Final Report

Stefano Graziosi Gabriele Molè Giovanni Carron Laura Lo Schiavo

Group 2

Research Question

We analyze two related dimensions of climate dynamics: the global temperature series and localized temperature behavior in San Francisco, California. First, we examine the NASA GISTEMP global temperature index by fitting a hidden Markov model to identify structural breaks, and by estimating a dynamic linear model to characterize the latent temperature process. Second, we focus on the San Francisco Downtown (DWTN) station from the Global Historical Climatology Network, modeling and forecasting daily maximum and minimum temperatures; we compare a suite of univariate and multivariate dynamic linear specifications to assess their relative performance in capturing local temperature dynamics. Our findings illustrate distinct regime shifts in the global series and demonstrate the value of multivariate state-space methods for short- and medium-term temperature forecasting at the station level.

1. Data acquisition and exploration

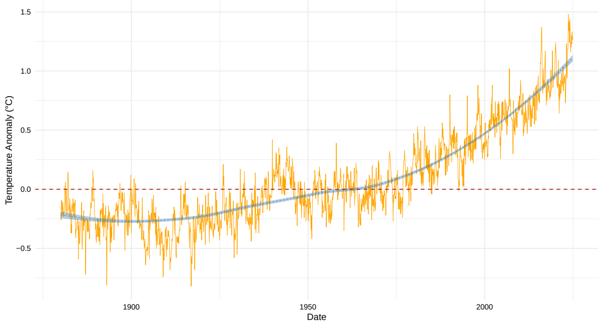
1.1. Trend

Climate: GISTEMP

The long-term temperature records reveal a persistent warming trend, with global anomalies progressively increasing from near-baseline values in the earlier periods to significantly elevated levels in recent decades, thereby suggesting a sustained and systematic shift in the planet's thermal equilibrium.

Furthermore, the data indicate that the rate of warming has not been uniform over time, as evidenced by periods of marked acceleration in recent decades compared to the more gradual increases observed during earlier intervals, implying an intensification of underlying climatic forcings.

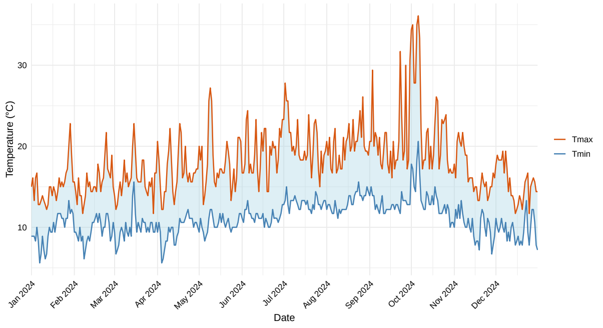
Figure 1-1. In orange, the temperature anomalies in °C compared to 1980. In blue, the long-term trend fitted by *Loess* smoothing methodology, including the standard error (shaded area).



Weather: GHCN

Historically, we observe a fairly consistent series with no apparent trend; despite some rare peaks above 35°C and even more rare drops below 0°C, there are no significant differences between years. As a narrower example, the maximum and minimum temperatures in San Francisco from January to December 2024 (plotted in 1-2 on the left) follow a seasonal cycle, with nighttime lows rising from 9 °C in winter to 15 °C in summer and daytime highs increasing from 15 °C to 23 °C, returning to winter.

Figure 1-2. In red, the maximum daily temperature; in blue, the minimum daily temperature.



1.2. SEASONALITY

CLIMATE: GISTEMP We notice how there is a distinct change between pre-1979 and post-1979 seasonality (calculated as `stl` decomposition seasonality for 30-year lapses), where the former is marked by more pronounced fluctuations; such structural change represent an empirical justification for allowing a time-varying seasonal component.

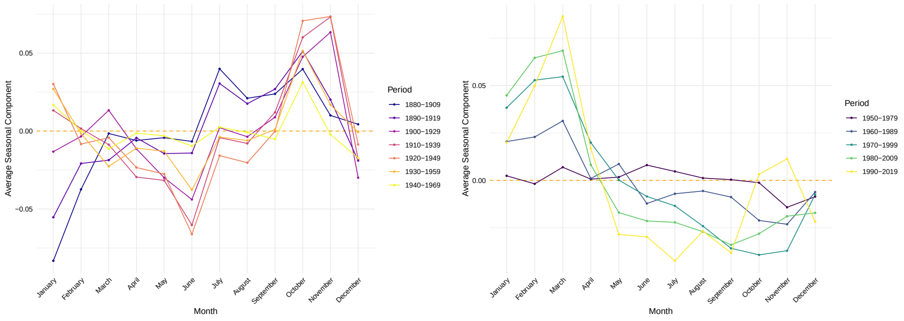


FIGURE 1-3. Comparison of the 30-year seasonal component before 1979 (left) and after 1979 (right).

WEATHER: GHCN The historical evolution of temperatures in San Francisco offers possibly less insights, as we observe the same stationary series with no significant deviations from a predicable pattern; this is evidenced by the below image plotting the smoothed behaviour of the temperature for the last 10 years:

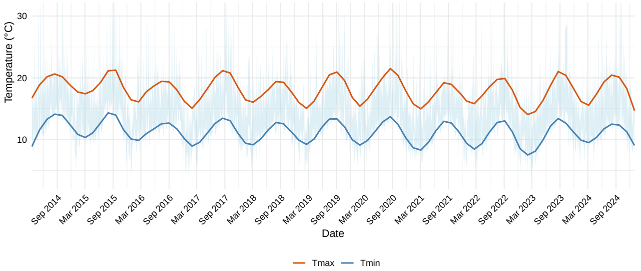


FIGURE 1-4. In red, the smoothed max daily temperature; in blue, the smoothed min daily temperature.

2. GISTEMP: CLIMATE DESCRIPTION

2.1. HIDDEN MARKOV MODELS

A graphical inspection of Figure 1-1 reveals visible change points in the level of temperatures, indicating that the series is not stationary. This motivates the use of an HMM to segment the series into distinct regimes. To better characterize these change points, and understand the evolution of warming across regimes, we specify an HMM with k latent states, a homogeneous transition matrix, and Gaussian emissions distribution. Our model is formulated as follows. Let Y_t denote the deseasonalised monthly anomaly at time t and $S_t \in \{1,\dots,k\}$ the unobserved regime. Conditional on $S_t = j$ we assume a Gaussian emission distribution $Y_t \mid S_t = j \;\sim\; \mathcal{N}\big(\mu_j, \sigma_j^2\big), \qquad j = 1,\dots,k,$; while the latent process follows a time-homogeneous Markov chain with transition matrix $A = (p_{ij})_{i,j \le k}$ and initial distribution $\boldsymbol{\pi}$. The model relies on two main assumptions. First, (A.1) the latent process $(S_t)_{t \ge 0}$ is assumed to be a time-homogeneous Markov chain with a finite state space $\mathcal{Y} = \{1,\dots,k\}$. Second, (A.2) given the latent process, the observations (Y_t) are conditionally independent, with each Y_t depending only on the current state S_t. All parameters, $\theta = \{\boldsymbol{\pi}, A, (\mu_j, \sigma_j^2)_{j \le k}\}$, are obtained by maximum likelihood via the Baum–Welch expectation–maximisation routine.

A crucial aspect of model specification in hidden Markov models (HMMs) is the choice of the number of latent states k. Increasing k generally improves in-sample fit in statistical criteria such as the Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC). However, this comes at the cost of interpretability, introducing regimes that lack clear economic interpretation or policy relevance, and the risk of overfitting. For this reason, and in line with the literature's emphasis on parsimony and interpretability, we focus on identifying structural breaks using two-state and three-state HMMs.

Table 2-1 shows that, with two regimes, the conditional means are neatly separated ($\hat{\mu}_1 = -0.164$, $\hat{\mu}_2 = 0.534$) and paired with distinct volatilities, yielding a clear "cool" versus "warm" interpretation. Notably, the warmer regime is also more volatile ($\hat{\sigma}_2 = 0.309$ vs. $\hat{\sigma}_1 = 0.184$), which is consistent with the idea that climate change is not only increasing average temperatures but also amplifying extremes. The three-state alternative introduces an intermediate regime whose mean (0.024) and variance resemble those of the cool phase, suggesting redundancy. For these reasons, we opt for the 2-regime specification: it is more parsimonious, easier to interpret, and sufficient to capture the key structural changes in the temperature series. Table 2-2 reports the estimated transition probabilities for our preferred model.

	$k=2$		$k=3$		
	State 1	State 2	State 1	State 2	State 3
state-dependent mean $\hat{\mu}$	-0.164	0.534	-0.283	0.024	0.629
	(0.006)	(0.007)	(0.007)	(0.004)	(0.004)
state-dependent standard deviation $\hat{\sigma}$	0.184	0.309	0.137	0.133	0.275
	(0.004)	(0.007)	(0.007)	(0.007)	(0.004)

TABLE 2-1. Maximum Likelihood Estimates of the HMM parameters for competing HMMs.

	To	
From	State 1	State 2
State 1	0.996	0.004
	(0.002)	(0.004)
State 2	0.007	0.993
	(0.002)	(0.004)

TABLE 2-2. Estimated transition matrix for the homogeneous two-state HMM.

The estimated state-dependent means suggest a clear interpretation of the regimes: State 1 corresponds to periods of negative or mildly positive temperature anomalies (mean -0.164), whereas State 2 reflects pronounced warming episodes (mean 0.534). The transition matrix (Table 2-2) reveals strong persistence in both regimes, with probabilities of remaining in the same state exceeding 99% each month. Specifically, the system remains in State 1 with probability $p_{11} = 0.996$ and in State 2 with $p_{22} = 0.993$, while the probabilities of switching between states are notably low ($p_{12} = 0.004$, $p_{21} = 0.007$). These results confirm that once the system enters a regime—either cool or warm—it tends to persist there, reinforcing the notion of slowly evolving climate patterns and validating the use of a hidden Markov model to capture persistent structural shifts in global temperature anomalies.

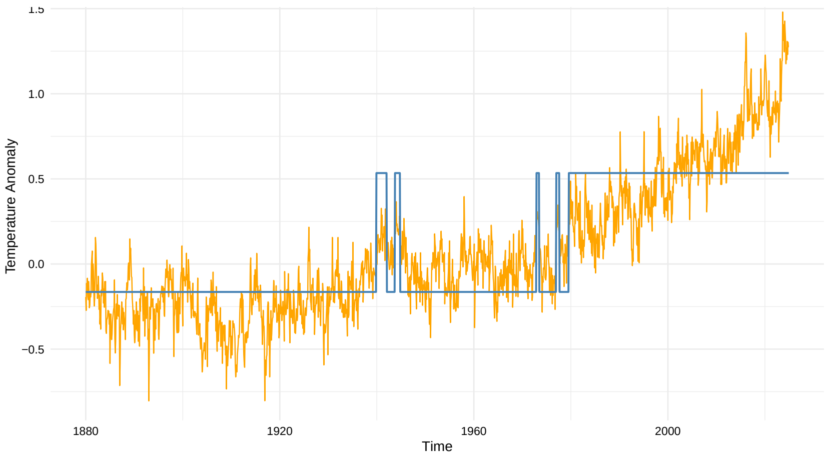


FIGURE 2-5. Graphical representation of the observed values (orange) and means μ_i of the hidden states (blue).

Figure 2-5 juxtaposes the observed anomaly series (orange) with the decoded state-dependent means (blue). The decoded path highlights infrequent but abrupt regime switches and sustained within-regime stability, in line with the visual inspection of the data, further supporting the adequacy of the two-state specification.

2.2. DYNAMIC LINEAR MODELS

While the HMM framework is effective in identifying discrete regime changes in global temperature anomalies, it lacks the ability to model smooth and gradual shifts in the underlying climate signal. To capture such dynamics, we now move to a continuous latent structure and specify a Random Walk Plus Noise (RWPN) model under the assumptions A.1 (with generalized state space) and A.2. This approach allows for sequential estimation of the evolving trend and provides an explicit quantification of uncertainty over time, therefore the model is particularly suited to describe persistent, non-stationary processes such as global warming.

Let Y_t be the deseasonalised anomaly and θ_t the unobserved level at month t. The model couples an observation equation with a random-walk state equation:

$$\begin{aligned} Y_t &= \theta_t + v_t, & v_t &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma^2), \\ \theta_t &= \theta_{t-1} + w_t, & w_t &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma_w^2), \end{aligned} \qquad \theta_0 \sim \mathcal{N}(m_0, C_0),\ \theta_0 \perp (v_t) \perp (w_t). \tag{DLM}$$

This local-level DLM accommodates gradual drifts (through σ_w^2) while retaining Kalman-filter tractability for sequential estimation and forecasting. The model was estimated via maximum likelihood by optimizing the concentrated log-likelihood. The resulting parameter estimates are summarized in Table 2-3, with SE on the log-variance scale.

Parameters	Estimate	Standard Errors (log)
Observation variance $\hat{\sigma}^2$	0.0060	0.0609
State variance $\hat{\sigma}_w^2$	0.0031	0.1062

TABLE 2-3. Maximum Likelihood Estimates of the RWPN parameters.

After estimating the parameters σ^2 (observation noise) and σ_w^2 (state process noise), we apply the Kalman filter for one-step-ahead forecasting, followed by Forward Filtering Backward Smoothing (FFBS) to recover the latent state trajectory. Figure 2-6 shows the smoothed estimates with 95% credible intervals for 1972–1977. We focus on this window for clearer visualization of the model's tracking performance. The RWPN model successfully captures the underlying signal, while preserving uncertainty quantification.

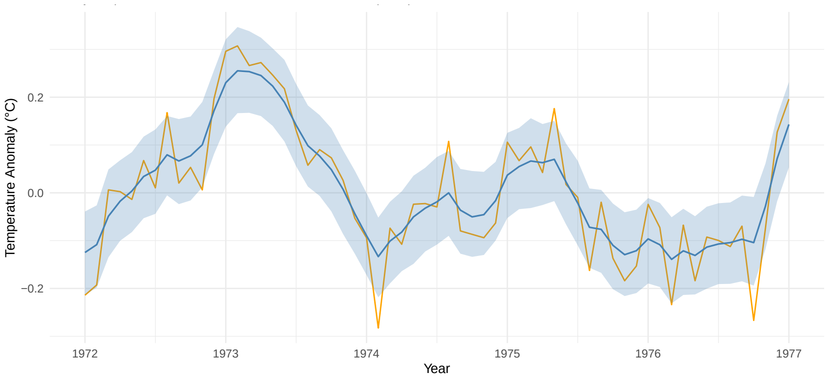


FIGURE 2-6. Observed values (orange) and RWPN one-step-ahead forecasts (blue) with 95% confidence intervals (shaded area).

With $\hat{\sigma}_w^2 = 0.0031$ and $\hat{\sigma}^2 = 0.0060$, the latent state evolves more smoothly than the noisy observations, yet remains sufficiently adaptive to incorporate new information. The Kalman gain indicates a modest weighting toward past states, producing a trend that balances responsiveness with stability. One-step forecast performance is strong, producing RMSE $= 0.207$ and MAE $= 0.183$; the 95% predictive interval captures every held-out value (Coverage$_{95} = 100\%$).

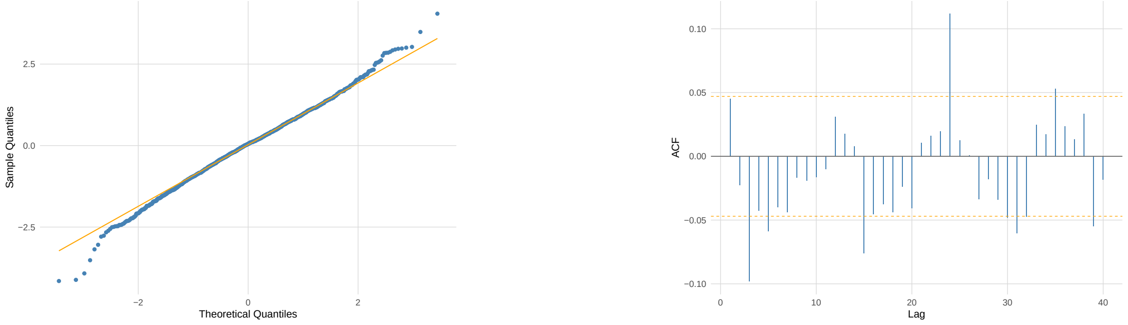


FIGURE 2-7. Normal Q–Q plot (left) and sample autocorrelation function (right) of the standardized one-step-ahead prediction errors. Lag 0 is omitted from the ACF to better visualize potential autocorrelation.

To assess model fit, we analyze the standardized residuals. As shown in Figure 2-7, standardised residual checks reveal that residuals are not fully white noise. The Q–Q plot indicates light deviations from normality in the tails, while the ACF plot—rescaled to highlight low-level structure—shows remaining temporal dependence. This is statistically confirmed by the Ljung–Box test, which rejects the null of no autocorrelation at lag 20 ($\chi^2 = 64.1$, $p < 10^{-5}$).

These residual patterns suggest that the RWPN may be too simplistic to capture the full dynamics of the temperature series. We also considered a local linear trend model as an intermediate alternative. However, it failed to address residual autocorrelation and yielded poorer fit compared to the seasonal specification. The slope variance was nearly zero, and the Ljung–Box test rejected the white noise hypothesis, indicating insufficient flexibility.

2.3. ARIMA(2,1,1) in state-space form

While the Random Walk Plus Noise (RWPN) model effectively captures long-run variability in global temperatures, its structural simplicity leaves short-run autocorrelation unmodeled. We therefore consider an ARIMA$(2,1,1)$ specification within the DLM framework, which remedies this while preserving Kalman-filter convenience.

Let $\Delta Y_t = Y_t - Y_{t-1}$. The model

$$\Delta Y_t = \phi_1 \Delta Y_{t-1} + \phi_2 \Delta Y_{t-2} + \varepsilon_t + \theta_1 \varepsilon_{t-1}, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma^2) \tag{ARMA(2,1)}$$

admits a canonical state-space representation and follows an ARMA$(2,1)$ process.

Maximum likelihood estimation produces well-identified parameters with tight confidence intervals, as reported in Table 2-4. All roots lie outside the unit circle, ensuring invertibility and causal stability. Residual diagnostics indicate model adequacy: the autocorrelation function shows no significant structure, and the Ljung–Box test at lag 20 does not reject the null of no autocorrelation ($p = 0.085$), suggesting the residuals resemble white noise.

Parameters	**Estimate**	**Standard Errors**
ϕ_1	0.474	0.029
ϕ_2	0.190	0.027
θ_1	-0.944	0.016
$\log \sigma^2$	-4.466	0.031

Table 2-4. Maximum Likelihood Estimates of the ARIMA$(2,1,1)$ parameters.

On a 60-month hold-out the model attains RMSE $= 0.118$, MAE $= 0.094$, and a Coverage$_{95\%}$ $= 92.5\%$, thus outperforming the RWPN on residual whiteness while retaining competitive predictive sharpness.

Combining the continuous-drift RWPN with the richer short-term dynamics of ARIMA yields a coherent pair of benchmarks: the first highlights the slow warming signal; the second offers an all-purpose forecasting baseline with well-behaved residuals.

2.4. COMPARISON

The HMM and DLM frameworks capture fundamentally different types of dynamics. The HMM is designed to detect regime shifts, allowing us to identify structural breaks in global temperature anomalies. In contrast, the DLM assumes a gradual evolution of latent states, which better reflects smooth, persistent changes such as long-run warming trends. While the HMM is useful for pinpointing sudden shifts, its stepwise structure may fail to capture gradual trends. On the other hand, the DLM (in particular, the seasonally-augmented RWPN) provides a more accurate fit and smoother estimates, though it still leaves short-term autocorrelation unmodeled. Overall, both models are complementary: the HMM offers a structural segmentation of the time series, whereas the DLM provides continuous tracking over time.

3. GHCN: Weather prediction

To forecast next-day *maximum* and *minimum* temperatures we model the deseasonalised series with three variants of the Random Walk plus Noise (RWPN) state–space model. The RWPN is the simplest non-stationary DLM, featuring a local-level state that drifts as a random walk while observations fluctuate around it.

Because our two weather series are measured at the same site and time, we explore increasingly coupled specifications that allow for cross-series information sharing.

3.1. Independent random walk plus noise models

We first use an independent random walk plus noise model:

$$\begin{aligned} &\mathbf{Y}_{SF,t} = \boldsymbol{\theta}_{SF,t} + \boldsymbol{v_{SF,t}}, \quad v_{SF,t} \sim \mathcal{N}(\mathbf{0}, V), \quad \boldsymbol{\theta}_{SF,t} = \boldsymbol{\theta}_{SF,t-1} + \boldsymbol{w_{SF,t}}, \quad w_{SF,t} \sim \mathcal{N}(\mathbf{0}, W), \\ &\boldsymbol{\theta}_{SF,t} = \begin{bmatrix} \theta_{SF,t,1} \\ \theta_{SF,t,2} \end{bmatrix}, \quad V = \begin{bmatrix} \sigma^2_{v,1} & 0 \\ 0 & \sigma^2_{v,2} \end{bmatrix}, W = \begin{bmatrix} \sigma^2_{w,1} & 0 \\ 0 & \sigma^2_{w,2} \end{bmatrix}. \end{aligned} \tag{I-RWPN}$$

where $(\boldsymbol{Y}_{SF,t})_{t \geq 1}$ are the observed time series for maximum and minimum temperatures for San Francisco station $(\boldsymbol{\theta}_{SF,t})_{t \geq 0}$ is the latent Markov process. Level and noise variances are estimated separately. V is the variance-covariance matrix for the innovation in the observation process. W is the variance-covariance matrix for the innovation in the state process, where this model assumes a 0-covariance restriction. This means that the model treats the innovations of the state process as uncorrelated.

Table 3-5 shows the estimated coefficients by MLE and their standard errors computed through the Hessian method. Estimates for maximum temperature $\sigma^2_{w,2}$, $\sigma^2_{v,2}$ are overall noisier compared to the ones for minimum temperature $\sigma^2_{w,1}$, $\sigma^2_{v,1}$. Both series observe noisier state innovation compared to observation noise ($\sigma^2_{w,1}/\sigma^2_{v,1} > 1; \sigma^2_{w,2}/\sigma^2_{v,2} > 1$). Hence, by the formula of Kalman filtering, predictions will rely more on recent observations.

3.2. Seemingly unrelated random walk plus noise models

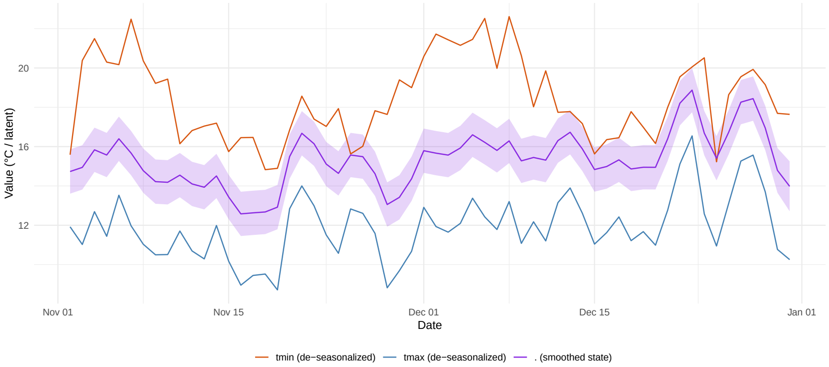
The SUTSE-RWPN model couples the two local levels through a joint system-noise covariance while keeping distinct observation noises. This borrows strength across series without forcing a common trend.

$$\begin{aligned} \mathbf{Y}_{SF,t} &= \boldsymbol{\theta}_{SF,t} + v_{SF,t}, \quad v_{SF,t} \sim \mathcal{N}(\mathbf{0}, V), \quad \boldsymbol{\theta}_{SF,t} = \boldsymbol{\theta}_{SF,t-1} + w_{SF,t}, \quad w_{SF,t} \sim \mathcal{N}(\mathbf{0}, W), \\ \boldsymbol{\theta}_{SF,t} &= \begin{bmatrix} \theta_{SF,t,1} \\ \theta_{SF,t,2} \end{bmatrix}, \quad V = \begin{bmatrix} \sigma^2_{v,11} & 0 \\ 0 & \sigma^2_{v,22} \end{bmatrix}, \quad W = \begin{bmatrix} \sigma^2_{w,11} & \sigma^2_{w,12} \\ \sigma^2_{w,21} & \sigma^2_{w,22} \end{bmatrix}. \end{aligned} \tag{SU-RWPN}$$

where all parameters are as before with the exception of the W matrix where correlation between state innovations is allowed to be non-zero. As shown in Table 3-5, the observation process becomes noisier, while part of the noise in the state process is captured by the significant covariance (3.98) between the innovation of minimum and maximum temperature. The signal-to-noise ratio still favours recent observations, as variances for the observation process are smaller for both time series.

3.3. Random walk plus noise common state process

Figure 3-8. Graph of Maximum and Minimum Temperature with latent state process for CSP-RWPN model.



The Common-trend RWPN model lets both extremes share a single latent level but have their own observation noises. This is the most restrictive specification and tests the hypothesis that daily minima and maxima co-move perfectly once seasonality is removed.

$$\begin{aligned} &\mathbf{Y}_{SF,t} = F\,\boldsymbol{\theta}_{SF,t} + v_t, \quad v_{SF,t} \sim \mathcal{N}(\mathbf{0}, V), \quad \boldsymbol{\theta}_{SF,t} = \boldsymbol{\theta}_{SF,t-1} + \begin{pmatrix} 0 \\ w_{SF,t} \end{pmatrix}, \quad w_{SF,t} \sim \mathcal{N}(0, \sigma^2_w), \\ &\boldsymbol{\theta}_{SF,t} = \begin{bmatrix} 1 \\ \xi_{SF,t} \end{bmatrix}, \quad F = \begin{bmatrix} \alpha_1 & \beta \\ \alpha_2 & \frac{1}{\beta} \end{bmatrix}, \quad V = \begin{bmatrix} \sigma^2_{v,11} & 0 \\ 0 & \sigma^2_{v,22} \end{bmatrix}. \end{aligned} \tag{CSP-RWPN}$$

The common factor model uses a single latent state (ξ_t) driving both Tmin and Tmax through loadings, capturing shared variation. Unlike independent or SUR versions, it imposes a simple low-dimensional statistical structure, enabling information pooling and co-movement modeling, while series-specific observation variances remain. From a purely theoretical standpoint, we expect superior joint out-of-sample predictive accuracy.

The estimated $\hat{\beta}$ is 0.87. This means that the common factor increases the minimum temperature (as $Y_{1,t} = \hat{\alpha_1} + \hat{\beta} * \xi_t$) but at a lower rate compared to the increase in maximum temperature ($1/\hat{\beta} = 1/0.87 = 1.15$). Estimated intercepts ($\hat{\alpha_i}, i \in \{1, 2\}$) are virtually zero while the other parameters remain comparable to previous models.

Model	Parameter	Estimate	Std. Error	Model fit		
				LogLik	AIC	BIC
Independent random walk plus noise models	$\hat{\sigma}^2_{v,1}$	0.4356605	0.027470	−88 999.87	178 007.7	178 041.9
	$\hat{\sigma}^2_{v,2}$	2.0729580	0.033720			
	$\hat{\sigma}^2_{w,1}$	0.8947723	0.020332			
	$\hat{\sigma}^2_{w,2}$	5.2174376	0.021330			
Seemingly unrelated random walk plus noise models	$\hat{\sigma}^2_{v,1}$	0.5278103	0.010169	−85 126.19	170 262.4	170 305.1
	$\hat{\sigma}^2_{v,2}$	3.0026580	0.065885			
	$\hat{\sigma}^2_{w,11}$	0.7895330	0.013022			
	$\hat{\sigma}_{w,12}$	3.9187366	0.072781			
	$\hat{\sigma}^2_{w,22}$	4.0335776	0.079385			
Random walk plus noise common state process	$\hat{\alpha}_1$	−0.0002721	—	−88 573.57	177 159.1	177 210.4
	$\hat{\alpha}_2$	0.0002314	0.238165			
	$\hat{\beta}$	0.8703086	0.003676			
	$\hat{\sigma}_{v,1}$	0.6931712	0.007793			
	$\hat{\sigma}_{v,2}$	2.8369856	0.011589			
	$\hat{\sigma}_w$	1.0657750	0.008478			

TABLE 3-5. **Parameter Estimates, Standard Errors, and Model Fit Statistics**.
Parameters for series 1 are for minimum temperature. Parameters for series 2 are for maximum temperature.

3.4. DISCUSSION

Table 3-6 reports the forecast accuracy of the three Dynamic Linear Models (DLMs), assessed both in-sample and out-of-sample (December 2024 left aside for testing), and three findings stand out. First, moving to the hold-out month raises errors for the minimum-temperature series (e.g., in the independent random-walk-plus-noise model RMSE rises from 1.289 to 1.546) yet slightly reduces them for the maximum-temperature series (RMSE falls from 2.987 to 2.406). This asymmetric pattern suggests that recent observations carry most of the predictive information for extreme daily highs, so the Kalman filter shrinks T_{max} forecasts more aggressively. Second, errors for T_{max} remain consistently larger than those for T_{min} across every specification—-even the best out-of-sample RMSE for T_{max} (2.406) exceeds its T_{min} counterpart (1.546) by roughly 55 %. The gap mirrors the higher state-innovation variance estimated earlier and the growing prevalence of hot-temperature extremes. Third, while in-sample performance is near-identical, out-of-sample accuracy now marginally favours the independent specification; the common-factor random-walk-plus-noise records out-of-sample RMSEs of 2.517 (T_{min}) and 2.529 (T_{max}). Nonetheless, a preliminary analysis of longer-term forecasts (6 and 12 months) suggested a better performance (That we ought to omit as it does not fall within the scope of this assignment), and this finding further reinforces our intention to proceed in a deeper analysis of the Common Factor Model.

Although not the top performer as we previously anticipated, it still offers the advantage of capturing the shared dynamics in a single interpretable latent state. Figure 3-9 depicts its within-sample fit and the pronounced weight it assigns to the most recent data.

Model	Sample	Tmin			Tmax		
		RMSE	MAE	95% Coverage	RMSE	MAE	95% Coverage
Independent Random Walk plus Noise	Within-sample	1.289	0.959	94.5%	2.987	2.162	93.9%
	Out-of-sample	1.546	1.176	100.0%	2.406	1.988	100.0%
"Seemingly unrelated" Random Walk plus Noise	Within-sample	1.252	0.935	94.9%	2.999	2.190	93.7%
	Out-of-sample	1.580	1.181	100.0%	2.760	2.285	100.0%
Latent-Factor Random Walk plus Noise	Within-sample	1.269	0.944	94.8%	3.264	2.454	94.4%
	Out-of-sample	2.517	2.042	100.0%	2.529	2.050	100.0%

TABLE 3-6. **Forecast performance for each DLM model**.
Out-of-sample forecasts are tested on December 2024 and trained on all the remaining data. Tmin and Tmax are the time series for minimum and maximum temperature respectively. RMSE is the Root Mean Squared Error. MAE is the Mean Absolute Error. 95% coverage is the proportion of times that the true, realized value falls inside the model's 95% credible interval.

These results fit naturally within a climate-change narrative. The persistently higher, more volatile errors for daily highs reflect the increasing frequency of unprecedented heat events, justifying the larger process variance for T_{max}. At the same time, the latent-factor DLM remains appealing: its common state can be linked directly to exogenous climate indicators, facilitating a structural interpretation of the warming trend at only a modest cost in predictive accuracy.

The observed 100 % empirical coverage of the nominal 95 % intervals in December 2024 is simply a consequence of the flat one-step-ahead forecast at the sample boundary, which widens the predictive density.

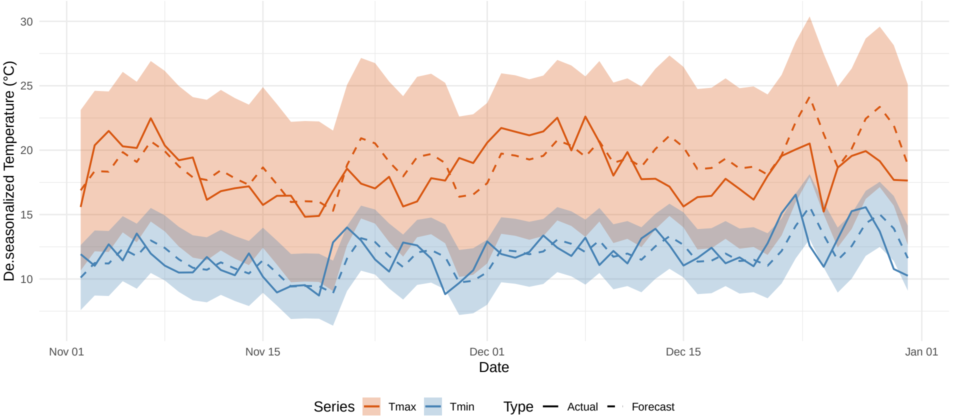
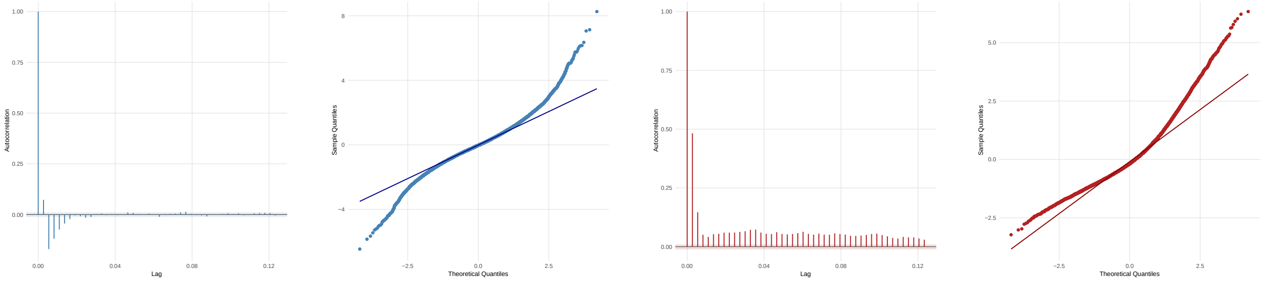


FIGURE 3-9. Prediction within sample for the Random walk plus noise common state process.

We come to the assessment of the assumptions of the Common Factor Random Walk plus noise model. We plot both the QQ-Plot and the ACF functions of the innovations. The first shows some slight deviations from normality though different for the two time series. Innovations for the "Tmin" time series has fatter tails compared to a normal distribution while the ones for "Tmax" are right-skewed. The Jarque–Bera Test unsurprisingly rejects the null of Normality at any conventional level for both time series. While they do not corroborate model assumptions they are informative on the kind of departures from the ideal situation, pointing to extreme situations compatible with a climate change framework. The ACF plot shows that autocorrelation decays quickly after a few lags. While the Ljung–Box test provides evidence for significant autocorrelation, the graph shows that the magnitude is effectively null in magnitude.



ACF of Std. Resid. (Tmin) QQ-Plot of Std. Resid. (Tmin) ACF of Std. Resid. (Tmax) QQ-Plot of Std. Resid. (Tmax)

FIGURE 3-10. Residual diagnostics for the CF model: ACF and QQ plots arranged in a single row.

As the assessment of innovations suggest that further improvements in model specification would be suitable, we estimate also a spatial model that is presented in the following section. Natural expansion of this work would be formally investigating the nature of the common factor by regressing the state stemming from the DLM analysis on potential explanatory covariates. Further work can also handle the limitations in the residuals by relying on Stochastic Volatility DLM (SV-DLM), Common Trend DLM with Correlated Innovations, and Student-t DLM that allow for changing variances while accounting for the previously observed fat tails.

3.5. SPATIAL MODEL

Spatial models allow to incorporate spatial correlation among different stations to improve forecast performance. The more realistic assumption of spatial dependence effectively controls for idiosyncratic shocks while also allowing to give more robust estimates of a potential common factor by relying on different stations. We first use all 13 stations available in the dataset [1] but found worse forecast performance than previous models both wihtin sample and out-of-sample . This is likely due to the incorporation of stations that are too different with respect to seasonal cycles, temperature volatility, and exposure to regional climate shocks. We hence pick those stations whose temperature patterns should closely resemble San Francisco DWTN, namely Paso Robles Muni AP, Baker City AP, and Prescott Love FLD.

We estimate a univariate Spatial-Factor Dynamic Linear Model (DLM) using seasonalized data. To ease computational burden we treat each time series (Tmin and Tmax) separately. The observation process for each station j is such

[1] Kalispell Glacier AP, Little Rock, Martinsburg E W Virginia RGNL, Wilmington Intl AP, Augusta State AP, Houlton AP, Pellston RGNL AP, Mason City Muni AP, Prescott Love FLD, San Francisco DWTN, Laramie AP, Baker City AP, Paso Robles Muni AP

that:

$$Y_{j,t} = F_t\theta_t + \varepsilon_{j,t}, \quad \varepsilon_{j,t} \sim \mathcal{N}\big(0, V_{\text{obs}}\big), \quad j = 1, \ldots, m.$$

where $(Y_{j,t})_{t\geq1}$ is the observed time series (either Tmax or Tmin) for station j, F_t is the common factor matrix and θ_t, the state vector made of L_t, level at day t, B_t, slope (trend) at day t, and Fourier pairs accounting for yearly seasonal components $(S_{1,t}S^*_{1,t})$, $(S_{2,t}S^*_{2,t})$. The F_t contains a $\lambda_j = \exp\Big(-d_j/\rho\Big)$, a fixed loading based on that station's distance d_j from San Francisco. In other words, it is the "spatial" component of the model. The state process follows the following dynamics:

$$\boldsymbol{\theta}_t = G\,\boldsymbol{\theta}_{t-1} + \mathbf{w}_t, \quad \mathbf{w}_t \sim \mathcal{N}\big(\mathbf{0}, W\big).$$

The G matrix is fully presented in the Appendix. Estimand parameters are hence W_{level} V_{obs}, the variance of the process noise and observation-noise variance, respectively. These are assumed to be the same for all stations.

The MLE estimates for the parameters of the "Tmax" time series are $\hat{W_{level}} = 9.54$ and $V_{obs} \hat{=} 1.00$. Again the signal to noise ratio favours reliance on recent observations as these present a much lower noise. We assessed the model by both its in-sample and out-of-sample performance for the San Francisco station. The out-of-sample metrics are obtained by testing the model on December 2024 after being trained on previous data. The within sample accuracy lowers compared to previous models (RMSE = 11.253; MAE = 9.665) while out-of sample (RMSE = 4.131; MAE = 3.665) see a noticeable improvement. Figure 3-11 shows the predicted and actual values for the out-of-sample forecast.

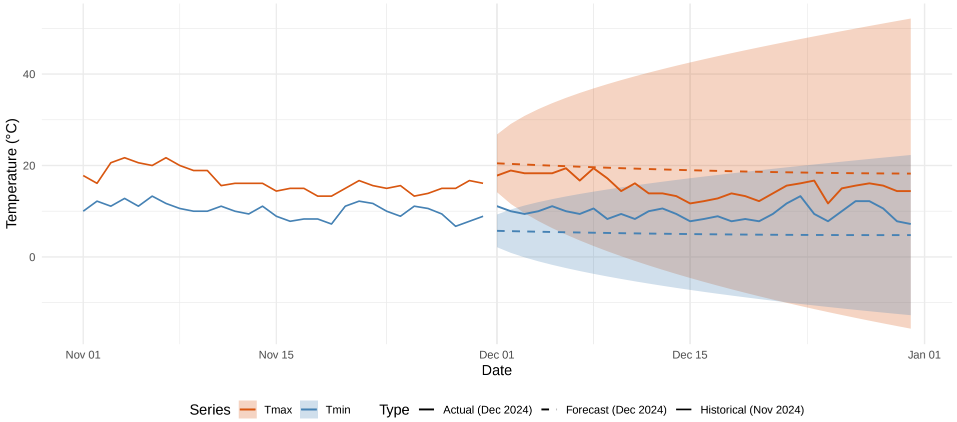


FIGURE 3-11. 30-day forecast of maximum and minimum temperatures for the Spatial DLM model.

Less promising results are obtained for the Tmin series. The estimated parameters by MLE are $\hat{W_{level}} = 2.52$ and $V_{obs} \hat{=} 1.00$. Similar to previous models, the state process for Tmin is less noisy than Tmax, while the signal to noise ratio remains favorable to recent observations. However, notice that this result was obtained by imposing an upward constraint on the variance of the state process to ensure convergence of MLE algorithm and numerical stability. The forecast accuracy worsen both in-sample (RMSE = 3.465, MAE = 2.889) and out of sample (RMSE = 4.786, MAE = 4.548) compared to previous models.

The errors in both time series still suffer from non-normality. The QQ Plot and the Kolmogorov–Smirnov test for Normality document significant departures from white noise errors, while the ACF and the Ljung-Box test give evidence of significant autocorrelation (results shown in the Appendix).

4. CONCLUSIONS

In this report, we relied on two distinct datasets to investigate long term changes in global temperature and short-term local behaviour in the San Francisco Downtown station: we consider the evolution of temperature anomalies reported in NASA GISTEMP and daily temperatures at San Francisco Downtown.

Concerning the former, a two-state Hidden Markov Model on the global series indicates a persistent shift to a warmer, more volatile regime. Among state-space alternatives, an ARIMA(2,1,1) in Random-Walk-plus-Noise form delivered the sharpest one-step forecasts, although residual diagnostics reveal remaining structure that warrants heavier-tailed or heteroscedastic innovations.

For the local record, three multivariate Random-Walk-plus-Noise designs were compared. The common-state specification, augmented with a parsimonious spatial–seasonal term, yielded the lowest out-of-sample errors, yet diagnostics again exposed mild autocorrelation and non-Gaussianity. Overall, the results underscore both the intensification of global temperature variability and the feasibility of short-lead urban forecasts, while highlighting the need for future work on richer error models, covariate-driven dynamics, and broader spatial pooling to ensure decision-relevant reliability.

A. Spatial model

A.1. Model Specification

We observe four stations $y_{j,t}$ at day t $(j = 1, \dots, 4)$ and posit a common 6-dimensional latent state $\boldsymbol{\theta}_t$. The model is

$$\boldsymbol{\theta}_t = G\,\boldsymbol{\theta}_{t-1} + w_t, \quad w_t \sim \mathcal{N}\big(0,\, W\big), \tag{1}$$

$$\mathbf{y}_t = F_t\,\boldsymbol{\theta}_t + v_t, \quad v_t \sim \mathcal{N}\big(0,\, V\,I_4\big), \tag{2}$$

where $\mathbf{y}_t = (y_{1,t}, y_{2,t}, y_{3,t}, y_{4,t})^\top$.

Transition matrix G and W. Write $\boldsymbol{\theta}_t = [\,\theta_t^{(\mathrm{lvl})},\, \theta_t^{(\mathrm{tr})},\, \theta_t^{(\cos 1)},\, \theta_t^{(\sin 1)},\, \theta_t^{(\cos 2)},\, \theta_t^{(\sin 2)}\,]^\top$. Then

$$G = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \cos\frac{2\pi}{365} & \sin\frac{2\pi}{365} & 0 & 0 \\ 0 & 0 & -\sin\frac{2\pi}{365} & \cos\frac{2\pi}{365} & 0 & 0 \\ 0 & 0 & 0 & 0 & \cos\frac{4\pi}{365} & \sin\frac{4\pi}{365} \\ 0 & 0 & 0 & 0 & -\sin\frac{4\pi}{365} & \cos\frac{4\pi}{365} \end{pmatrix}, \qquad W = \mathrm{diag}\big(\sigma^2_{\mathrm{lvl}},\, \sigma^2_{\mathrm{tr}},\, \sigma^2_{\mathrm{sea}},\, \sigma^2_{\mathrm{sea}},\, \sigma^2_{\mathrm{sea}},\, \sigma^2_{\mathrm{sea}}\big).$$

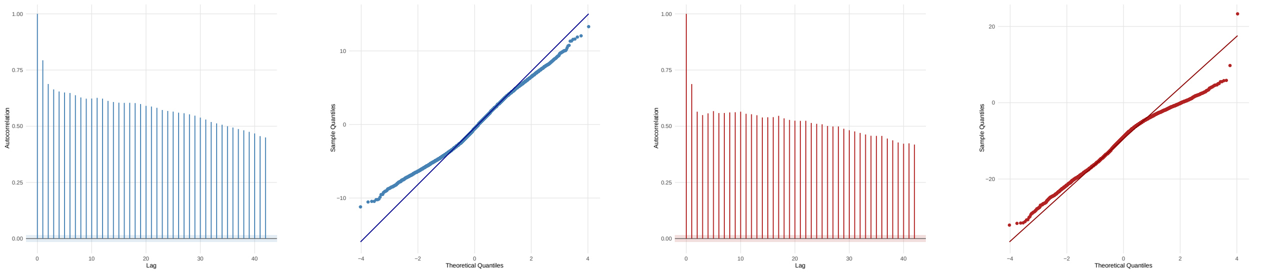
Observation matrix F_t and V. Define the "base" row $F_{\mathrm{base},t} = [\,1,\ 0,\ \cos(2\pi t/365),\ \sin(2\pi t/365),\ \cos(4\pi t/365),\ \sin(4\pi t/365)\,]$. With spatial loadings $\lambda_j = \exp(-d_j/\rho)$ (distance d_j to SF, $\rho = 500$), set

$$F_t = \begin{pmatrix} \lambda_1\, F_{\mathrm{base},t} \\ \lambda_2\, F_{\mathrm{base},t} \\ \lambda_3\, F_{\mathrm{base},t} \\ \lambda_4\, F_{\mathrm{base},t} \end{pmatrix} \in \mathbb{R}^{4\times 6},$$

Initial prior. $\boldsymbol{\theta}_0 \sim \mathcal{N}\big(0,\ 10^6 I_6\big)$.

In summary, parameters $\{\sigma^2_{\mathrm{lvl}},\, \sigma^2_{\mathrm{tr}},\, \sigma^2_{\mathrm{sea}},\, V\}$ are estimated by maximum likelihood.

A.2. Residuals



ACF of Std. Resid. (Tmin) QQ-Plot of Std. Resid. (Tmin) ACF of Std. Resid. (Tmax) QQ-Plot of Std. Resid. (Tmax)

Figure A-12. Residual diagnostics for the CF model: ACF and QQ plots arranged in a single row.

Both graphical evidence and formal test point to failure of standard assumptions of white noise of errors.

The ACF and the Ljung-Box give evidence of significant and relevant autocorrelation (that might be amplified by the spatial nature of the model). This holds for both the time serie for maximum and minimum temperature.

The QQ-Plot and the Kolmogorov-Smirnov test give evidence of significant departures from Normality. We adopted the Kolmogorov-Smirnov test as the Shapiro Wilk is not suitable for large n. Notice that this model give a different picture compared to previous residuals as extreme events seem to be less frequent.