Analyzing dynamical circulations in the tropical tropopause layer through empirical predictions of cirrus cloud distributions

John W. Bergman1,2, Eric J. Jensen3, Leonhard Pfister3, Laura L. Pan2, and Shawn Honomichl2

1Bay Area Environmental Research Institute, Sonoma, California, USA, 2Atmospheric Chemistry Division, National Center for Atmospheric Research, Boulder, Colorado, USA, 3Earth Science Division, NASA Ames Research Center, Moffett Field, California, USA

Abstract We explore the use of nonlinear empirical predictions of thin cirrus for diagnosing transport through the tropical tropopause layer (TTL). Thirty day back trajectories are calculated from the locations of CALIPSO cloud observations to obtain Lagrangian dry and cold points associated with each observation. These historical values are combined with “local” (at the location of the CALIPSO observation) temperature and specific humidity to predict cloud probability using multivariate polynomial regression. We demonstrate that our statistical sample (seven seasons) is sufficient to retrieve the full nonlinear relationship between cloud probability and its predictors and that substantial information is lost in a purely linear analysis. The best cloud prediction is obtained by the two-variable combination of local temperature and humidity, which reflects the close relationship between clouds and relative humidity. However, single-variable predictions involving parcel histories are better than those based solely on the individual local fields, indicating the existence of reliable dynamical information content within parcel trajectories. Thermal fields are better cirrus predictors during boreal winter than summer primarily due to poor predictions over the Asian summer monsoon region, revealing that the functional relationship over southern Asia differs from the rest of the tropics; in short, TTL cirrus formation over regions of active maritime convection, such as the West Pacific, is thermally dominated, indicating an environment in which in situ cirrus are readily formed, while TTL cirrus of southern Asia is moisture dominated, indicating a more direct connection between convective injection of moisture and thin cirrus.

1. Introduction

The tropical tropopause layer is a transition zone that merges the upper tropical troposphere with the lower stratosphere. The tropical tropopause layer (TTL) is important primarily as a gateway through which tropospheric air enters the stratosphere and, as such, exerts a strong influence on constituent concentrations in the stratosphere [Fueglistaler et al., 2009, and references therein]. Understanding transport through the TTL has its challenges. Foremost among these is lack of in situ measurements. A large portion of the TTL is over ocean and the landed regions within the tropics are typically within countries that do not maintain extensive meteorological observational networks. In addition, tropical convection is difficult to simulate in general circulation models compromising the integrity of “analyzed data sets” (i.e., operational analysis and reanalysis) that depend on those models.

While the TTL suffers from the lack of in situ measurements, it is situated above most thick clouds and, as such, is well observed by remote sensors aboard satellites. In this context, cloud distributions are key to understanding dynamical circulations in the TTL because of two important properties: they interact strongly with radiative fluxes, and they form when the relative humidity exceeds roughly 1.0. In addition to being readily observed from satellite, their impact on radiative fluxes makes clouds important for diabatic heating rates, which in the TTL are largely in balance with adiabatic cooling associated with vertical motion [e.g., Randel et al., 2008] tying clouds directly to large-scale vertical transport. For example, idealized radiative transfer calculations estimate that the “level of zero radiative heating” (the minimum altitude to which air parcels must be lofted by convection in order to reach the stratosphere) is elevated by 1 km if the radiative effect of clouds is neglected [Gettelman et al., 2004; Corti et al., 2005]. The relationship between cloud formation and humidity together with the nearly exponential dependence of saturation vapor pressure on
temperature has profound implications for the TTL as it does for most of the lower atmosphere. For example, it helps set upper limits on the humidity of air passing through the tropical tropopause where atmospheric temperatures minimize [Brewer, 1949; Newell and Gould-Stewart, 1981; Jensen et al., 1996], and it links cloud formation to circulation features in which moist air cools. Since the uplift of air is an important factor if not dominant mechanism for cooling in the troposphere, clouds help us identify important dynamical events such as convection, wave activity, and slow ascent.

The prevailing paradigm for boundary-layer-to-stratosphere transport is that air is first lifted into the TTL by convection and then transported slowly by the large-scale uplift that is in balance with radiative heating into the stratosphere over a period of 30–60 day [Fueglistaler et al., 2009; Ploeger et al., 2010]. Once air parcels are detrained from convection in the TTL, their circulation patterns seem to be largely unaffected by the small-scale interactions typical of turbulent flows that exist in lower regions of the atmosphere (cf. discussion in Fueglistaler et al. [2004]). This simplification makes Lagrangian particle models that track the individual air parcel trajectories important tools for understanding transport through the TTL and chemical distributions in the stratosphere [e.g., Jackson et al., 2001; Legras et al., 2003; Bonazzola and Haynes, 2004; Jensen and Pfister, 2004; Fueglistaler et al., 2004; James et al., 2008; Liu et al., 2010; Ploeger et al., 2011; Schoeberl and Dessler, 2011]. While precise tracking of air parcels is not feasible for times typical of transport through the TTL, Lagrangian calculations can, in principle, supply probability distributions from which useful statistical measures of transport can be derived. For example, an ensemble of properly formulated back trajectories, which trace parcel paths backward in time from a target location \( x_0 \) and time \( t_0 \) to a source \( (x_{src},t_{src}) \), can yield a statistical description of the conditions that determine the atmospheric state at \( (x_0,t_0) \).

While their potential as diagnostic tools is high, trajectory calculations face obstacles that can undermine their utility. To determine air parcel trajectories requires global “forcing fields” at high temporal and spatial resolution. For kinematic trajectories that are formulated in pressure coordinates, forcing fields are the horizontal winds \( u \) and \( v \) and pressure velocity \( \omega_p \); for diabatic trajectories that are formulated in isentropic, or \( \theta \), coordinates, forcing fields are \( u, v \), and diabatic heating rates \( q \). Direct observations of the horizontal winds are too sparse to determine parcel trajectories even in the most observation-rich regions of the atmosphere, while vertical motion is, for practical purposes, too weak to observe. As a result, trajectories are typically calculated using analyzed data sets for forcing fields. Analyzed fields are valuable, state-of-the-art products that incorporate the top observational and modeling technologies. Nevertheless, these data suffer from uncertainties that are difficult to assess, particularly in the TTL, where the coverage of in situ data is thin and satellite observations of temperature offer poor constraints on the wind fields due to the small influence of the Coriolis force.

In light of the data uncertainties, this study does not assume that any source of data, whether observational or model derived, is accurate. Instead, we diagnose circulations in the TTL by examining the consistency between data sources. That is, we quantify the mutual information contained in two fields (e.g., the probability of cloud occurrence and temperature) to diagnose circulations. The underlying assumption is that two sets of data from “independent” sources (i.e., neither data set is constructed using information that was also used to construct the other) will only be consistent if there is a real physical connection between them. As will become apparent, our analysis allows us to extract dynamical information from data that has potentially large uncertainties associated with it and allows us to compare the relative reliability of different data sources for the same physical quantity.

For this study, we analyze the TTL circulation in terms of distributions of thin cirrus clouds in a vertical layer near the tropical tropopause. That is, we quantify dynamical consistency via the comparison of thin clouds observed from CALIPSO (described in section 2) to empirical predictions of thin clouds. Our analysis method (schematically outlined in Figure 1) follows the standard practice of deriving (training) a prediction model using observed values of predictors and predictands from one time span (e.g., winter 2008–2013) and validating the prediction model using observations from an independent time span (e.g., winter 2007). Note, our use of the term “prediction” corresponds to the calculation of one quantity from its statistical relationship to others; it is not be confused with predictions of future values. The method proceeds as follows: We use multivariate, nonlinear regression to determine the functional relationship between the probability of CALIPSO observing TTL cirrus and a set of predictors (obtained from thermal and moisture fields in the TTL). We then apply the regression model to an independent set of predictors to calculate the corresponding predicted cloud probabilities and validate those predictions against CALIPSO observations.
The power of this analysis method is derived from the following properties:

1. All analyses are based on comparisons of predicted to observed fields. This allows us to use linear comparison measures such as the correlation coefficient $r$ and the explained variance $r^2$ without ambiguity. That is, we know a priori that the relationship between the predicted and observed fields is not only linear but that the two fields are, ideally, identical. This also allows us to perform all of the analysis in a consistent, well-defined context.

2. While the relationship between predicted and observed cloud fields is linear, the predictions are derived from nonlinear regressions. That is, within the constraints of statistical sampling, we obtain all the cloud information available from a specified quantity. This allows us to draw meaningful comparisons between the predictive capabilities of two different quantities. Note that this level of rigor would only be possible from a linear prediction model if the nonlinear terms were insignificant. Our analysis does, however, require that the functional relationship between cloud distributions and their predictors be “smooth” (infinitely differentiable).

This method allows us to perform two important tasks that are demonstrated in this study. We can diagnose dynamical circulations, and we can compare the relative validity of two data sources. The latter allows us to both compare sources of observational data and to assess the performance of models. Within this context, this manuscript addresses the following questions:

1. How important are the thermal histories of air parcels compared to “local” (i.e., at the location of cloud formation) properties for determining the probability of TTL cirrus? It seems reasonable that given enough information (temperature, humidity, turbulent kinetic energy, etc.), one should be able to accurately predict cloud formation from local quantities alone. However, there are advantages to investigating the parcel history as well. For one, knowledge of parcel histories can compensate for missing local information. For example, the local humidity is influenced by the thermal history of air parcels; particularly the Lagrangian cold point, which sets an upper limit on a parcel’s humidity. In addition, understanding the parcel’s thermal history helps us understand dynamical interactions that led to the local atmospheric state. This is a motivating factor, for example, for studies that predict water vapor concentrations in the stratosphere based on Lagrangian dry points [Bonazzola and Haynes, 2004; Fueglistaler et al., 2004;]
2. Experimental Design, Models, and Input Data

2.1. Overview: Experimental Design

The data used to train and validate the cloud predictions are bin averages, where each bin represents a vertical layer in the altitude range 15.2 km ≤ z ≤ 17.0 km (~100–125 hPa) with horizontal/time dimensions of 2° × 2° × 6 h. The specified altitude range is chosen because it is wide enough to provide ample cloud sampling but also narrow enough to limit the variety of cloud formation processes. In this regard, we note that cloud predictions deteriorate noticeably if the layer is enlarged to 13.4–17.0 km (~100–150 hPa). Cloud probability is determined from the fraction of CALIPSO profiles that have thin clouds (optical depth τ < 0.3) within the specified altitude range; that is, a CALIPSO profile is considered to contain thin cirrus if any part of the altitude range overlaps any part of a thin cloud layer or layers. Our criterion for thin clouds follows the definition of thin cirrus set by Sassen and Cho [1992] and includes subvisible cirrus (optical depth τ < 0.03).

For each of the bins we calculate a set of 30 day back trajectories initialized at the bin latitude, longitude and time centers, and at evenly spaced vertical intervals (~95 m apart; 20 trajectories for each bin). Cloud predictor values for each bin are, with the exception of Microwave Limb Sounder (MLS) humidity, obtained from averages over values associated with each back trajectory. The predictors used is this study are: the “local temperature” T₀ (i.e., the initial temperature for each back trajectory), the “local specific humidity” qᵥ, the “Lagrangian cold point differential” δT = T₀ − Tₘᵦᵣₑᵦ, which is the difference between the local temperature and the Lagrangian cold point Tₘᵦᵦₑᵦ, and the “Lagrangian dry point differential” δqᵥ = qᵥₘᵦᵦₑᵦ / qᵥ₀,sᵃᵗ, which is the Lagrangian dry point qᵥₘᵦᵦₑᵦ normalized by the local saturation humidity. Local temperature, the Lagrangian cold point, and the dry point are derived from temperature, humidity, and pressure along each trajectory interpolated from analyses (section 2.3.3). Local humidity for most results is specified from Aura MLS (described in section 2.3.2).

For this study we analyze data from seven winters (January–February 2007–2013) and seven summers (July–August 2006–2012). Predictions for individual seasons (e.g., winter 2007) are based on regressions trained on the complimentary seasons (e.g., winter 2008–2013). Our results are determined from averages over seven seasonal predictions.

2.2. The Trajectory Model

Trajectories are calculated with 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 analyzed meteorological fields to advance the parcel trajectories back in time with the fourth-order Runge-Kutta method at a time step of 30 min. This work employs both kinematic trajectories, which use pressure velocity for vertical motion, and diabatic trajectories, which use diabatic heating.
2.3. Data Sources

2.3.1. CALIPSO Cloud Observations

The observed cloud probabilities are obtained from Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation (CALIPSO) [Winker et al., 2010; Powell et al., 2011] Level 2 5 km cloud layer profiles. These data provide top and base altitudes and optical depth for cloud layers with optical depths as low as 0.01 at 30 m vertical and 5 km horizontal resolution. CALIPSO is a polar orbiter and the cloud profiler is a nadir viewing instrument and, as such, sampling is limited; during a single (2 month) season each 2° bin is sampled, on average, during 5.2 passes of the satellite with approximately 45 observations in each pass. Diurnal sampling is restricted to ~1:30 A.M. and P.M. local time; how this affects our results is uncertain, but the effect could be minimal since the air parcels of interest are typically separated from active convection by more than 24 h [Riihimaki et al., 2012] and in some case more than a week [Pfister et al., 2001]. These data were obtained from the Atmospheric Science Data Center administered by the National Aeronautical and Space Administration (NASA) Langley Research Center.

2.3.2. Aura MLS Water Vapor

Aura Microwave Limb Sounder (MLS) [Schoeberl et al., 2006; Livesey et al., 2011] water vapor mixing ratio at 100 mbar is used as a measure of the local humidity $q_{v0}$. Here we use only MLS observations that are “colocated” with CALIPSO observations. That is, we use MLS observations that are within 3° and 3 h of the bin centers. We extend the bin size to enhance colocation statistics; with the extended spatial dimensions, there is at least one MLS observation for 96% of the CALIPSO observations. The 3° region is consistent with the along track resolution of the Aura MLS data, so there is little degradation to data quality by expanding the bin width for MLS. Aura MLS data for this work were obtained from the Goddard Earth Sciences Data and Information Center administered by the NASA Goddard Space Flight Center.

2.3.3. Meteorological Data

Meteorological data used for the thermal and moisture fields as well as the forcing data for the trajectory calculations are obtained from three sources:

1. Modern-Era Retrospective analysis for Research and Applications (MERRA) [Rienecker et al., 2008; Rienecker et al., 2011] provides three hourly data at 1.25° horizontal resolution on 42 pressure levels (25 levels from 1000 to 100 mbar and 17 levels at pressures < 100 mbar). These data were obtained from Goddard Earth Sciences Data and Information Center.

2. The Global Forecast System (GFS) operational analysis produced at the National Centers for Environmental Prediction provides six hourly data at a horizontal resolution of 1° on 26 pressure levels (21 levels from 1000 to 100 mbar and 5 levels at pressures < 100 mbar). GFS data were obtained from the Computational and Information Systems Laboratory (CISL) Data Support Section at the National Center for Atmospheric Research (NCAR).

3. The ERA-interim reanalysis [Dee et al., 2011] is produced by the European Centre for Medium Range Forecasts and provides six hourly data at 0.7° horizontal resolution with vertical levels at 50 mbar, 70 mbar, and 25 mbar intervals from 100 mbar to the surface. These data were obtained from the CISL Data Support Section at NCAR.

3. The Regression Model

3.1. The Multivariate Nonlinear Regression

The prediction model is derived using multivariate polynomial regression (see details in Appendix A) of the form

$$C(Y_1, Y_2, \ldots, Y_n) = \sum_{m_0=0}^{M} \sum_{m_1=0}^{M-m_0} \sum_{m_2=0}^{M-(m_1+m_2)+\ldots+m_k} \alpha_{m_0, m_1, \ldots, m_k} \left(Y_1^{m_0} Y_2^{m_1} \cdots (Y_n)^{m_k}\right),$$

where $C$ is the predicted cloud fraction, $M$ is the “order” of the regression, $Y_i$ form a set of $n$ predictors, $Y_i'$ are the corresponding normalized perturbations (see Appendix A), and $\alpha_{m_0, m_1, \ldots, m_k}$ are the regression coefficients. One benefit of using polynomial regression is the direct correspondence to the Taylor expansion. This property provides the conceptual foundation that allows us to unambiguously determine, within well-defined uncertainty criteria (derived in Appendix B), the functional dependence of cloud probability on the specified predictors provided:
1. The function is, for practical purposes, infinitely differentiable.
2. We have sufficient sampling statistics. Constraint 1 should not pose any problem as long as our analysis remains at space and time scales for which the fluid approximation for the atmosphere holds (i.e., at scales large enough to aggregate discrete molecular motions into continuous functions) and constraint 2 can be tested. Using the analogy to the Taylor expansion, we know that the regression converges as its order $M$ becomes large. In fact, it will converge quickly using the normalized predictors $Y_i'$. The downside of this approach is that the total number of regression coefficients $N$ can be computationally prohibitive for even moderate values of $M$ and $n$, following the functional dependence

$$N = \sum_{l=0}^{M} \frac{(l + n - 1)!}{l!(n - 1)!}. \quad (2)$$

For example, a sixth-order expansion has 7 terms for a one-variable regression, 28 terms for two variables, and 84 terms for three variables. With our computing resources, the practical limit for performing three-variable regressions is $M = 9$.

3.2. Convergence and Uncertainty

Figure 2 examines the convergence properties of the prediction model in terms of two important comparison metrics: the explained variance $r^2$ (solid lines) and the root-mean-square (RMS) prediction error (RMSE) (dashed lines). These metrics are displayed as functions of regression order for a one-variable (local temperature $T_0$; black lines), a two-variable ($T_0$ and $\delta T$; red lines), and a three-variable regression ($T_0$, $\delta T$, and $\delta q_v$; blue lines) as functions of regression order. Close approximations of asymptotic values (i.e., at large values of regression order) are obtained at a regression order of 5 and the regressions are stable out to 9 orders (the practical limit of the calculations).

In addition to the uncertainty of determining the asymptotic values of $r^2$ and RMSE, there are uncertainties associated with the limited data sample from which we calculate the regression. We estimated that uncertainty using a Monte Carlo method (Appendix B) and found that the combined uncertainty associated with order convergence and statistical sampling for results shown in this study is less than 0.01 for $r^2$ and less than 0.001 for RMSE.

While explained variance and RMS error allow us to examine differences of amplitude and patterns, they are often redundant in that large correlations accompany small RMS differences as demonstrated in Figure 2. For this study, we have examined both metrics but this manuscript focuses solely on explained variance as a metric; we obtained no additional insight from the analysis of RMSE.

4. Main Results

This section examines the relationship between temperature, specific humidity, and TTL cirrus. This might seem odd since cloud formation is above all a function of local relative humidity; clouds form when the relative humidity exceeds ~1.0. If we were merely seeking a reliable prediction scheme for TTL cirrus, it would
make sense to focus on local relative humidity first. However, we are seeking insights into the dynamics of the TTL, and it is the relative cloud information content of our four predictors and of combinations of these predictors that provides this insight. For example, the relative predictive skill of local temperature and local absolute humidity is found here to be a distinguishing feature of thin cirrus formed under different dynamical conditions, whereas relative humidity is a reliable predictor throughout the tropics. In this analysis, we first compare the cloud information content derived from nonlinear regression to linear regression and then compare the cloud information contained in local temperature relative to Lagrangian cold point. Examining the source of prediction error then allows us to identify regions where the functional dependence of cloud differs, indicating that different dynamical interactions take place. We then take a closer look at interesting dynamical differences that distinguish TTL cirrus formation over the West Pacific during boreal winter from that over southern Asia during summer. The results in this section were obtained using kinematic trajectories forced by MERRA winds. Local temperature and the Lagrangian cold and dry point are calculated by interpolating MERRA temperature and humidity onto trajectories, 100 hPa humidity from Aura MLS determines the local humidity, and CALIPSO observations determine cloud probabilities.

### 4.1. Information Content in Thermal Fields

That cloud formation and temperature should have a close relationship is demanded by the nearly exponential dependence of saturation vapor pressure on temperature and is well established in the TTL [e.g., Clark, 2005; Dessler et al., 2006; Liu et al., 2007; Virts et al., 2010; Riihimaki et al., 2012]. The obvious similarities between spatial patterns of upper tropospheric temperature (Figure 3a) and observed TTL cirrus (Figure 3d) confirm that fact, as does the spatial variance explained by the linear relationship ($r^2 = 0.64$; corresponding to the linear correlation coefficient $r = 0.80$). This value would be considered large in many meteorological contexts, but nevertheless falls short of quantifying how close the relationship between clouds and temperature actually is. By employing a nonlinear regression model (Figure 3b), we find that variance explained by local temperature increases to 0.85; an additional 0.21 over the linear relationship. Figure 3c displays the two-variable cloud prediction using the local temperature and the cold point differential $\delta T$. While adding $\delta T$ to the regression model clearly improves the cloud prediction, the impact is not large; the variance explained by the multivariate regression (0.90) is only 0.05 larger than that explained...
by $T_0$ alone. Does this mean that the thermal history is a poor predictor of clouds? No, on the contrary, $\delta T$ is actually a better predictor of cloud distributions ($r^2 = 0.87$; not shown) than $T_0$. That the cold point does not improve the prediction much is due, in part, to the high correlation between $T_0$ and $\delta T$ ($r^2 = 0.84$). This close relationship can be explained, in part, by the following: on average $\delta T$ must be larger ($\delta T$ can never be negative) for warm $T_0$ than for cold $T_0$ because, for cold $T_0$, $\delta T$ must be small (i.e., temperatures much colder than $T_0$ do not exist).

The predictions of cloud distributions during boreal summer (Figure 4) provide an interesting contrast to the wintertime example shown in Figure 3. There is a respectable linear relationship between local temperature (Figure 4a) and the observed distribution of TTL cirrus (Figure 4d) during summer ($r^2 = 0.42$), although it is much weaker than the corresponding relationship during winter ($r^2 = 0.64$). However, there are significant differences in the added variance explained by nonlinear terms and by the Lagrangian cold point for the two seasons. Whereas nonlinear terms provide a substantially better cloud prediction during winter (compare Figures 3a and 3b), the full nonlinear prediction during summer (Figure 4b; $r^2 = 0.45$) is not much better than the linear one. And whereas adding $\delta T$ to the prediction model only increases the explained variance by 0.05 (from 0.85 to 0.90) during winter, the corresponding increase is 0.20 during summer (from 0.45 to 0.65; Figures 4b and 4c, respectively). These comparisons highlight differences that likely indicate that different dynamical processes are responsible for the formation of TTL cirrus during the two seasons. It is also worth noting that for both seasons the cold point differential is an important predictor; it explains more cloud variance than the local temperature and adding $\delta T$ to the regression model reduces unexplained variance by about 30% (from 0.56 to 0.36 during summer; 0.15 to 0.10 during winter).

4.2. Information Added by Humidity

To summarize the cirrus information content within thermal and moisture fields, Table 1 displays the explained variance by the individual predictors (bold faced values along diagonals) and by pairs of predictors (off diagonal values) for winter and summer. The values in Table 1 are, in many respects, consistent with current cloud formation paradigms.

1. The combination of local temperature and local humidity is the best predictor of TTL cirrus during both seasons. This almost certainly reflects the fact that local relative humidity is predominantly a function of $T_0$ and $q_0$. The differences between these predictions and those by other selected combinations (notably, $T_0$ and $\delta T$ during winter; $q_0$ and $\delta q_0$ during summer) are small, but statistically significant. In addition, no sets of predictors, including three-variable sets (not shown), were found to provide better

![Figure 4](https://example.com/figure4.png)

**Figure 4.** As in Figure 3 except for seven-summer (July–August) averages. In addition to the different spatial patterns that favor cloud formation over southern Asia, differences with the winter results from Figure 3 include that there is only a small amount of additional information in the nonlinear terms of the local temperature-cloud relationship compared to the linear relationship and the cold point differential adds substantially to the explained variance.
Thus, predictions based on $T_0$ and $q_{v0}$ represent the standards to which we should compare all other predictions.

2. The best single-variable cloud predictor is $\delta q_v$ although during winter $\delta T$ is nearly as good. That is, the Lagrangian cold and dry point differentials are better predictors than either of the local quantities despite being derived from trajectories calculated from analyzed fields in an observation-poor region. It, perhaps, should be no surprise that $\delta T$ and $\delta q_v$ are the best individual predictors since they were selected because of their conceptual connection to cloud formation. That is, they each measure how moist an air parcel is relative to local saturation, whereas the local values measure how moist a parcel is relative to tropical mean conditions. Nevertheless, the fact that these results are consistent with our conceptual framework is important; it helps substantiate our analysis.

4.3. Diagnosing Dynamical Differences

The cloud predictions based on thermal fields shown in Figures 3c and 4c are each derived from a single functional relationship (one for each season) applied to the entire tropics. As a result, prediction errors arise in regions that have anomalous functional dependencies of cloud probability on thermal fields; that is, the dynamics that lead to cloud formation are somehow different in these regions than over most of the tropics. In that context, the notable property of the error fields (observed minus predicted cloud probability; Figure 5) is that the anomalous cloud formation is associated primarily with continental convection; near south equatorial Africa and South America during boreal winter (Figures 5a) and over South Asia during summer (Figures 5b). It is well established that continental convection differs from maritime convection; for example, each has a different characteristic diurnal cycle [e.g., Bergman and Salby, 1996]. It is reasonable, then, to speculate that the prediction errors displayed in Figure 5 arise, in part, from circulation differences.

Table 1. The Explained Variance by the Individual Predictors and by Pairs of Predictors for Winter and Summer

<table>
<thead>
<tr>
<th></th>
<th>$T_0$</th>
<th>$\delta T$</th>
<th>$q_{v0}$</th>
<th>$\delta q_v$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Winter</td>
<td>0.85</td>
<td>0.90</td>
<td>0.88</td>
<td>0.88</td>
</tr>
<tr>
<td>Summer</td>
<td>0.45</td>
<td>0.65</td>
<td>0.83</td>
<td>0.72</td>
</tr>
</tbody>
</table>

(a) Winter

(b) Summer

Figure 5. The spatial distribution of seven-season prediction errors by the combined local temperature and cold point differential regression for (a) boreal winter 2007–2013 and (b) summer 2006–2012. Contour intervals mimic those in Figures 3 and 4; negative values are represented with dashed contour lines. The regions outlined by thick dashed lines are references for Figure 6. The largest prediction errors are concentrated near continental locations; equatorial South America and Africa during winter, southern Asia during summer.
To delve deeper into this issue, Figure 6 compares predicted to observed temporal variations of regional cloud probability for (a) a region where the tropics-wide cloud prediction works well (the West Pacific during boreal winter; outlined with thick dashed lines in Figure 5a) and (b) for a region where the tropics-wide prediction is less successful (South Asia during summer; outlined with thick dashed lines in Figure 5b). For each panel, the gray line represents the smoothed daily time series of cloud probability determined from CALIPSO observations; the black line is the cloud probability predicted from local temperature and humidity, and the blue and red lines represent predictions from the local temperature and local humidity alone. The prediction from the combination of local temperature and humidity tracks the observed variability for both regions; this is not surprising given the expected close relationship between cloud formation and relative humidity. The interesting result, and one from which we can extract dynamical information, is found in the comparison of the individual temperature and humidity predictions. For the West Pacific, local temperature explains a substantial fraction (0.53) while the local humidity explains a very small fraction (0.03) of the cloud variance. For South Asia during summer, the reverse is true; local humidity is a better predictor ($r^2 = 0.26$) than local temperature ($r^2 = 0.09$). This phenomenon deserves more detailed analysis, but we can nevertheless speculate that these differences reflect the importance of two fundamentally different types of TTL cirrus identified in previous studies using the temporal separation between convection and cirrus formation [Pfister et al., 2001; Riihimaki et al., 2012]. The West Pacific results are consistent with “in situ” TTL cirrus formation in which cirrus formation occurs at least a few days after the convective injection of moisture into the TTL [Jensen et al., 1996]. In this case, since the specific humidity of the air was set days in the past, the timing of cloud formation coincides with temperature fluctuations. The closer ties between humidity and cirrus formation evident in South Asia are consistent with a more immediate connection between TTL cirrus and convection, one in which TTL cirrus form as the result of the direct injection of moisture near the tropopause by deep convection. Such an injection can lead to strong relative humidity fluctuations in the absence of strong temperature fluctuations, which reduces the predictive capabilities of temperature. This analysis, then, provides an insightful complement to the distinction between types of TTL cirrus obtained from previous studies.

5. Model Sensitivities

This section contains results of additional calculations that help put the results of the previous section into a better-defined context, explore the robustness of our results, and diagnose the relative reliability of different data sources.

5.1. Interannual Variations

Since the regression model is trained over a specified time period and geographical region, the results we quote do not necessarily generalize to all times and regions; our results must be interpreted within the context they
are derived. Here we use interannual variations to explore the range of applicability of the results shown in Figures 3 and 4. Figure 7 displays the variance explained by local temperature alone (blue bars) and by the two-variable regression using $T_0$ and $\delta T$ (gray bars) for predictions of cloud distributions during individual seasons as well as the seven-season average. Interannual variability represents a larger source of uncertainty than either sampling or limited regression order; the standard deviation associated with each set of predictors is $\sim 0.03$. Nevertheless, the enhancement of explained variance associated with combining the cold point differential with local temperature to predict clouds cited in section 4 is a robust result that is evident in all years. The larger values of $r^2$ for 7 year averages compared to the individual seasons is the expected consequence of averaging.

5.2. Data Source Sensitivity

The final issue we address is the sensitivity of our results to the data source used. Such sensitivities serve both to test the robustness of results and to evaluate the relative reliability of data sources. Figure 8 displays the (summer) explained variance by two-variable ($T_0$ and $\delta T$) predictions using different data sources for temperature and forcing fields. These are kinematic trajectories using MERRA (blue), diabatic trajectories using MERRA (gray), kinematic trajectories using the GFS operational analysis (red), and kinematic trajectories using ERA-interim (black). The three kinematic calculations all give comparable results (seven-season averages of $r^2$ in the range 0.62–0.67), the lone outlier being the diabatic calculation ($r^2 = 0.50$). Since the MERRA diabatic and kinematic predictions are derived from identical temperature data, their difference must be due to differences in the Lagrangian cold point, which means that in the context of predicting cloud fields the trajectories derived from MERRA, diabatic calculations are inferior to those derived from MERRA kinematic calculations. It is tempting to use this result to weigh in on the discussion regarding the relative merits of the kinematic formulation versus the diabatic formulation.
formulation [Danielsen, 1961; Schoeberl et al., 2003; Ploeger et al., 2010, 2011; Schoeberl and Dessler, 2011]; however, diabatic heating rates from MERRA have known weaknesses—particularly during summer [Bergman et al., 2013; Wright and Fueglistaler, 2013]. Furthermore, there is no significant difference in the cloud predictions by the four data sets during winter (not shown; seven-season averages of \( r^2 \) are in the range 0.90–0.91 for all four data sources).

Figure 9 compares cloud predictions using local humidity from MERRA (blue bars) and Aura MLS (gray bars) for winter. Note that MLS is a better predictor for TTL cirrus for all years. In this context, it is important to note that MERRA assimilates moisture sensitive radiances from HIRS, Stratospheric Sounding Unit, microwave sounding unit, advanced microwave sounding unit, Special Sensor Microwave Imager, GOES, and Atmospheric Infrared Sounder—but not MLS. Does this mean that humidity from MLS is superior to these other sources? Perhaps, but not necessarily; MLS measurements are more sensitive to water vapor variability near the tropical tropopause [Schoeberl et al., 2006] and, so, problematic model components such as the convective parameterization might lead to water vapor errors in MERRA. However, it is important to note that Aura MLS data are (by design) colocated with CALIPSO observations, whereas the data sources used by MERRA are not. Thus, the better predictions using MLS are, in part, artifacts of strategic sampling and do not necessarily make MLS the better data for all applications.

6. Summary and Conclusions

This manuscript explores the use of nonlinear regression models that predict distributions of cirrus clouds for diagnosing transport through the Tropical Tropopause Layer. To this end, we first calculate 30 day back trajectories from the locations of CALIPSO observations to obtain Lagrangian dry and cold points for each time/location. Using these historical values in addition to the local temperature and humidity, empirical predictions for cloud probability are derived and compared to observed clouds. Important properties of this analysis are (1) all results are obtained from the comparison of predicted and observed fields, which permits the use of linear comparison metrics such as correlation and puts all results into a single context, easing interpretation. (2) The prediction models are nonlinear, which arguably extracts all meaningful dynamical information from the relationship between TTL cirrus and its predictors. This allows us to unambiguously assess the information in one set of predictors relative to another, which in turn allows us to diagnose dynamical interactions as well as assess the relative validity of different data sources.

Our results demonstrate that statistical sampling allotted by CALIPSO observations over seven seasons is sufficient, within a small uncertainty (less than 0.01 for \( r^2 \)), to retrieve the full nonlinear relationship between cloud probability and its predictors and that substantial information is lost in a linear analysis. During boreal winter, a single nonlinear function involving the local temperature and historic cold point explains 90% of the variance for spatial variations of the seven-season cirrus probability. While local temperature by itself explains the bulk of cloud variance, the cold point is still important; the cold point differential explains more cirrus variability than the local temperature and, when combined with local temperature, reduces the unexplained variance by a factor of one third. The predictive capabilities of thermal fields are lower during summer primarily due to poor predictions over the Asian summer monsoon region. However, those poor predictions do not result solely from a weak relationship between clouds and thermal histories but rather because there is a different functional relationship over southern Asia compare to the rest of the tropics. Analyses that isolate variations within the West Pacific during boreal winter and southern Asia during summer reveal a striking contrast between the dynamical conditions that promote the formation of TTL cirrus in these two regions; in short, TTL
cirrus formation over regions of active maritime convection, such as the West Pacific, is thermally dominated, indicating an environment in which in situ cirrus are readily formed, while TTL cirrus of southern Asia is moisture dominated, indicating a more direct connection between convective injection of moisture and thin cirrus. Sensitivity experiments show that qualitative aspects of our results (i.e., the relative amounts of cloud information in different sets of predictors) are robust to interannual variations and to the choice of analysis data set. The notable exceptions are the poor predictions based in the cold point derived from diabatic trajectories using heating rates from MERRA. This likely reflects a well-known weakness of diabatic heating rates in the tropical upper troposphere from MERRA. We also found that water vapor distributions from Aura MLS are better predictors of CALIPSO cloud fields than those from MERRA. It is important to remember that these assessments are performed in the context of TTL cirrus formation and, while this is an important context, they do not necessarily generalize to all applications. Overall, our results demonstrate a potentially valuable analysis method that will allow us to make strong statements regarding dynamical interactions in a region where data uncertainties are large.

Appendix A

The regression is performed as follows. Consider training vectors for cloud observations \( Y \) and predictors \( X \), where \( j = 1, 2, \ldots, n \) and \( n \) is the number of predictors. The elements of \( Y \) and \( X \) correspond to values for each \( 2^\circ \times 2^\circ \times 6 \) h bin in the training data sets. To promote rapid convergence with regression order \( M \), we remove the mean and normalize by the standard deviation of each \( X_j \) and use the perturbation quantities \( X_j' \) in the regression.

\[
X_j' = \frac{X_j - \overline{X}_j}{\sigma(X_j)}
\]

where \( \overline{X}_j \) are the elements of \( X_1 \), \( \overline{X}_j' \) are the elements of \( X_j' \), and \( N_j \) is the number of those elements. The predicted cloud probability \( C \) as a function of predictor values \( (Y_1, Y_2, \ldots, Y_n) \) is then,

\[
C(Y_1, Y_2, \ldots, Y_n) = \sum_{m_0=0}^{M-1} \sum_{m_1=0}^{M-1} \cdots \sum_{m_n=0}^{M-1} \alpha_{m_0, m_1, \ldots, m_n} (Y_1)^{m_0} (Y_2)^{m_1} \cdots (Y_n)^{m_n},
\]

where the \( \alpha \)'s are regression coefficients. Note, there is a regression coefficient for each permutation of the exponents \( m_i \). We use \( Y_j \) for predictors to distinguish the values used to predict cloud probability from the \( X_j' \), which are used to train the regression model. Perturbations values for the predictors are calculated from the mean and standard deviation of training predictors

\[
Y_j' = \frac{Y_j - \overline{Y}_j}{\sigma(Y_j)}.
\]

Appendix B

A goal of this analysis method is to obtain well-defined and conceptually meaningful metrics with which to make quantitative comparisons. In order to achieve that, we need two measures of uncertainty: the first is a standard measure of sampling uncertainty, that is, the uncertainty of metrics \( r^2 \) and RMSE due to the finite size of the training data set. To calculate this source of uncertainty we (1) take 200 random subsamples (each composed of one tenth of the total samples) of the observed and predicted cloud data, (2) calculate the metrics \( r^2 \) and RMSE for each subsample, and (3) calculate the standard deviations of the metrics over the 200 estimates. Since this estimate of uncertainty is valid for a sample size one tenth of the size of the training data, the calculated sampling uncertainty overestimates the true uncertainty by a factor of \( \sqrt{10} \).
The second source of uncertainty results from estimating the asymptotic values of $r^2$ and RMSE from a finite number of regression coefficients. To estimate the asymptotic values, we examine the difference

$$\Delta y_m = y_m - y_{m-1} \quad (B1)$$

where $y$ represents one of the metrics and $y_m$ is the estimate of $y$ using an $m$-order regression. With this definition, the functional dependence of $\Delta y_m$ with increasing $m$ defines the convergence properties of the polynomial regression. Figure B1 shows this functional dependence for the six quantities plotted in Figure 2 in terms of $\log_{10}(\Delta y_m)$. For orders 2–9, the decrease of $\log_{10}(\Delta y_m)$ with increasing order is approximately linear, and thus, $\Delta y_m$ decreases exponentially with order. That being the case, we can use values of $\Delta y_m$ to estimate the asymptotic values $y_m$ using

$$y^m_0 = y_m + \frac{\Delta y^2_m}{\Delta y_m} \quad (B2)$$

where $y^m_0$ is an estimate of $y_m$ based on an $m$-order regression. Note that if the convergence is precisely exponential then $y^m_0 = y^m$ for all $m > 1$.

For all results, we estimate the asymptotic values of $r^2$ and RMSE and the corresponding uncertainties as follows:

1. Calculate values $y_m$ and $y^m_0$ for orders 2–9.
2. Determine the highest relevant order as the highest value of $m$ for which the sign of $\Delta y_m$ indicates continued convergence. That is, the smallest $m$ for which $\Delta y_{m+1} < 0$ for $y = r^2$ ($\Delta y_{m+1} > 0$ for RMSE).
3. Specify $y_m$ as the mean of $y^m_0$ over the highest four relevant orders.
4. Specify the corresponding uncertainty as the standard deviation of $y^m_0$ over the highest four relevant orders.

**References**


**Acknowledgments**

This work benefited from helpful conversations with and comments on early versions of the manuscript from W. Randel and three anonymous reviewers. J. Bergman and E. Jensen were visitors at the Atmospheric Chemistry Division 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.


