), (d) of source-adjusted models to the use of vertical motion in the trajectory calculations (errors using diabatic or kinematic trajectories minus those using adiabatic trajectories), and (e) of source-adjusted models to the use of WACCM-based chemical production (errors with WACCM chemistry minus those without). The blue bars identify models using: vertical motion (in Figures 8a, 8c, and 8e), adjusted ozone sources (in Figure 8b), and WACCM-based chemical production (in Figure 8d). These figures provide a consistency argument that identifies a preferred region of parameter space; that is, the source adjustment is important for reliable ozone prediction, if chemistry is an important component then vertical motion must be as well, and if vertical motion is important then chemistry probably is also important.

trajectories. However, not all science questions are removed from consideration, and more importantly, we have a quantitative measure for determining which questions are well posed. An interesting feature of Figure 9 is that model errors are not strongly tied to the length of the trajectory. It is true that the shorter trajectories (5 and 10 days) have smaller errors than the longer trajectories (20 and 30 days), and in particular, the models with the largest errors (worst 5) are all long trajectories; however, we also find long trajectories among predictions with the smallest model errors (e.g., three out of the top six).

Figure 10 revisits the sensitivities to source adjustment, vertical motion, and chemistry for models within the reduced parameter space. Figure 10a shows the error difference between predictions by the models with full physics and models that differ only in the lack of source adjustment, Figure 10b displays Δe_{RMS} for full physics compared to adiabatic calculations, and Figure 10c shows Δe_{RMS} for full physics compared to models with no chemistry. All panels display these sensitivities in the context of which data set is being used for source concentrations; i.e., the gray bars correspond to source ozone from ERA-Interim, and the blue bars correspond to WACCM data. As expected from the results in Figure 8, all of the models from the reduced space

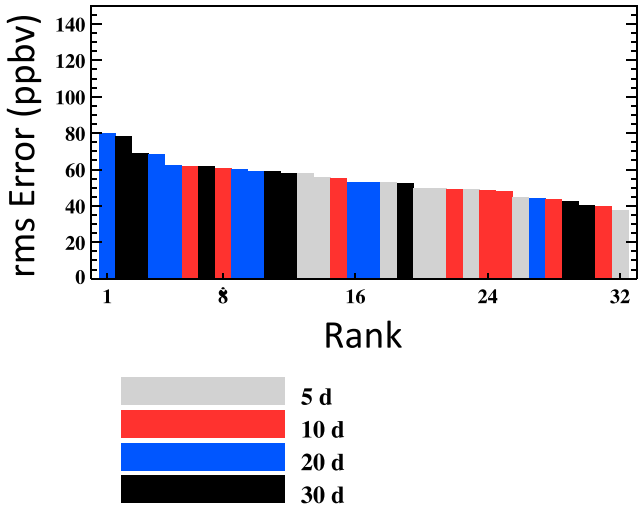

Figure 9. The bar graph representation of total RMS error for models with full physics. The bars are colored according to the length of the trajectories: the gray color represents models using 5 day trajectories, the red color represents 10 day trajectories, the blue color represents 20 day trajectories, and the black color represents 30 day trajectories. Limiting models to this reduced parameter space reduces the range of model errors from ~40–270 ppbv to ~40–80 ppbv.

are degraded by removing the source adjustment and there is a clear distinction based on which ozone data set is used; the source adjustment has a relatively small impact (<30 ppbv) for WACCM data compared to ERA-Interim data (40–120 ppbv). Similarly, all models from the reduced parameter space are degraded by a switch to adiabatic trajectories. The impact of vertical motion is also a function of the source ozone, with models using WACCM ozone being more sensitive to vertical motion than those using ERA-Interim. In fact, models using WACCM ozone are actually more sensitive to vertical motion than to the source adjustment. WACCM ozone that undergoes a smaller source adjustment than ERA-Interim (due to the smaller systematic errors in the unadjusted WACCM ozone) is no doubt an important contributor to these features. The sensitivity to the use of chemistry is expectedly smaller and more ambiguous than the source adjustment or vertical motion sensitivities. However, all models in the reduced parameter space that use adjusted ERA-Interim ozone data for source concentrations are degraded by removing chemistry. That is, in terms of these sensitivities, models using ERA-Interim ozone are perfectly consistent (i.e., there are no exceptions) with our determination of the preferred model parameter space.

An important consideration when performing trajectory calculations is deciding which forcing data and which trajectory formulation to use. Figure 11 explores this issue within the confines of the reduced parameter space comparing prediction errors using different wind-field data (Figure 11a), different trajectory formulations (Figure 11b), and different source ozone data (Figure 11c). These differences are small (typically

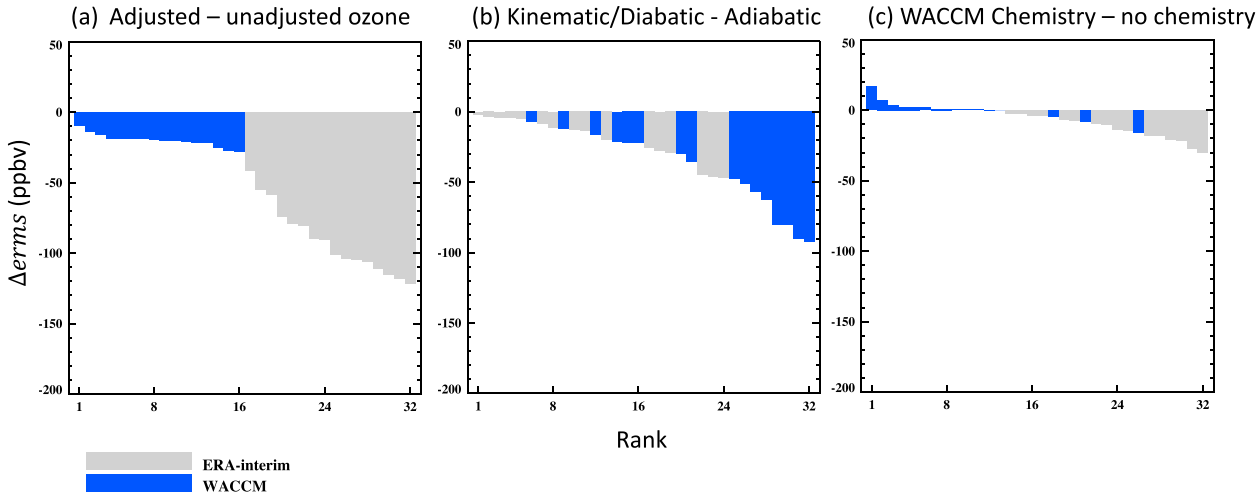

Figure 10. The bar graph representations of total RMS error sensitivities for full physics models: (a) sensitivities to the ozone source adjustment (error using the adjustment minus without the adjustment), (b) to the use of vertical motion in the trajectory calculations (errors using diabatic or kinematic trajectories minus those using adiabatic trajectories), and (c) to the use of WACCM-based chemical production model (errors with WACCM chemistry minus those without). The blue bars identify models that use WACCM source ozone. All full physics models are degraded by removing the source adjustment, all full physics models are degraded by removing vertical motion, and all full physics models using ERA-Interim source ozone are degraded by removing the WACCM-based chemical production model.

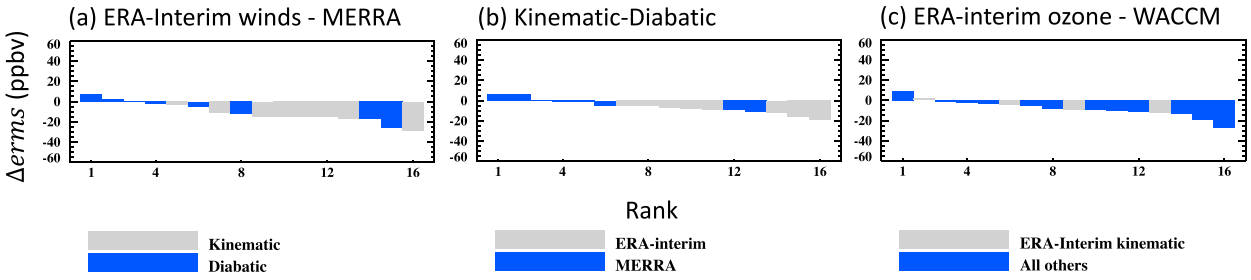

Figure 11. The bar graph representations of total RMS error sensitivities for full physics models: (a) the sensitivity to the choice of trajectory forcing data (ERA-Interim minus MERRA), (b) to the choice of trajectory formulation (kinematic minus diabatic), and (c) to the choice of source ozone data (ERA-Interim minus WACCM. The gray bars identify the kinematic calculations in Figure 11a, the use of ERA-Interim winds in Figure 11b, and the kinematic calculations that use ERA-Interim winds in Figure 11b. A consistency argument identifies the kinematic calculation using ERA-Interim winds, adjusted ERA-Interim source ozone, and WACCM-based chemical production as the best model configuration. A note of caution: this result is based, in part, on error sensitivities that do not pass a significance test.

<20 ppbv), are *not* likely to be statistically significant, and are somewhat ambiguous in their interpretation. Note that *most*, but not all, models from our reduced parameter space have smaller prediction errors when using ERA-Interim winds compared to MERRA winds, and *most*, but not all, have smaller errors for kinematic trajectories than for diabatic trajectories, and *most*, but not all, models using ERA-Interim ozone have smaller errors than those using WACCM ozone. Nevertheless, there is a set of parameters that consistently survives scrutiny; if we further constrain our parameter space we find that *all* kinematic calculations have smaller errors using ERA-Interim winds than for MERRA winds (gray bars in Figure 11a), and *all* models using ERA-Interim winds have smaller errors for kinematic calculations than for diabatic calculations (gray bars in Figure 11b), and all but one of the kinematic calculations using ERA-Interim winds have smaller errors using adjusted ERA-Interim ozone than those using adjusted WACCM ozone. The lone exception is a 30 day trajectory with a meager Δe_{RMS} (WACCM-minus-ERA) of 1.9 ppbv. Trajectories using ERA-Interim winds that provide better predictions than those using MERRA are consistent with our previous findings [*Bergman et al.*, 2013, 2015] that indicate problems with trajectories using MERRA winds and those of *Wright and Fueglistaler* [2013] that find problems with diabatic heating rates from MERRA. However, our finding that the kinematic calculations provide better ozone predictions than diabatic models conflicts with other studies [e.g., *Schoeberl et al.*, 2003; *Ploeger et al.*, 2011, 2012]. Experimental uncertainty aside, these conflicts are not necessarily contradictions; instead, they demonstrate that the choice of trajectory formulation should not be an a priori decision but rather the result of sensitivity experiments.

4.4. Misled by Correlations—Again

So far, we have used an amplitude measure, RMS error, to assess model performance. Correlation is also an important measure because it relates how well variations of the predictand are tracked by the predictor; i.e., it measures the phase information of the predictor. Typically, RMS error and correlation complement each other and together provide a useful characterization of model error, although they are not entirely independent (for example, a prediction with no RMS error must have a correlation coefficient $r = 1.0$) and it can be tempting to assume that the model with a highest correlation with observations is the best model. However, for our model comparisons, higher correlation is not associated with smaller RMS errors. This is demonstrated in Figure 12a, which shows the source adjustment sensitivity in terms of an explained variance difference Δr^2 (values from models using the source adjustment minus those from models that do not). Despite the large improvement that the source adjustment has on RMS error (exceeding 100 ppbv in some cases; Figure 8a), no model experiences a change of explained variance larger than 0.05 and those changes are typically negative. That is, source adjustments nearly always reduce RMS error (indicating model improvement), yet typically also reduce explained variance (indicating model degradation).

These contradictory assessments are due to a systematic relationship between the explained variance and the slope of the regression line (see equation (2)). As shown in Appendix A, predictions that experience a systematic reduction of the slope also experience a reduction of explained variance provided that the random error does not change much. Since our sensitivity tests make strong systematic changes to the prediction, this artifact of slope reduction dominates the change of explained variance. If we remove this artifact by calculating the slope-corrected explained variance (i.e., by calculating the explained variance

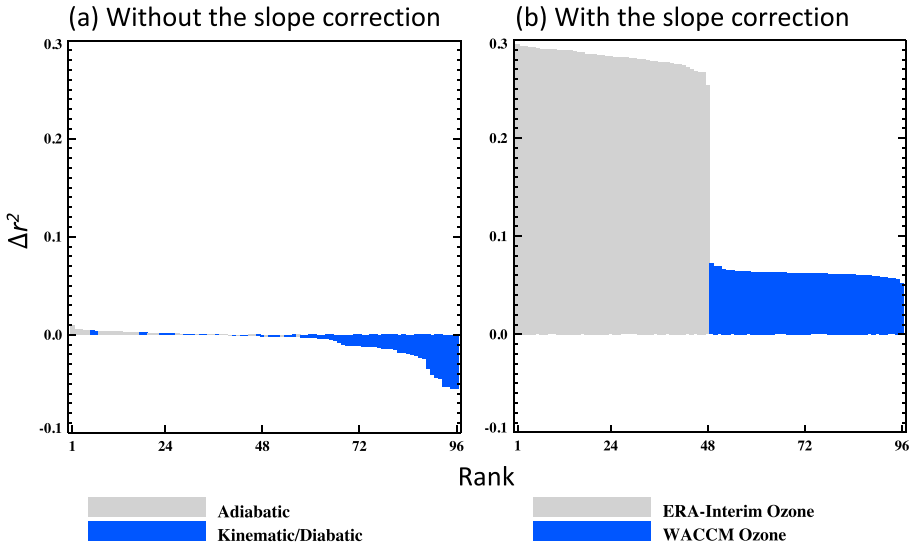

Figure 12. The bar graph representations of explained variance r^2 sensitivities to the use of the source adjustment for (a) a standard calculation of r^2 and (b) the slope-corrected r^2. Without the slope correction, explained variance indicates a degradation of model predictions with the use of the source adjustment, whereas total RMS error indicates improved predictions. The slope-corrected explained variance indicates that the source adjustment improves ozone predictions for all models.

after removing systematic error; see Appendix A) then explained variance is enhanced for all models by the source adjustment (Figure 12b). The difference between the conventional measure of explained variance and the slope-corrected measure is substantial and serves as a stark reminder of the pitfalls of relying solely on correlation as a measure of model performance.

5. Summary and Conclusions

The viability and limits of applicability of Lagrangian analysis for diagnosing the interplay between chemistry and dynamics were investigated by comparing model predictions of ozone mixing ratios to observations along flight tracks in the tropical UTLS during ATTREX. Trajectories were initiated at target locations along ATTREX flight tracks and traced backward in time to source locations. Chemical production models were initialized at the source locations with ozone mixing ratios obtained from analysis data and integrated forward in time along parcel pathways to the target locations where model predictions of ozone mixing ratios were compared to UCATS observations. The analysis of model error focused on three model components: source ozone concentrations, trajectory formulation, and chemical production along the parcel path. To diagnose model sensitivities, ozone predictions were calculated from 192 different model configurations, utilizing two sets of wind fields (ERA-Interim and MERRA) to calculate air parcel trajectories, three trajectory formulations (adiabatic, diabatic, and kinematic), four sets of source ozone mixing ratios (standard and adjusted; WACCM and ERA-Interim), two chemical production models (WACCM chemistry and no chemistry), and four trajectory lengths (5, 10, 20, and 30 days). For each of these model configurations, 59,400 values were compared to observations over 165 h of flight tracks.

The direct comparison of analysis ozone data to observations reveals large random errors for "instantaneous" values (10 s averages) that are substantially reduced by averaging data over 75 min (~800 km) flight segments. Our Lagrangian predictions of ozone mixing ratios are likewise improved by averaging (predictions of instantaneous values were not shown in the manuscript), reinforcing our expectation that such predictions are only valid for averages over sufficiently large ensembles. Averaging also exposes pronounced systematic errors that motivate a linear adjustment to the source mixing ratios obtained from analysis data.

The use of extensive sensitivity experiments and data averaging is critical aspects of this work that tame large random and systematic uncertainties that would otherwise mask important relationships. Nevertheless, there are caveats that should be considering when interpreting of our results. We cannot rule out the importance of sources of error that we have not considered. This point is made clear by Figure 8b, which shows that models

not using the source adjustment are more sensitive to the addition of vertical motion than source-adjusted models; in this case, the importance of vertical motion is exaggerated if we do not first adjust the source ozone mixing ratios. Similarly, the importance of the source adjustment is exaggerated if we examine predictions from adiabatic trajectories, and the importance of chemical production only emerges if we limit our scope to source-adjusted models with vertical motion (compare Figures 8c and 8e). That is, if we are not examining the correct parameter space, our results could be misleading. Since our trajectory calculations rely on resolved wind fields from reanalysis data, they neglect or underrepresent potentially important physical processes such as tropical convection and turbulent mixing, our isolation of a preferred parameter space could be inappropriate and lead to misleading results. Second, we note that model performance is a function of the error metric used. In that context, we found that a naïve application of correlation analysis leads to a misleading diagnosis of model performance due to the functional relationship between correlation and the slope of the line of best fit.

The sensitivity experiments perform two primary functions: they reveal which of the model components most affects ozone predictions and they help us identify full physics models (i.e., those that use the source adjustment, vertical motion, and WACCM chemistry) as consistently representing the most realistic models. In this capacity, model sensitivities provide partial derivatives in parameter space that determine a local minimum for prediction errors. This method is more robust than using the prediction error alone to identify the most realistic model configuration; it relies on the combined results of many sensitivity tests, which reduces the impact of random errors, and it relies on tests that alter one parameter at a time, which reduces the misleading influence of systematic error cancellation that occurs for some combinations of unrealistic model components.

From the sensitivity experiments we conclude the following:

1. The ozone source adjustment leads to a major improvement in the model simulation of ozone for either ozone source data set. This fact indicates that the specification of source ozone mixing ratios is the largest source of uncertainty for these Lagrangian prediction models. It also validates the use of the linear source adjustment, which in turn validates the contention that ATTREX data provide a reasonable sample of ozone mixing ratios in the tropical upper troposphere.
2. Diabatic or kinematic trajectories generally give better results than isentropic trajectories. This result is reassuring because it represents partial validation for vertical velocities and diabatic heating rates from reanalysis data sets, for which direct validation with observational data is not viable.
3. Compared to the use of the source adjustment and vertical motion, predictions are not very sensitive to the use of active ozone chemistry along the parcel path.
4. Full physics models (source-adjusted models using vertical motion and active chemistry) are equally sensitive (±10 ppbv; see Figure 11) to the choice of wind data (MERRA versus ERA-Interim), of trajectory formulation (diabatic versus kinematic), or of ozone source data (WACCM versus ERA-Interim).
5. ERA-Interim winds give slightly better results than MERRA winds in most cases using full physics.
6. ERA-Interim ozone gives slightly better results than WACCM ozone in most cases using full physics.

The "most realistic" models, as identified by the preferred parameter space, have prediction errors on the order of 50%, which is large enough to limit their utility for many scientific applications. However, this does not remove all science questions from consideration; more important, it provides a quantitative measure for distinguishing which science questions can be addressed with Lagrangian calculations. As a cautionary note, model sensitivities such as those investigated here are likely to be context dependent and one would be ill advised to use our results alone to choose a model configuration for an unrelated scientific investigation.

Appendix A: Slope-Corrected Correlation

For a linear prediction model of the form

$$y_i = a + bx_i + \zeta_i,$$

the explained variance

$$r^2 = \frac{b^2\sigma_x^2}{b^2\sigma_x^2 + \sigma_\zeta^2}$$

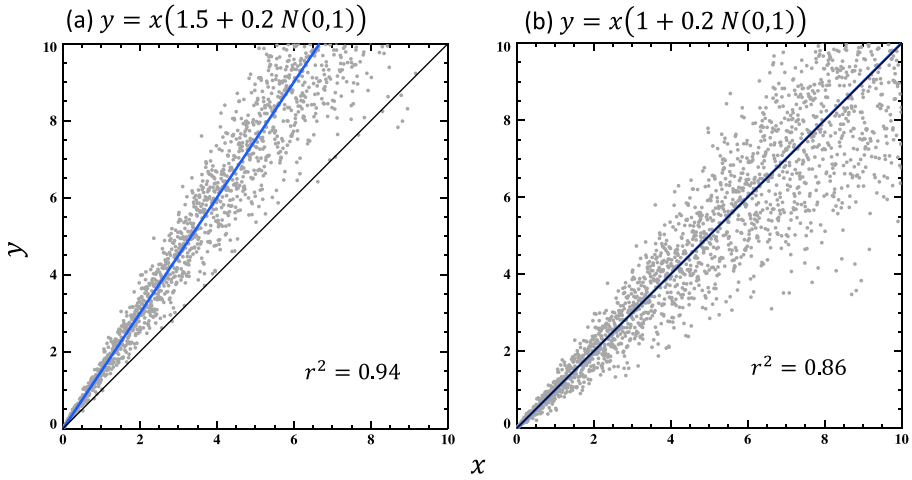

Figure A1. (a and b) Scatterplots of predictions y versus actual values *x* that demonstrate the reduction of explained variance that can accompany a reduction of systematic error despite identical random errors. The only difference between the two prediction models is that Figure A1a contains a systematic error with regression slope $b = 1.5$ and Figure A1b has no systematic error. The variance explained by the model with a systematic error is larger ($r^2 = 0.94$) than that for the model with no systematic error ($r^2 = 0.86$).

depends on slope *b* of the line of best fit. If that slope is greater than 1, a reduction of the systematic error that leaves the random error unchanged reduces the explained variance of the prediction. This is demonstrated in Figure A1, which shows scatterplots (prediction *y* versus actual *x*) for two prediction models that differ only in the value of the regression slope *b*. In this case, despite having identical random errors, the model with no systematic error ($b = 1.0$; Figure A1b) explains less variance ($r^2 = 0.86$) than the model that has systematic errors ($r^2 = 0.94$; Figure A1a). We can remove this artifact if we construct a slope-corrected model

$$y_i' = y_i - a - (b-1)x_i = x_i + \zeta_i,$$

for which correlations and explained variance

$$r'^2 = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_\zeta^2}$$

no longer depend explicitly on the regression slope.

Acknowledgments
The tenor of this work can be directly traced to discussions with P. Sardeshmukh to whom we are particularly grateful. This work also benefitted from discussions with and comments by A. Conley, E. Jensen, S. Madronich, W. Randel, J.F. Lamarque, and two anonymous reviewers. J. Bergman was a visitor at the Atmospheric Chemistry Observations and Modeling Laboratory of NCAR during the execution of this study. The National Center for Atmospheric Research is operated by the University Corporation for Atmospheric Research, under sponsorship of the National Science Foundation. Supporting information for *Kinnison et al.* [2007] can be found online at http://onlinelibrary.wiley.com/doi/10.1029/2006JD007879/abstract. UCATS and NOAA-CSD ozone data can be found at NASA's ESPO data archive: https://espoarchive.nasa.gov/archive/browse/attrex. ERA-Interim data were obtained from the CISL Data Support Section at the National Center for Atmospheric Research (NCAR). MERRA data were obtained from the Goddard Earth Sciences Data and Information Services Center. For other data and models used in this study contact J. Bergman (j.w.bergman921@gmail.com).

References

Ashfold, M. J., N. R. P. Harris, E. L. Atlas, A. J. Manning, and J. A. Pyle (2012), Transport of short-lived species into the tropical tropopause layer, *Atmos. Chem. Phys.*, *12*, 6309–6322, doi:10.5194/acp-12-6309-2012.

Avallone, L. M., and M. J. Prather (1996), Photochemical evolution of ozone in the lower tropical stratosphere, *J. Geophys. Res.*, *101*, 1457–1461, doi:10.1029/95JD03010.

Bergman, J. W., E. J. Jensen, L. Pfister, and Q. Yang (2012), Seasonal differences of vertical transport efficiency in the tropical tropopause layer: On the interplay between tropical deep convection, large scale vertical ascent, and horizontal circulations, *J. Geophys. Res.*, *117*, D05302, doi:10.1029/2011JD016992.

Bergman, J. W., F. Fierli, E. J. Jensen, S. Honomichl, and L. L. Pan (2013), Boundary layer sources for the Asian anticyclone: Regional contributions to a vertical conduit, *J. Geophys. Res. Atmos.*, *118*, 2560–2575, doi:10.1002/jgrd.50142.

Bergman, J. W., L. Pfister, and Q. Yang (2015), Identifying robust transport features of the upper tropical troposphere, *J. Geophys. Res. Atmos.*, *120*, 6758–6776, doi:10.1002/2015JD023523.

Bergman, J. W., E. J. Jensen, L. Pfister, and T. P. Bui (2016), Air parcel trajectory dispersion near the tropical tropopause, *J. Geophys. Res. Atmos.*, *121*, 3759–3775, doi:10.1002/2015JD024320.

Cariolle, D., and H. Teyssèdre (2007), A revised linear ozone photochemistry parameterization for use in transport and general circulation models: Multi-annual simulations, *Atmos. Chem. Phys.*, *7*(9), 2183–2196, doi:10.5194/acp-7-2183-2007.

Dee, D. P., et al. (2011), The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, *Q. J. R. Meteorol. Soc.*, *137*, 553–597.

Folkins, I., R. Chatfield, D. Baumgardner, and M. Proffit (1997), Biomass burning and deep convection in southeastern Asia: Results from ASHOE/MAESA, *J. Geophys. Res.*, *102*(D11), 13,291–13,299, doi:10.1029/96JD03711.

Fueglistaler, S., A. E. Dessler, T. J. Dunkerton, I. Folkins, Q. Fu, and P. W. Mote (2009), Tropical tropopause layer, *Rev. Geophys.*, *47*, RG1004, doi:10.1029/2008RG000267.

Fujiwara, M., et al. (2017), Introduction to the SPARC Reanalysis Intercomparison Project (S-RIP) and overview of the reanalysis systems, *Atmos. Chem. Phys.*, *17*, 1417–1452, doi:10.5194/acp-17-1417-2017.

Gao, R. S., J. Ballard, L. A. Watts, T. D. Thornberry, S. J. Ciciora, R. J. McLaughlin, and D. W. Fahey (2012), A compact, fast UV photometer for measurement of ozone from research aircraft, *Atmos. Meas. Tech.*, *5*, 2201–2210, doi:10.5194/amt-5-2201-2012.

Jensen, E. J., et al. (2017), The NASA Airborne Tropical TRopopause EXperiment (ATTREX): High-altitude aircraft measurements in the tropical western Pacific, *Bull. Am. Meteorol. Soc.*, *98*, 129–143, doi:10.1175/BAMS-D-14-00263.1.

Kinnison, D. E., et al. (2007), Sensitivity of chemical tracers to meteorological parameters in the MOZART-3 chemical transport model, *J. Geophys. Res.*, *112*, D20302, doi:10.1029/2006JD007879.

Lamarque, J. F., et al. (2012), CAM-Chem: Description and evaluation of interactive atmospheric chemistry in the Community Earth System Model, *Geosci. Model Dev.*, *5*, 369–411, doi:10.5194/gmd-5-369-2012.

Law, K. S., et al. (2010), Air mass origins influencing TTL chemical composition over West Africa during 2006 summer monsoon, *Atmos. Chem. Phys.*, *10*(22), 10,753–10,770, doi:10.5194/acp-10-10753-2010.

Marsh, D., M. J. Mills, D. E. Kinnison, J.-F. Lamarque, N. Calvo, and L. M. Polvani (2013), Climate change from 1850 to 2005 simulated in CESM1 (WACCM), *J. Clim.*, *26*, 7372–7391, doi:10.1175/JCLI-D-12-00558.1.

Pfister, L., et al. (2001), Aircraft observations of thin cirrus clouds near the tropical tropopause, *J. Geophys. Res.*, *106*, 9765–9786, doi:10.1029/2000JD900648.

Pisso, I., E. Real, K. S. Law, B. Legras, N. Bousserez, J. L. Attié, and H. Schlager (2009), Estimation of mixing in the troposphere from Lagrangian trace gas reconstructions during long-range pollution plume transport, *J. Geophys. Res.*, *114*, D19301, doi:10.1029/2008JD011289.

Ploeger, F., et al. (2011), Insight from ozone and water vapour on transport in the tropical tropopause layer (TTL), *Atmos. Chem. Phys.*, *11*, 407–419, doi:10.5194/acp-11-407-2011.

Ploeger, F., P. Konopka, R. Müller, G. Günther, J.-U. Grooß, C. Schiller, F. Ravegnani, A. Ulanovski, and M. Riese (2012), Back-trajectory reconstruction of water vapour and ozone in-situ observations in the TTL, *Meteorol. Z.*, *21*, 239–244.

Randel, W. J., and E. J. Jensen (2013), Physical processes in the tropical tropopause layer and their roles in a changing climate, *Nat. Geosci.*, *6*, 169–176, doi:10.1038/ngeo1733.

Rienecker, M. M., et al. (2008), The GEOS-5 Data Assimilation Dystem—Documentation of versions 5.0.1 and 5.1.0, and 5.2.0,. *NASA Tech. Rep. Series on Global Modeling and Data Assimilation, NASA/TM-2008-104606*, vol. 27, 92 pp., Goddard Space Flight Cent., Greenbelt, Md.

Rienecker, M. M., et al. (2011), MERRA-NASA's Modern-Era Retrospective Analysis for Research and Applications, *J. Clim.*, *24*, 3624–3648, doi:10.1175/JCLI-D-11-00015.1.

Schoeberl, M. R., and L. C. Sparling (1995), Trajectory modeling, diagnostic tools in atmospheric science, in *Proc. Int. School of Phys. Enrico Fermi*, vol. 124, pp. 289–305, IOS Press, Amsterdam.

Schoeberl, M. R., A. R. Douglass, Z. Zhu, and S. Pawson (2003), A comparison of the lower stratosphere age spectra derived from a general circulation model and two data assimilation systems, *J. Geophys. Res.*, *108*(D3), 4113, doi:10.1029/2002JD002652.

Ueyama, R., E. J. Jensen, L. Pfister, G. S. Diskin, T. P. Bui, and J. M. Dean-Day (2014), Dehydration in the tropical tropopause layer: A case study for model evaluation using aircraft observations, *J. Geophys. Res. Atmos.*, *119*, 5299–5316, doi:10.1002/2013JD021381.

Wang, T., R. W. Randel, A. E. Dessler, M. R. Schoeberl, and D. E. Kinnison (2014), Trajectory model simulations of ozone (O_3) and carbon monoxide (CO) in the lower stratosphere, *Atmos. Chem. Phys.*, *14*, 7135–7147, doi:10.5194/acp-14-7135-2014.

Wright, J. S., and S. Fueglistaler (2013), Large differences in reanalyses of diabatic heating in the tropical upper troposphere and lower stratosphere, *Atmos. Chem. Phys.*, *13*, 9565–9576, doi:10.5194/acp-13-9565-2013.