Using Machine Learning to Predict Near-Surface Air Temperature

Nathan Friesenhahn
AOS C204 Fall 2025

Introduction

The relationship between near-surface air temperature, humidity, and pressure is governed by fundamental atmospheric thermodynamics. In particular, the Clausius–Clapeyron relation describes the exponential increase in saturation vapor pressure with temperature, implying that warmer air masses can generally hold more water vapor. At global scales, this results in a strong and physically meaningful covariance between specific humidity and air temperature. The objective of this project is to examine the extent to which this relationship can be captured by simple machine learning models. Specifically, I investigate whether global-mean near-surface air temperature can be predicted from a small set of global-mean atmospheric variables, evaluate several regression approaches, and explore whether explicit encoding of the seasonal cycle improves predictive skill. Although global averaging removes spatial structure, it provides a simplified dataset for evaluating how much of Earth's large-scale temperature variability can be inferred from basic thermodynamic relationships.

Rather than constructing a comprehensive forecasting system, the purpose of the analysis is to assess to what extent Earth's large-scale temporal variability can be reconstructed from a very limited feature set, and to examine what these models reveal about the underlying structure of the climate system.

Data

The analysis uses the ECCO Version 4 Release 4 (V4r4) surface atmospheric forcing dataset spanning 1992–2017. For each monthly time step, the dataset provides global gridded fields of near-surface air temperature, specific humidity, surface pressure, and wind variables on a regular latitude–longitude grid. To focus exclusively on temporal evolution, all fields were converted to global, area-weighted means, yielding a time series of global-mean temperature, specific humidity, and surface pressure, each containing 312 samples.

Inspection of the resulting time series (*Fig. 1*) reveals several features that motivated the modeling choices used later in the study. Global-mean temperature exhibits a pronounced and highly regular annual cycle with a peak-to-trough amplitude of several Kelvin, superimposed on

a small long-term warming trend. When temperature is plotted against humidity (*Fig. 2*), the points fall along an approximately linear curve with moderate scatter. These characteristics suggest that (1) humidity should contain strong predictive information about temperature, and (2) much of the variance in the dataset is associated with the seasonal cycle. Together, these observations inform the model designs described in the following section. I also plotted temperature against pressure to check for the expected relationship given by the ideal gas law. However, because surface pressure varies only by ~1% globally and is largely determined by column mass rather than temperature alone, its global-mean fluctuations are too small to show the ideal-gas-law relationship visible at regional scales.
This plot is shown in *Fig. 3*, which shows no meaningful relationship between these two variables. I will discuss this more in the discussion section, but for now, I will keep pressure as a possible second feature due to the physically expected proportionality.

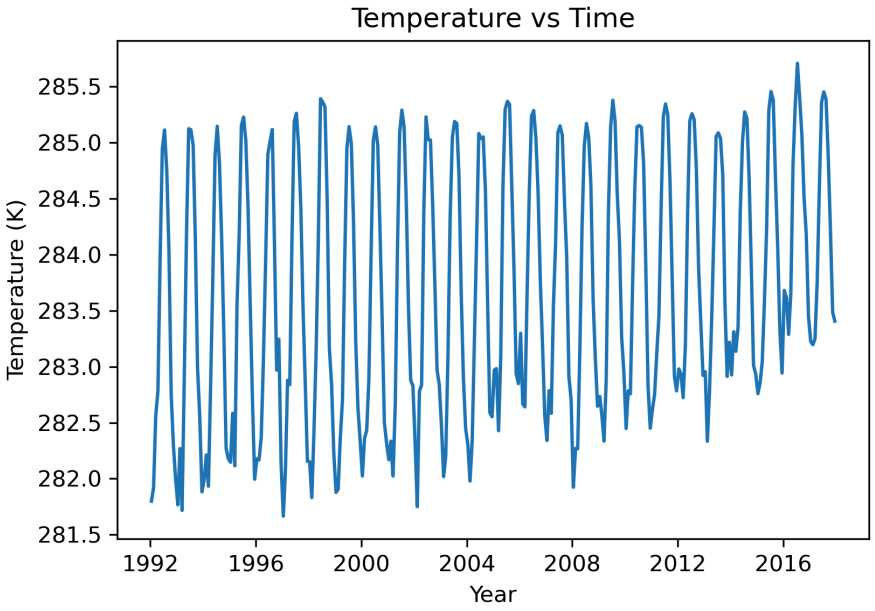


Fig. 1 - Plot of global-mean near surface air temperature in Kelvin vs. time in years.

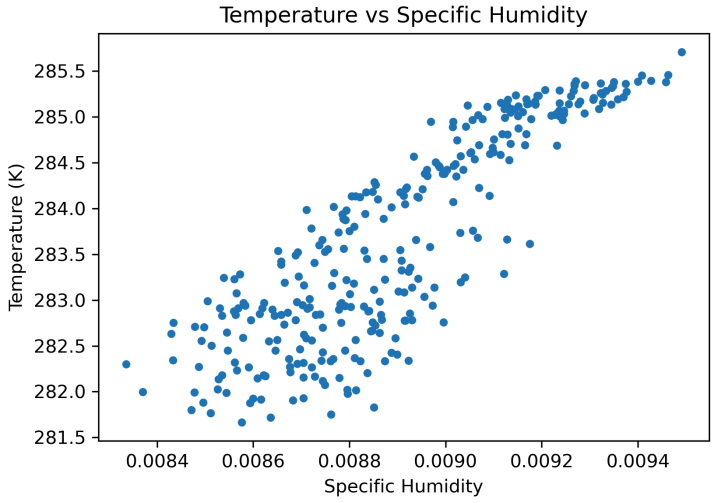


Fig. 2 - This is a plot of global-mean temperature (K) vs. specific humidity. That shows a strong linear proportionality with moderate scattering.

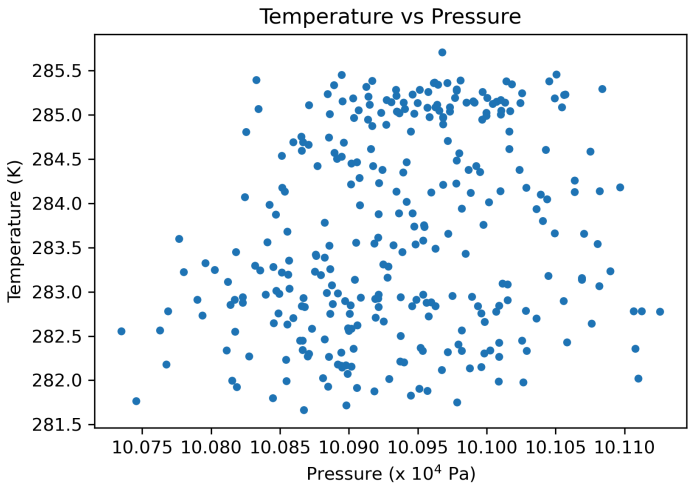


Fig. 3 - This is a plot of global-mean temperature (K) vs. Pressure (x10^4 Pa). There is large scatter and no visible relationship between the two variables despite the underlying known physical relationship.

The first 250 samples (1992–2012) were used for model training, while the remaining 62 samples (2013–2017) served as the test set to simulate out-of-sample forecasting. Where appropriate, input features were standardized using statistics computed only from the training period.

Methods

The modeling framework was designed to reflect both the physical relationships present in the atmosphere and the statistical structure observed in the data. Three regression approaches were tested: ordinary least squares linear regression, ridge regression with L2 regularization, and support vector regression (SVR) with a radial basis function kernel. These models were first trained using specific humidity as the sole predictor, given the strong apparent relationship between humidity and temperature identified in the exploratory inspection of the dataset.

The goal of this project is to test how well simple, physically interpretable relationships can reproduce global-mean temperature variability, so linear regression is an appropriate first model. The Clausius–Clapeyron relationship between temperature and specific humidity is approximately exponential at the molecular level, but in the narrow range of globally averaged humidity values (~0.0084–0.0094 kg/kg) the relationship is effectively linear. As a result, a linear model can capture most of the variance without requiring nonlinear kernels or deep learning approaches.

Ridge regression introduces an L2 penalty to prevent coefficient inflation when models overfit or when features are highly collinear. Because our model uses only a single predictor and the relationship is nearly linear, the ordinary least squares solution is already stable. As a result, the

Ridge solution closely matches the unregularized linear model, and performance differences are negligible.

SVR is most effective when nonlinear structure is present or when high-dimensional feature interactions matter. In this dataset, however, the temperature–humidity relationship is nearly linear and the predictor space is one-dimensional. SVR's ϵ-insensitive loss discards small but meaningful variations, and the RBF kernel provides no advantage when the underlying relationship does not deviate significantly from linearity. With limited sample size and no extensive hyperparameter tuning, the model underfits, leading to worse performance than simple linear regression.

Because surface pressure exhibits relatively modest global-mean variability, but may still contain weak covariances with temperature through hydrostatic balance or circulation changes, each model was then retrained using both humidity and surface pressure as predictors. Although temperature and pressure are physically connected through the ideal gas law, global-mean surface pressure varies by only ~1% and is dominated by mass redistribution rather than thermodynamic changes. As a result, the global-mean pressure series shows no meaningful correlation with temperature. Adding pressure therefore introduces noise rather than information, slightly degrading model performance relative to humidity alone.

Finally, the pronounced seasonal cycle visible in all variables motivated the construction of a model incorporating the month of year as an explicit predictor. This "phase-folded" formulation allows the regression to learn seasonally varying temperature–humidity relationships and mirrors techniques used in the analysis of periodic astrophysical signals, where observations are mapped onto a single representative cycle to highlight underlying structure. The phase-folded model was implemented using a simple linear regression in two variables (specific humidity and month index), trained on the same chronological training set.

Model performance on the test set was quantified using root-mean-square error (RMSE) and the coefficient of determination (R^2), representing the fraction of variance explained.

Results

The humidity-only models (*Fig. 4*) performed well. Linear regression achieved an RMSE of 0.561 K and an R^2 of 0.638, while ridge regression gave nearly identical results. These models accurately reproduce the seasonal cycle and capture more than sixty percent of the variance in the test data. SVR performed less effectively, with a higher RMSE and a noticeably lower R^2,

likely reflecting the challenges of fitting a nonlinear model in a very low-dimensional, relatively small dataset.

Including surface pressure as an additional predictor did not improve performance for the linear or ridge models (*Fig. 5*); in fact, both models performed slightly worse than their humidity-only counterparts. SVR showed modest improvement when pressure was added, but still lagged behind the simpler linear models.

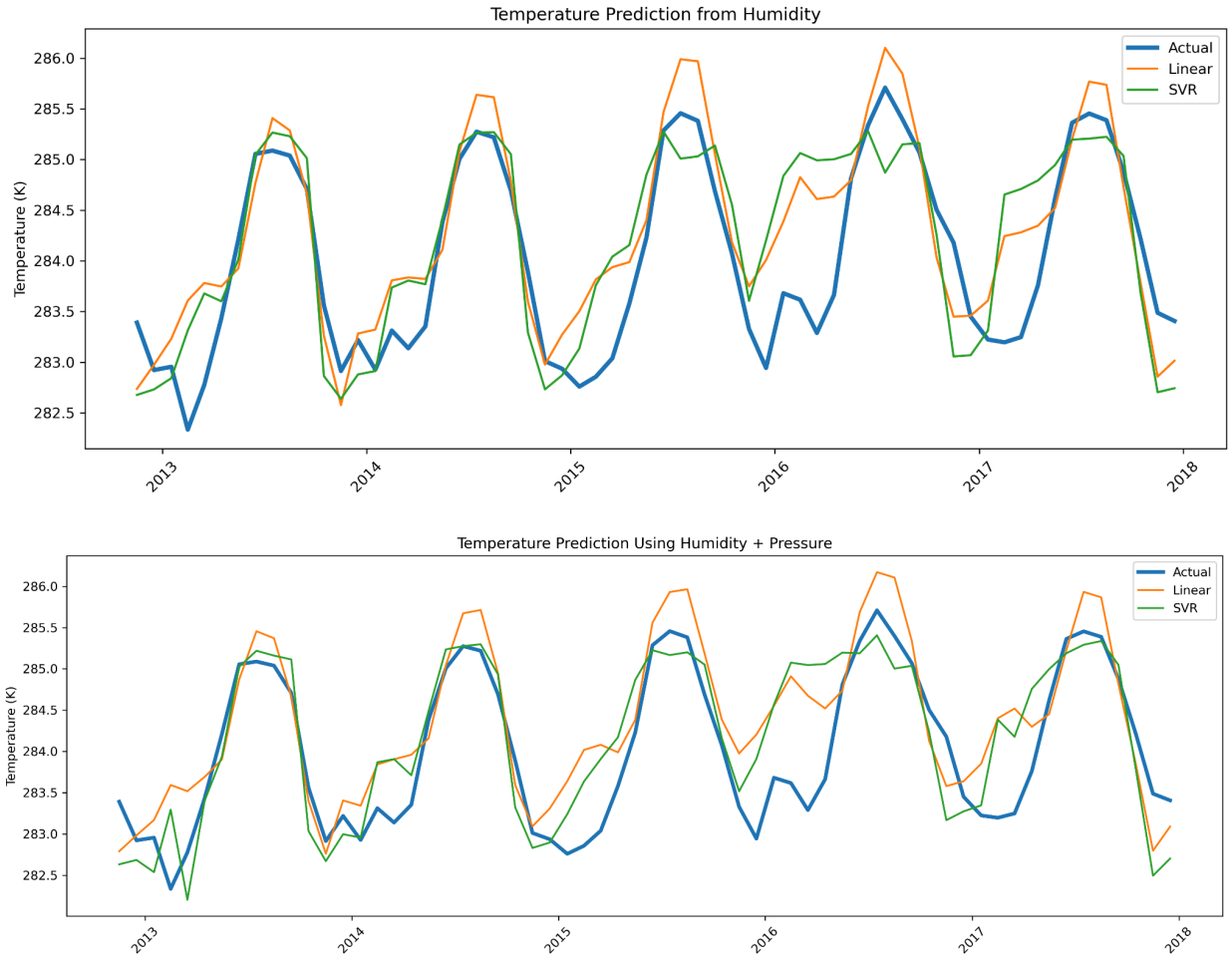
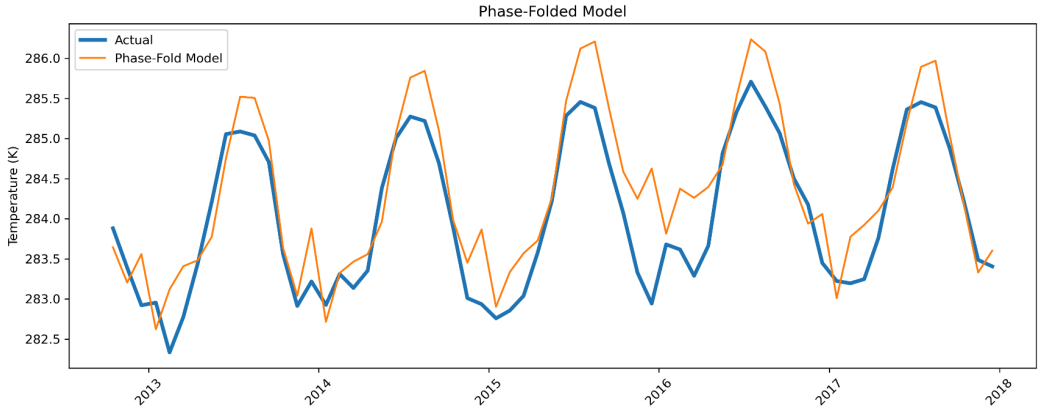


Fig. 4 *(top)* - This plot overplots both the linear (orange) and SVR (green) models using humidity as the feature. The Actual temperature is shown in blue. The models overshoot temperatures earlier in the year, seen very clearly in 2016, whereas they are more accurate in the latter half of the year. Interestingly, the linear model overshoots the peak temperatures, while the SVR model undershoots it.
Fig. 5 *(bottom)* - Similar to the previous figure, this plot shows the linear and SVR models trained using both humidity and pressure as features.

***Fig. 6** (below)* - This plot shows the predicted temperature from the phase-folded model (orange) over the actual temperature (blue). Visually, this model is more accurate than the previous models, specifically in the first half of the year.



The linear phase-folded model (*Fig. 6*), by contrast, produced a clear improvement over all other approaches. By including the month of year along with humidity, the linear regression achieved an RMSE of 0.509 K and an R^2 of 0.698 (*Fig. 7*). The model not only reduced overall error but also more accurately tracked the amplitude and timing of the seasonal cycle when evaluated against the real chronological data. This result indicates that the temperature–humidity relationship varies systematically through the year and that explicitly encoding the position within the seasonal cycle provides the model with physically meaningful information that enhances predictive skill.

***Fig. 7** (below)* - This table shows the RMSE and R^2 values for each model of the previously discussed models. Specific humidity is referred to as “qh” and pressure is “p”.

Model Performance

Model	RMSE	R^2
Linear(qh)	0.561	0.638
Ridge(qh)	0.559	0.641
SVR(qh)	0.661	0.498
Linear(qh+P)	0.611	0.572
Ridge(qh+P)	0.607	0.577
SVR(qh+P)	0.602	0.584
Phase-fold	0.509	0.698

Discussion

The strong performance of the humidity-only linear models underscores the powerful thermodynamic constraint imposed by Clausius–Clapeyron: global-mean near-surface temperature and specific humidity co-vary in a nearly linear fashion, reflecting the basic physics of water vapor capacity in the atmosphere. Because the seasonal cycle dominates the variance in these global means, models that reproduce this relationship capture much of the observed behavior.

However, the diminished performance of the two-feature models indicates that surface pressure does not contribute meaningful additional information for this prediction task. Surface pressure exhibits relatively weak global-mean variability compared to temperature and humidity (*Fig. 3*), and its small fluctuations do not appear to correlate strongly with temperature in a way that benefits linear regression.

The enhanced performance of the phase-folded model highlights the importance of seasonal structure. By explicitly categorizing the data by month of the year, the model is able to learn distinct temperature–humidity relationships for different parts of the annual cycle. This is consistent with physical expectations: the climate system responds to humidity differently in different seasons due to changes in insolation, circulation patterns, and land–sea contrasts. The success of this simple seasonal predictor suggests that incorporating physical insight, in this case the periodic nature of Earth's orbit, can significantly improve even the simplest machine learning models.

Nevertheless, the framework has clear limitations. The analysis treats Earth as a single, globally averaged box and therefore ignores spatial structure that is essential to real climate dynamics. The dataset provides only about 300 samples, limiting the complexity of models that can be trained. Furthermore, because the models predict total temperature rather than anomalies, they primarily learn the seasonal cycle rather than interannual variability. Processes such as ENSO and decadal oscillations remain essentially invisible in this setup.

Conclusion

Simple machine learning models are able to capture a large fraction of the temporal variability in global-mean near-surface temperature using only specific humidity as a predictor. Linear and ridge regressions explain approximately 64 percent of the variance, consistent with expectations derived from thermodynamic theory. Incorporating global-mean surface pressure does not

improve performance, but encoding the seasonal cycle through the month of year substantially enhances predictive skill, raising the explained variance to nearly 70 percent.

These results demonstrate that even minimal models can reveal physically meaningful relationships within the climate system when paired with appropriate preprocessing and basic physical insight. Future work could explore modeling temperature anomalies rather than the full seasonal cycle, incorporating lagged predictors to represent atmospheric memory, or applying similar techniques to spatially resolved datasets.

References

ECCO Consortium. (2017). *ECCO Ocean State Estimate – Release 4, Version 4 (ECCO4v4)* [Dataset]. NASA Physical Oceanography Distributed Active Archive Center (PO.DAAC). https://podaac.jpl.nasa.gov/