ENSAE PARIS – INSTITUT POLYTECHNIQUE DE PARIS

Research Report — Main

Stochastic and Machine-Learning Temperature Models for Weather-Derivative Pricing

with an extension to natural-gas volatility coupling

July 2026

AUTHOR:

Nayel BENABDESADOK

Discover **Boreas**, my weather-derivatives pricing desk built on statistical and machine-learning models:

huggingface.co/spaces/nayelsdk1/boreas-weather-derivatives

SELF-DIRECTED QUANTITATIVE RESEARCH PROJECT

Abstract

This project studies temperature forecasting models applied to *weather derivatives*, with broader applications to the commodities market, notably *power* and *gas trading*. Two families of models are compared: the Ornstein–Uhlenbeck mean-reverting process, which captures the seasonality, trend and seasonal volatility of temperature, and three learning models (Ridge regression, Random Forest and XGBoost), trained by nested cross-validation on seventy years of data (1950–2019) and then evaluated, at several horizons relevant to a desk, using metrics tailored to weather indices.

The results prove contrasted across forecasting horizons and hard to separate, since weather shocks are strongly correlated between geographically close cities. These models are then applied to the valuation of calls on *Heating Degree Days* (HDD) and confronted with a retrospective economic test over the five test winters (2020–2025). The approach substantially reduces the complexity of classical weather models and offers pricing that is more robust and faster for commodity traders. An extension, coupling temperature anomalies, natural gas storage (Henry Hub) and the volatility of its price, will be published as a complement.

About this project This personal project, without academic supervision, was carried out over four months (from March to early July 2026), alongside my gap-year internship as an AI Engineer at Corpy&Co., in Tokyo. The idea came to me while traveling in the Kyushu archipelago, renowned for its weather shocks: typhoons, the rainy season, intense cold spells and humid summers. This research gradually drew me down the rabbit hole of weather derivatives, with the desire to specialize in them in the future. A wonderful interlude in my life in Tokyo. I warmly thank everyone I met during this period; they will recognize themselves.

Travel memories, Japan 2026: Kumamoto, Beppu, Miyazaki and Kagoshima

Contents

Notations

Table 1: List of symbols used in the report.

Symbol	Description
T_t	Daily average temperature on day t: DAT = (TMIN + TMAX)/2.
Train / test	Estimation and training window (1950–2019) versus out-of-sample test window (2020–2025), fixed once in section 4.
$\theta(t)$	Deterministic seasonal mean: constant + linear trend + Fourier series (section 4).
X_t	Deseasonalised residual; state variable of the OU process.
ε_t	OU innovation.
$\kappa,\ \phi$	Mean-reversion speed and AR(1) coefficient $\phi = e^{-\kappa}$.
$\sigma_X(t)$	Seasonal (day-of-year) volatility of the OU residual.
c	Conversion factor $c = \sqrt{2\kappa/(1-e^{-2\kappa})}$.
W_t	Standard Brownian motion.
$\mathbb{P},\ \mathbb{Q}$	Historical and risk-neutral probability measures.
λ	Market price of weather risk (Girsanov change of measure), (Cao and Wei 2004).
$\mathscr{F}_s$	Information available at the valuation date s (filtration).
$\mathscr{H}$	Winter assumption $\mathbb{P}(T_t > T_{\mathrm{ref}}) \approx 0$ (section 6).
HDD, CDD	Heating / cooling degree days, base 18 °C.
H_n	Cumulative HDD over the n-day contract window $[t_1, t_n]$ (equation (2)).
CAT_n	Cumulative average temperature over $[t_1, t_n]$: $\mathrm{CAT}_n = \sum_{t=t_1}^{t_n} T_t$ (section 5.4).
$\mu_H,\ \sigma_H$	Mean and standard deviation of the simulated cumulative-HDD distribution.
$s,\ n$	Valuation date and contract length in days ($t_n = t_1 + n - 1$).
$K,\ \alpha$	Strike and monetary tick of the contract ($ per degree day).
r, TTM	Risk-free discount rate and time to maturity (TTM $= t_n - s$).
$C_s(n),\ P_s(n)$	Closed-form price of a weather call / put (section 6).
$\mu_s(n),\ \Lambda_s(n)$	Conditional mean and standard deviation of the index H_n under the OU model.
$\Phi,\ \varphi$	CDF and pdf of the standard normal $\mathcal{N}(0,1)$.
X_{t+1}	One-day-ahead residual: the learning target.
x_t	Feature vector ($d = 50$) built at date t (table 4).
MAPE, MAE	Mean absolute percentage error (in Kelvin, tuning scorer) and mean absolute error.
APE	Absolute percentage error on cumulative HDD (evaluation metric).
m	Model index, $m \in$ {clim, burn, OU, Ridge, RF, XGB}.
$K_{c,w}$	Burn strike for city c, winter w: mean cumulative HDD of prior winters (equation (23)).
$\Pi_m(c,w),\ \widehat{\Pi}_m(c,w)$	Contest price under model m and its Monte-Carlo estimator.
$Y_{c,w}$	Realised payoff of the contract once the winter is known.
$e_m(c,w)$	Signed pricing error $\Pi_m(c,w) - Y_{c,w}$.
RMSE_m	Aggregate pricing error of model m over the 35 contracts (equation (25)).

1 Introduction

In the commodities sector, weather has always held a central place, particularly in energy, whether in production (hydroelectricity, solar) or consumption (heating), but also in fields such as agriculture, tourism, or transportation. In the United States, it is estimated that roughly 30% of the economy is directly affected by weather conditions. More broadly, close to a third of the GDP of developed economies is thought to be exposed to climate variability. This exposure keeps growing, driven by the challenges of the energy transition and by the increasing frequency of extreme events.

Yet part of this risk lies beyond the reach of traditional insurance. Insurance compensates for rare and severe losses accompanied by material damage (water damage to an industrial site, a frozen harvest), but it does not cover the economic losses caused by merely unfavorable weather conditions, even though these can be considerable. A mild winter damages nothing, but it collapses heating demand and the revenue of a gas distributor. A cool summer reduces electricity demand. A dry season cuts agricultural output. It is in this context that weather derivatives appeared on the over-the-counter (OTC) market in 1997, against the backdrop of deregulation of the U.S. energy market and anticipation of a strong El Niño episode. As early as 1999, the Chicago Mercantile Exchange (CME) launched the first listed futures and options on climate indices (HDD, CDD, CAT, ...).

A weather derivative is a contract whose payoff depends not on the price of a financial asset, but on a weather variable (temperature, precipitation, wind speed) liable to affect the supply or demand of a commodity. Since the payout is independent of any damage actually incurred, these products eliminate the moral hazard inherent to insurance and constitute tradable, standardized instruments.

This market, considered a niche, is experiencing a marked resurgence. After an initial peak in the mid-2000s (the *Weather Risk Management Association* then estimated the traded notional at close to 45 billion dollars), volumes have been setting new records since 2023, driven by two structural forces: the intensification of climate volatility and the rise of intermittent renewable energies, whose production depends directly on wind and sunshine. The CME thus saw the average volumes of its climate derivatives jump by more than 260 % relative to 2022, with open interest growing by 48 % year over year (CME Group 2024). Weather risk is therefore once again becoming a first-order financial risk for energy and commodity desks.

In this context, having additional information about the weather would offer a twofold competitive advantage to market participants (hedge funds, energy and trading companies, ...): an informational edge on future conditions would allow both better management of the risk associated with this index (and the design of better-suited derivative products) and anticipation of price movements on the related physical and financial markets, such as natural gas or electricity. This advantage, however, runs into a fundamental limit: chaos theory, formulated by Edward Lorenz in 1963 under the name of the "butterfly effect," establishes that beyond roughly a week weather forecasts become intrinsically unstable, and therefore of little reliability for traditional climatology based on physical models. Recent advances in statistics and machine learning nonetheless make it possible to better grasp these dynamics.

The objective of this work is to propose an in-depth analysis of the weather derivatives sector, centered on the behavior of seven U.S. cities, and to establish a benchmark of different temperature forecasting models, particularly over the winter period. In parallel, a dedicated application for pricing these products is developed and makes it possible to compare the performance of the models presented here.

2 Literature review

The modeling and valuation of weather derivatives have long been dominated by statistical approaches to temperature. The founding framework, due to Alaton, Djehiche, and Stillberger (2002), describes daily temperature by a mean-reverting Ornstein-Uhlenbeck process driven by a Brownian motion and centered on a deterministic seasonal mean; the market price of risk, indispensable in an incomplete market, is formalized by Cao and Wei (2004). This foundation was then progressively enriched to better account for the stylized facts of temperature. Benth and Šaltytė Benth (2007) replace the Brownian motion with Lévy processes and refine the modeling of seasonal volatility, while Alfonsi and Vadillo (2023) (co-authored with the École des Ponts and AXA Climate) add stochastic volatility in order to be more cautious in the face of extreme events.

More recently, machine learning has entered this field. Barnor, Kampouridis, and Kanellopoulos (2026) compare a broad range of models, both statistical and machine-learning-based, and single out the best performer with respect to a battery of forecasting criteria, suggesting that these methods can outperform parametric approaches in anticipating the temperature underlying the contracts.

This interest in forecasting is not incidental: its economic value is attested by trading practice. Kuppelwieser and Wozabal (2023) show that an intraday trading strategy on the electricity market, based solely on superior weather-forecast information, generates significant profits. This result invites a deeper exploration of temperature forecasting as a source of advantage, beyond mere actuarial valuation.

This modeling nonetheless runs into two difficulties. On the one hand, the dynamics of temperature are nonlinear and exhibit heavy tails, which Brownian models capture poorly and which justify the extensions mentioned above (stochastic volatility, Lévy processes) as well as specific attention to tail risk (Cheng, Poreddy, and Chen 2025). On the other hand, and more fundamentally, it is delicate to isolate the influence of temperature on commodity prices: this asset class is simultaneously governed by supply, demand, storage levels, and geopolitical tensions, and its valuation relies on full-fledged commodity-price models (Schwartz and Smith 2000). A robust link does exist between temperature, storage, and natural-gas futures prices, but disentangling the purely meteorological signal from the rest remains difficult.

This is precisely what motivates the present work: the informational gain of such forecasts is all the more valuable because it is difficult to obtain and could capture dynamics not yet exploited by the market. Whether for investment banks seeking to better manage their positions in certain commodities, for reinsurance companies seeking to better price and hedge their climate-exposed commitments (parametric hedges, diversification of their risk portfolio), or for hedge funds that use weather models to generate trading signals, the value of additional information on the weather is manifest. It is this paradigm, still little explored, that we humbly propose to investigate through the modeling of temperature to price weather derivatives and possibly propose a coupling with gas.

3 Data Analysis

This section presents the temperature data used and the processing applied to it. An exploratory analysis then draws out their main characteristics and highlights the difficulties that temperature modeling poses.

3.1 Choice of Cities and Data Used

Although weather derivatives are now traded in many regions of the world, 13 of the 18 cities covered by CME contracts (detailed in appendix A) are located in the United States.

We first selected four cities to study their behavior: Chicago O'Hare, New York JFK, Boston Logan, and Minneapolis–St. Paul.

The choice of these cities was primarily driven by their gas consumption: they are located in the Midwest and Northeast of the United States, the country's two main regions of winter natural gas consumption. They are also among those for which the CME has offered standardized weather contracts since 1999.

In order to enrich the study and observe contrasting climatic dynamics, we also added three cities that do not appear in the list covered by the CME: Los Angeles, Seattle, and Dallas.

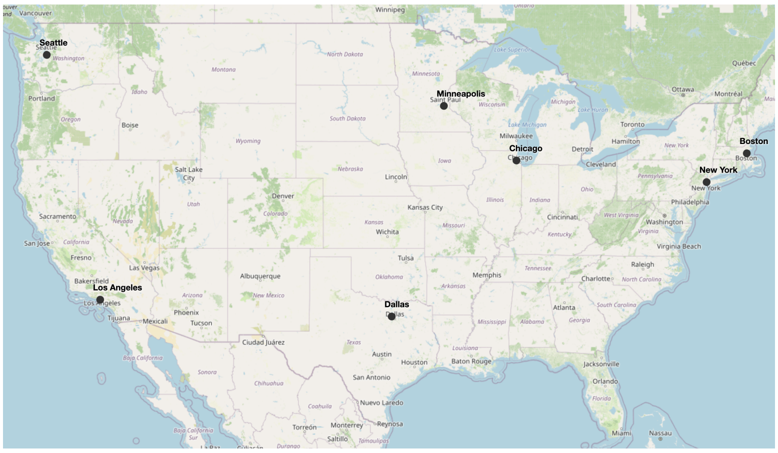

Figure 1: The seven cities selected for the study, represented by black dots.

The data used comes from the National Oceanic and Atmospheric Administration (NOAA), which provides daily weather observations for the seven selected cities over the period 1950–2025.

For each day t, the minimum ($T_{\min,t}$), maximum ($T_{\max,t}$), and mean (T_t) temperatures are available. Our analysis focuses on the daily mean temperature. When the latter is missing, it is reconstructed from the extreme temperatures according to $T_t = \frac{1}{2}\left(T_{\max,t} + T_{\min,t}\right)$.

All temperatures are expressed in degrees Celsius (°C). Since the raw NOAA data is provided in degrees Fahrenheit (°F), a conversion is applied to each observation according to $T_t^{(°C)} = \frac{5}{9}\left(T_t^{(°F)} - 32\right)$. The series are also cleaned of non-standard days (in particular February 29), so that each day t has the same number of observations over the study period. Finally, each date is associated with its day number in the year (doy $\in \{1, \ldots, 365\}$), and two seasonal indicators are defined: *Winter* (from November to March) and *Summer* (the other months). Strictly speaking, these windows should be adapted to the latitude of each city. We keep them identical for all seven cities in the context of this study.

3.2 Descriptive Analysis

As one might expect, temperature is governed above all by a cyclical physical dynamic, marked by the alternation between summer and winter and by transition months in spring and autumn.

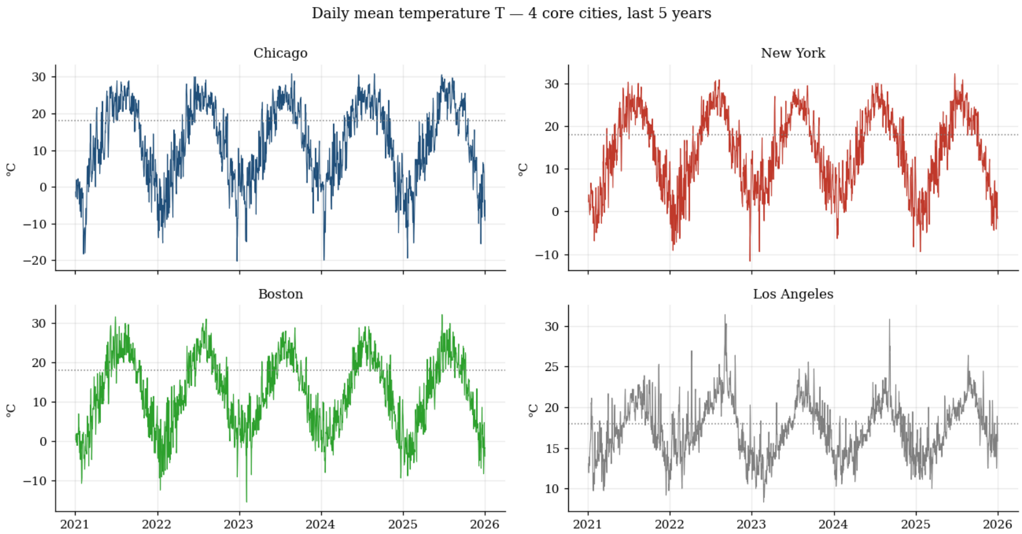

Figure 2: Daily mean temperature T in the four main cities, over the last five years.

Beyond this deterministic seasonal trend, examination of the recent history reveals a random component whose intensity varies according to the time of year and the location. Geography plays a decisive role here. Chicago and Minneapolis, in the heart of the continental Midwest, are exposed to cold air masses coming from Canada: their winters are harsh and highly volatile (the winter of 2023 thus experienced episodes of extreme cold across the Eastern United States), while their summers are noticeably more stable. New York and Boston, open to the Atlantic Ocean, benefit from a moderating maritime effect but remain subject to oceanic uncertainties. Los Angeles and Seattle, on the Pacific coast, exhibit the smallest thermal amplitudes, even though Los Angeles occasionally experiences episodes of intense heat or unexpected cold.

Thus, the climatic events to which cities are exposed depend essentially on their geographic location, which also translates into strong correlations between cities in the same region.

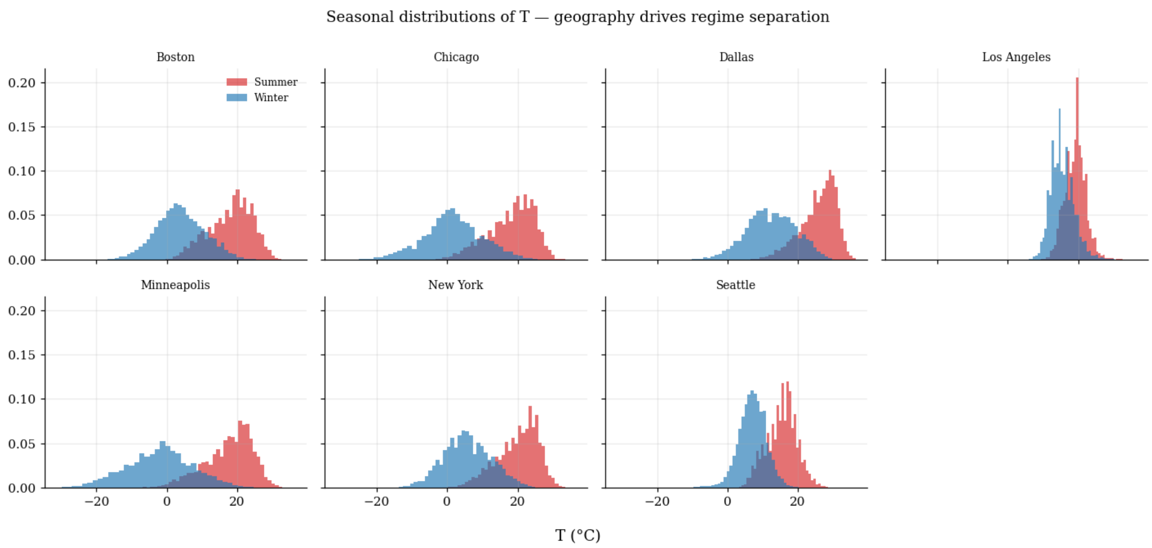

Figure 3: Seasonal distributions (winter/summer) of temperature T for the seven cities.

These differences can be read directly from the distribution of temperatures, examined separately in winter and summer. In winter, the four cities of the Northeast and Midwest exhibit very spread-out distributions: their high dispersion (winter standard deviation on the order of 8 to 9 °C in Chicago and Minneapolis) covers a wide range of temperatures, from severe cold spells to thaws. Their shape nonetheless remains close to Gaussian, nearly symmetric and without excess kurtosis. Cold seasons are thus distributed fairly uniformly between mild and harsh episodes. In summer, these same cities display a marked negative skewness, a sign of a heavier left tail, that is, of occasional episodes of cool summers.

Los Angeles stands out clearly: to its very small amplitude (the smallest standard deviation in the sample) are added a positive skewness and the highest kurtosis, reflecting a distribution tightly clustered around its mean but punctuated by episodes of intense heat. Seattle, for its part, exhibits a leptokurtic winter, with a pronounced peak accompanied by a few extreme values. A stricter breakdown, limited to the heart of winter (December to February) and summer (June to August), is reported in appendix B: it separates the two regimes more sharply but does not alter these conclusions. These observations are quantified in table 2.

	Winter				Summer			
City	Mean	Std. dev.	Skew.	Kurt.	Mean	Std. dev.	Skew.	Kurt.
Chicago O'Hare	1.61	8.25	−0.05	+0.01	18.22	6.47	−0.62	−0.15
New York JFK	5.74	6.81	+0.08	−0.28	20.13	5.83	−0.62	−0.16
Boston Logan	3.98	6.80	+0.08	−0.21	18.08	6.14	−0.45	−0.43
Minneapolis–St. Paul	−2.14	9.29	−0.07	−0.21	17.61	6.72	−0.68	+0.07
Dallas	12.31	6.80	−0.12	−0.29	25.88	5.01	−0.83	+0.37
Los Angeles	15.46	3.26	+0.53	+0.60	19.19	2.88	+0.13	+0.49
Seattle	7.06	3.89	−0.28	+0.90	15.46	4.22	−0.11	−0.26

Table 2: Descriptive statistics of the daily mean temperature T by season (winter from November to March, summer from April to October), over the period 1950–2025.

3.3 Trend and Volatility of Temperature

The ten-year moving average of temperature (figure 4) highlights the underlying trend, once the seasonal and inter-annual fluctuations are smoothed out. The seven cities exhibit a clear upward drift, particularly pronounced over the last two decades. The slope of this trend is on the order of +0.2 °C per decade, strongest in Minneapolis and Dallas (about +0.27 °C) and more moderate on the Atlantic coast (about +0.15 °C in Boston). Several factors may be at the origin of this drift, foremost among them the urban heat island effect, linked to the growing industrialization of cities and their surroundings, and the increase in greenhouse gas emissions since the 1950s, whose effect on surface temperatures is now widely documented.

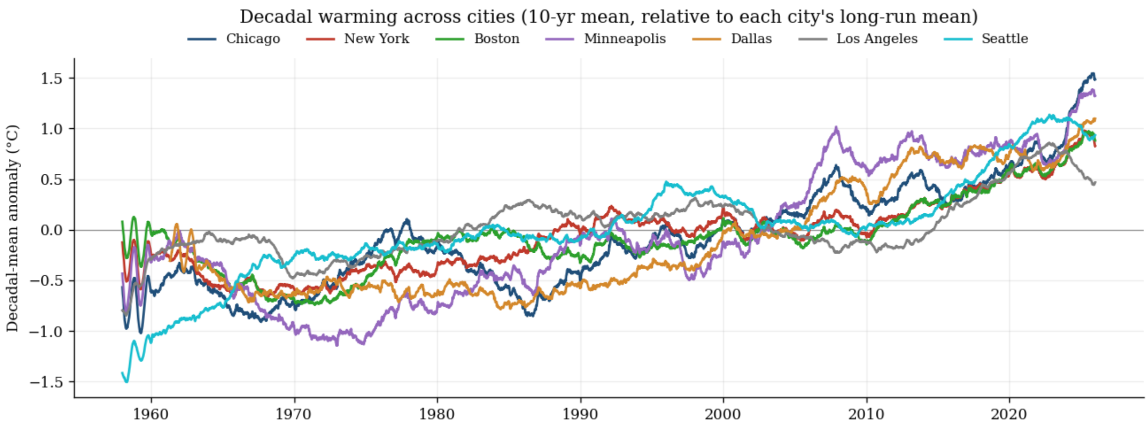

Figure 4: Anomaly of the ten-year moving average of temperature (°C)

Naturally, one would be tempted to ask whether this rise in temperatures also influences the occurrence of extreme events. This warming is not accompanied by an increase in long-term variance: over the entire period, the variance of the anomalies has remained stable for all seven cities (appendix D). Warming therefore shifts the level of temperature without increasing its variability, which will justify some of our assumptions about volatility.

While the variance remains stable from one decade to the next, it does, however, vary strongly over the course of the year. figure 5 shows the 30-day moving variance of temperature over the last fifteen years. It reveals a clearly cyclical behavior: for each city, calm periods, corresponding to summers in most cases, alternate with markedly more volatile periods, in particular the winters of the continental cities. In Chicago and Minneapolis, the winter variance thus regularly exceeds 50 °C^2, compared with a few units at the height of summer, while the Atlantic cities (New York, Boston) exhibit more contained oscillations. The three remaining cities are reported in appendix C.

The volatility of temperature is therefore neither constant nor purely random: it follows a regular seasonal cycle, which makes it modelable. It is precisely this risk, concentrated in the winter months for the regions of high gas consumption, that market participants seek to hedge by means of weather derivatives.

Figure 5: 30-day moving variance of temperature (°C^2), over the last fifteen years, for the four main cities.

3.4 Summary

The descriptive analysis brings out three properties of temperature that will guide its modeling. First, temperature exhibits a strong deterministic structure, both in its level and in its variance: its level follows a regular seasonal cycle, onto which a slow underlying drift is grafted, and its volatility likewise follows a clear annual cycle. Second, warming is slow but clearly perceptible, on the order of +0.2 °C per decade on the mean and visible on the moving average of all the cities. It calls for a trend component in the mean level of temperature. Finally, volatility is cyclical over the course of the year, but it has not been altered by the rise in temperatures: its long-term variance has remained stable. Its seasonality must therefore be modeled, while its stationarity over time can be assumed.

3.5 HDD and CDD, Climatic Underlyings

We define here the best-known climatic indices on which weather derivatives are based and which will be used in what follows.

As a reminder, a weather derivative is a financial contract whose underlying is not a classic financial asset (stock, interest rate) but a climatic index (temperature, wind, precipitation). The most widespread indices are the *Heating Degree Day* (HDD) and the *Cooling Degree Day* (CDD).

Let T_t be the mean temperature of day t and T_{ref} a reference temperature. We adopt $T_{\text{ref}} = 18\,°\text{C}$ as the reference temperature, used in °F in CME contracts. We define:

$$\text{HDD}_t = \max(T_{\text{ref}} - T_t,\ 0) \qquad \text{CDD}_t = \max(T_t - T_{\text{ref}},\ 0) \tag{1}$$

Intuitively, the HDD measures the excess of cold relative to the reference temperature: if $\text{HDD}_t > 0$, then $T_t < T_{\text{ref}}$, the temperature is below the threshold. Symmetrically, the CDD measures the excess of heat.

These indices are cumulated over the contract window $[t_1, t_n]$ (with n the number of days); the underlying is therefore:

$$H_n = \sum_{t=t_1}^{t_n} \text{HDD}_t \qquad \text{CDD}_{t_1}^{t_n} = \sum_{t=t_1}^{t_n} \text{CDD}_t \tag{2}$$

As an example, consider a real-estate portfolio manager whose heating costs soar during harsh winters. To hedge against this, he buys an HDD *call* with strike $K = 100$. At the end of a colder-than-expected winter, the index reaches $H_n = 120$.

Unlike classic options, where exercise may give rise to delivery of the underlying, a climatic index is not deliverable: the contract is **cash-settled** and directly pays out a monetary amount $\text{Payoff} = \alpha \max(H_n - K,\ 0)$, where α is the tick (the amount paid per degree day). The seller therefore pays 20α \$ to the buyer here.

Thus, when temperatures are lower than normal, the HDD increases and the contract generates a payment that offsets the additional heating cost: this mechanism allows a participant exposed to cold to hedge against unfavorable winter conditions.

4 Modelling Temperature via Ornstein–Uhlenbeck Processes

In this section, we introduce the first model of our work for temperature, based on the Ornstein–Uhlenbeck process (hereafter denoted OU). It will also allow us to establish the essential notation for the Machine Learning part.

The data split is fixed here: the data from January 1, 1950 to December 31, 2019 are used to estimate the parameters of the OU model and to train the learning models (section 5). The data from January 1, 2020 to the end of 2025 (never used during estimation or tuning) are kept as a test set, touched only once by the final model in order to assess its ability to generalise out-of-sample.

4.1 Specification

As we saw in the previous part, temperature is governed by two distinct components:

- a deterministic component, marked by strong seasonality and a dependence on the time of year. It can be modelled by a sinusoidal function of the form $\sin(\omega t + \varphi)$. To this is added the effect of global warming, which cannot be neglected. We will denote it $\theta(t)$;
- a stochastic component that reverts fairly quickly toward the deterministic mean (deviations dissipate within a few days depending on the city) and whose variance evolves with the seasons. We will denote this residual X_t.

We then model the temperature of a city on day t as the sum of these two components:

$$T_t = \theta(t) + X_t. \tag{3}$$

Let us see how to determine these two terms in order to reconstruct the dynamics of any city.

4.2 Deterministic component $\theta(t)$

Given the two components of θ, we proceed as follows:

- the seasonality is modelled by a Fourier series, a linear combination of harmonics $a_k \sin(k\omega t)$ and $b_k \cos(k\omega t)$, with $\omega = 2\pi/365.25$ (the average length of the year), $k \in \{1, \dots, N\}$ the order of the series and the coefficients $(a_k, b_k) \in \mathbb{R}^2$;
- the warming is represented by a linear trend $A + Bt$, with $(A, B) \in \mathbb{R}^2$. Other forms could have been explored (for example a quadratic trend to account for an acceleration of the warming), but we retain this approach.

The deterministic component is thus written:

$$\theta(t) = A + Bt + \sum_{k=1}^{N} \big(a_k \sin(k\omega t) + b_k \cos(k\omega t)\big). \tag{4}$$

To estimate the coefficients and determine the order N specific to each city, we build a design matrix $\mathbb{X}$ containing the explanatory variables, and we define the column vector of coefficients β:

$$\mathbb{X}_t = \begin{bmatrix} 1 & t & \sin(\omega t) & \cos(\omega t) & \cdots & \sin(N\omega t) & \cos(N\omega t) \end{bmatrix}, \quad \beta \quad = \begin{bmatrix} A & B & a_1 & b_1 & \cdots & a_N & b_N \end{bmatrix}^\top$$

Over the set of observations, the deterministic model is written $\theta = \mathbb{X}\beta$. Denoting x_t the covariate vector of day t (the t-th row of $\mathbb{X}$) and Y_t the observed temperature, the parameters are estimated by ordinary least squares (OLS), which minimise the sum of squared residuals. The estimator is then written as a sum:

$$\hat{\beta} = \arg\min_{\beta} \sum_t \left(Y_t - x_t^\top \beta\right)^2 = \left(\sum_t x_t x_t^\top\right)^{-1} \sum_t x_t Y_t.$$

By the law of large numbers, as the number of observations n grows, $\frac{1}{n}\sum_t x_t x_t^\top$ and $\frac{1}{n}\sum_t x_t Y_t$ converge to their respective expectations, so that $\hat{\beta}$ converges in probability to the population coefficient vector.

To determine the order N of the Fourier series, we seek the one that best fits the data without overfitting. To this end we resort to the Akaike and Bayesian information criteria (AIC and BIC):

$$\mathrm{AIC} = -2\log(\hat{L}_n) + 2\tilde{k} \qquad \mathrm{BIC} = -2\log(\hat{L}_n) + \tilde{k}\log(n),$$

where n is the number of temperature observations, $\tilde{k}$ the number of parameters ($2N$ harmonics plus the two constants A, B, i.e. $\tilde{k} = 2(N+1)$) and $\hat{L}_n$ the maximised likelihood of the model.

For both criteria, the first term rewards goodness of fit while the second penalises the complexity of the model. With $n \approx 25\,550$ estimation days, $\log(n) \approx 10.1$: the BIC therefore penalises the addition of parameters more heavily than the AIC. We retain the order N that minimises one of these criteria, as illustrated in figure 6: the gap to the minimum stabilises around $N = 3$. The detail of the AIC/BIC values for the seven cities and each order is reported in appendix F, and the estimated coefficients of $\theta(t)$ by city in appendix G.

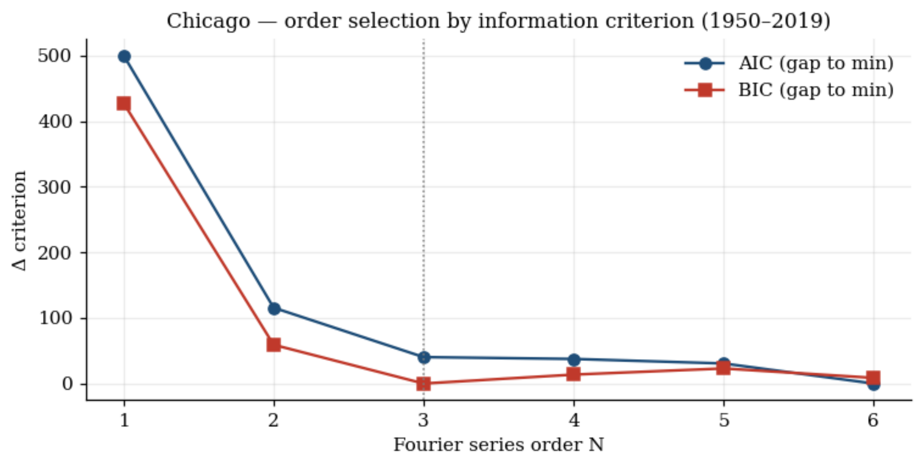

Figure 6: Selection of the order N of the Fourier series for Chicago by AIC and BIC (gap to the minimum, 1950–2019 window).

To evaluate the likelihood of the model, we observe that the residuals $X_t = T_t - \theta(t)$ follow a normal distribution. Their empirical distribution is examined in detail during the study of the stochastic component X_t.

After computation (detailed in appendix E), the maximum likelihood estimator of the variance is:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n} X_t^2.$$

By reinjecting $\hat{\sigma}^2$ into the maximised log-likelihood, the constant terms group together and we obtain the simplification:

$$\boxed{-2\log(\hat{L}_n) = n\log(\hat{\sigma}^2) + \text{constant}} \tag{5}$$

Examining the estimated coefficients city by city (table G.1), we find that the slopes B are consistent with what the descriptive analysis had highlighted (figure 4): a warming on the order of 0.10 to 0.30 °C per decade. The coefficients of the first harmonic, a_1 and b_1, largely dominate the higher orders. Some works in the literature indeed limit themselves to $N = 1$, which would not substantially change what follows. figure 7 superimposes the estimated deterministic component $\theta(t)$ on the observed temperature over a few years: the fit faithfully reproduces the seasonal cycle.

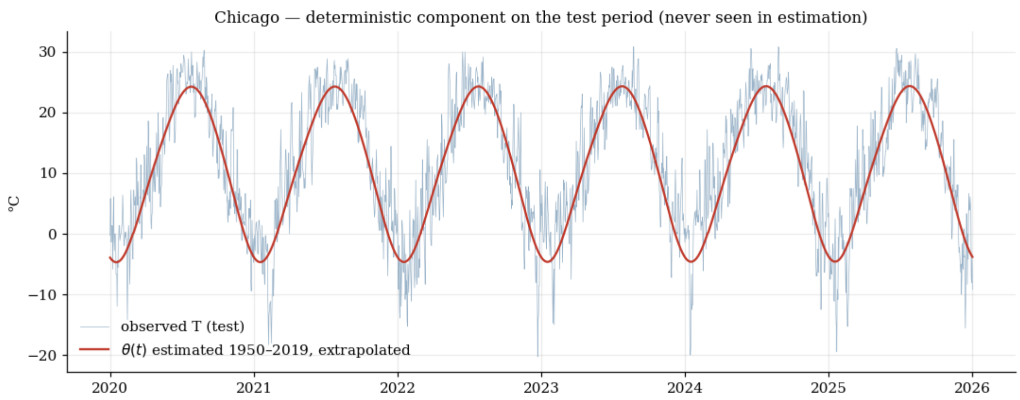

Figure 7: Estimated deterministic component $\theta(t)$ (light line) superimposed on the observed temperature, over the test period.

4.3 Stochastic component X_t

From there, we turn to the stochastic component, the residual $X_t = T_t - \theta(t)$. figure 8 presents an example for Chicago.

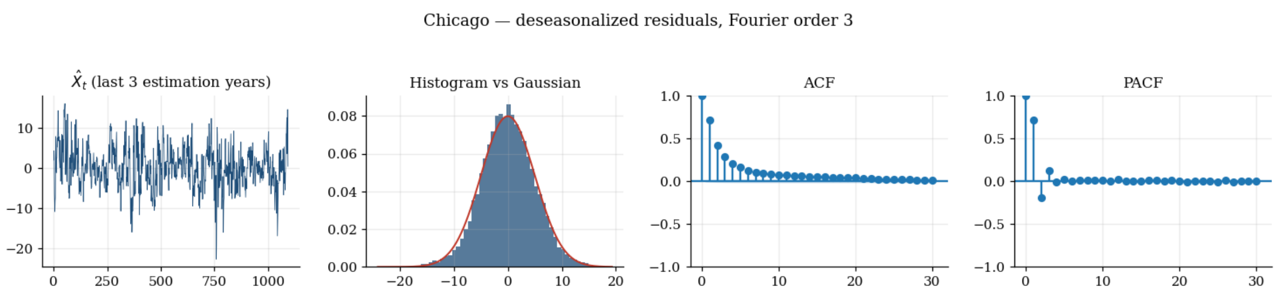

Figure 8: Deseasonalized residuals X_t of Chicago (Fourier order $N = 3$): time series, histogram compared to a normal distribution, ACF and PACF.

As announced, X_t is a series centred at zero, with a distribution close to a Gaussian (second panel), whose volatility depends on the time of year and on the city. The literature very widely models this component by an Ornstein–Uhlenbeck mean-reverting process, one of the most used processes in finance and physics, for example to describe the position of a particle subject to a restoring force (Langevin equation) or short interest rates (Vasicek model). In our case, the model is written:

$$dX_t = -\kappa X_t \, dt + \sigma_X(t) \, dW_t, \tag{6}$$

where $\kappa > 0$ is the speed of mean reversion: X_t converges in expectation to 0, which makes T_t converge to $\theta(t)$. The term $\sigma_X(t)$ is the volatility of the process (not constant in our case, in view of the previous analysis) and W_t a standard Brownian motion.

The challenge here is to determine the value of κ by solving equation equation (6), then to assign the model a volatility consistent with the time of year.

4.3.1 Solving the OU process and estimating κ

Solving equation equation (6) (detailed in appendix H) leads, for $s < t$, to:

$$\boxed{X_t = e^{-\kappa(t-s)}\,X_s + \int_s^t e^{-\kappa(t-u)}\,\sigma_X(u)\,\mathrm{d}W_u} \tag{7}$$

The same approach applied to $T_t = \theta(t) + X_t$, whose dynamics is $\mathrm{d}T_t = \big(\theta'(t) + \kappa(\theta(t) - T_t)\big)\mathrm{d}t + \sigma_X(t)\,\mathrm{d}W_t$, gives:

$$T_t = \theta(t) + \big(T_s - \theta(s)\big)e^{-\kappa(t-s)} + \int_s^t e^{-\kappa(t-u)}\,\sigma_X(u)\,\mathrm{d}W_u. \tag{8}$$

In our framework, we seek to forecast the temperature of the next day: we place ourselves in the filtration $\mathscr{F}_{t-1}$ and take $s = t-1$. The second term is a stochastic integral that encodes the unpredictability of the model. We denote it $\varepsilon_t = \int_{t-1}^t e^{-\kappa(t-u)}\,\sigma_X(u)\,\mathrm{d}W_u$. The process then simplifies to:

$$X_t = e^{-\kappa}\,X_{t-1} + \varepsilon_t, \tag{9}$$

which is exactly the relation of an AR(1) process. The empirical analysis confirms this (4^{th} plot of figure 8): the PACF of X_t exhibits a marked spike at the first lag followed by noise, while the ACF decays slowly, the signature of an AR(1).

To estimate κ, it therefore suffices to estimate the coefficient ϕ of the AR(1), then to set $\hat{\kappa} = -\ln(\hat{\phi})$. table 3 reports these values for the seven cities.

Table 3: Mean-reversion speed $\hat{\kappa} = -\ln\hat{\phi}$ by city (1950–2019 estimation), AR(1) coefficient $\hat{\phi}$ and half-life $\ln 2/\hat{\kappa}$ (in days).

City	$\hat{\kappa}$	$\hat{\phi}$	**Half-life (days)**
Chicago	0.332	0.718	2.1
New York	0.391	0.676	1.8
Boston	0.460	0.631	1.5
Minneapolis	0.293	0.746	2.4
Dallas	0.361	0.697	1.9
Los Angeles	0.270	0.763	2.6
Seattle	0.291	0.747	2.4

Reversion is fast everywhere (half-life of 1.5 to 2.6 days): a deviation from the seasonal norm dissipates within about ten days. It is fastest in Boston ($\hat{\kappa} = 0.46$) and slowest in Los Angeles ($\hat{\kappa} = 0.27$), with the coastal cities displaying slightly higher persistence.

4.3.2 Estimating $\sigma_X(t)$

As seen previously, the volatility is not constant and depends strongly on the time of year (figure 5). To estimate $\sigma_X(t)$, we exploit the relation defining the innovation $\hat{\varepsilon}_t$:

$$\hat{\varepsilon}_t = X_t - e^{-\hat{\kappa}}\,X_{t-1} = \int_{t-1}^t e^{-\hat{\kappa}(t-u)}\,\sigma_X(u)\,\mathrm{d}W_u.$$

Since σ_X is assumed deterministic, the Itô isometry gives:

$$\mathrm{Var}(\hat{\varepsilon}_t) = \int_{t-1}^{t} e^{-2\hat{\kappa}(t-u)}\, \sigma_X^2(u)\, \mathrm{d}u.$$

Assuming $\sigma_X(u)$ constant over $[t-1,t]$, we obtain $\mathrm{Var}(\hat{\varepsilon}_t) = \sigma_X^2(t)\,\frac{1-e^{-2\hat{\kappa}}}{2\hat{\kappa}}$, hence the estimator:

$$\boxed{\hat{\sigma}_X(t) = \sqrt{\frac{2\hat{\kappa}}{1-e^{-2\hat{\kappa}}}} \cdot \sqrt{\frac{1}{W}\sum_{u=t-W/2}^{t+W/2} \hat{\varepsilon}_u^2}} \tag{10}$$

where the sum runs over a sliding window of $W = 30$ days centred at t. figure 9 contrasts the innovations $\hat{\varepsilon}_t$ and the estimated volatility $\hat{\sigma}_X(t)$ for Chicago: the amplitude of the innovations (left) is much stronger in winter, which $\hat{\sigma}_X(t)$ (right) translates into a marked seasonal cycle.

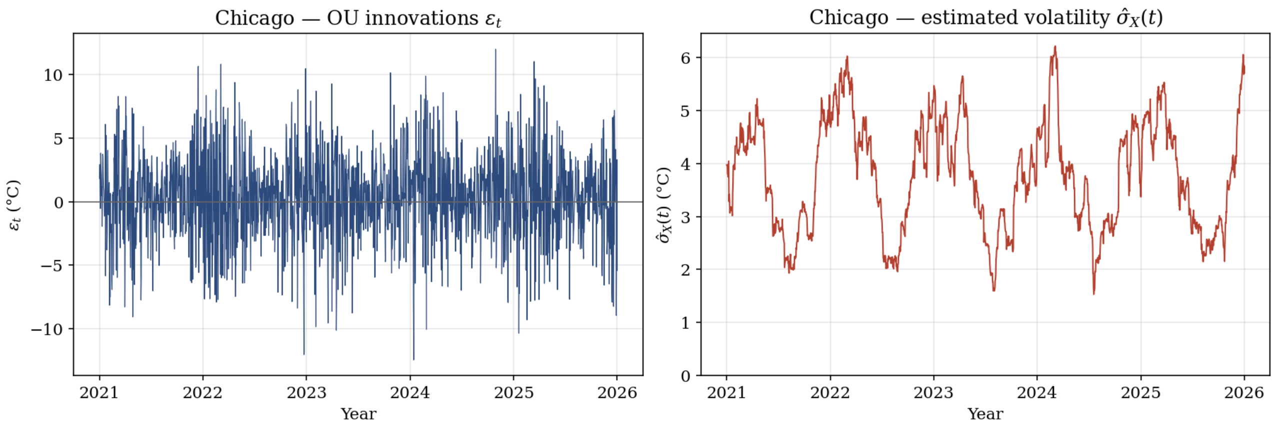

Figure 9: Chicago, 2021–2025: innovations $\hat{\varepsilon}_t$ of the OU process (left) and estimated volatility $\hat{\sigma}_X(t)$ from a 30-day sliding window (right).

To simplify our study and since no warming effect is to be noted, we will assume that the volatility is the same on each day of the year ($\sigma_X(t) = \sigma_X(t+365)$); this reduces our study to finding, for each city, a volatility over the 365 days of the year. In appendix C we recover this seasonal volatility profile for the additional cities. For our 4 main cities, the volatility in winter is much stronger than in summer, whereas Seattle and Los Angeles have a relatively constant volatility over the year.

Several methods for estimating this volatility have been explored.

The first assumes $\sigma_X(t)$ constant over each month: $\sigma_X(t) = \sigma_m$ for t belonging to month m, the twelve monthly levels $(\sigma_1, \ldots, \sigma_{12})$ being estimated by the empirical mean of the volatilities observed over each month.

The second models $\sigma_X(t)$ as a sum of harmonics, in the manner of a Fourier series:

$$\sigma_X(t) = \sigma_0 + \sum_{k=1}^{N_\sigma} \left(a_k \cos(k\omega t) + b_k \sin(k\omega t)\right).$$

The optimal order N_σ is determined as for θ, by retaining the one that minimises the BIC criterion.

A third approach, using B-splines, was also implemented in the notebook dedicated to volatility, but its results, very close to those of the Fourier representation, did not prove sufficiently convincing. One could also have modelled the volatility itself as an Ornstein–Uhlenbeck process, of the form $d\sigma_X(t) = -\kappa_\sigma\big(\mu - \sigma_X(t)\big)\,dt + \tilde{\sigma}_X(t)\,dW_t^\sigma$. We nevertheless keep the Fourier approach, which is simpler and more robust, the volatility not being the core of our study.

figure 10 compares these approaches on the Chicago profile: the Fourier series (order $K = 4$) hugs the seasonal cycle without overfitting the noise, and it is the one we retain.

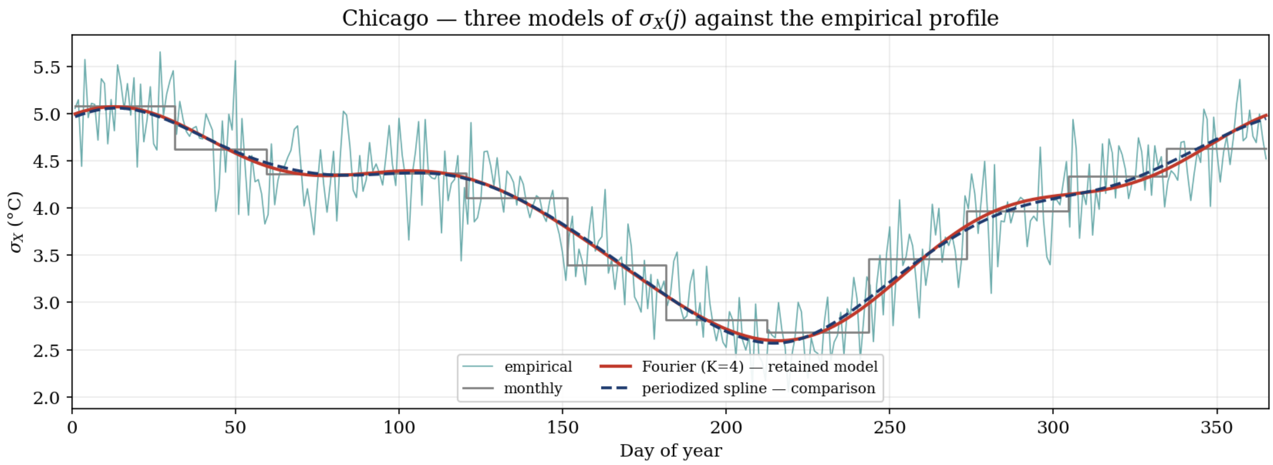

Figure 10: Chicago: volatility profile σ_X by day of the year, empirical profile fitted by monthly steps, Fourier series (order $K = 4$) and periodic spline.

Once θ, κ and σ_X are estimated, the model is fully specified and one can simulate temperature trajectories. With the equation 6, the continuous-time process is: $dT_t = \left(\theta'_t + \kappa(\theta_t - T_t)\right) dt + \sigma_t \, dW_t$ where $\theta'_t = \frac{d\theta}{dt}$ is the analytical derivative of the seasonal function. The Euler-Maruyama discretization with step $\Delta t = 1$ day gives:

$$T_{t+1} = T_t + \left(\theta'_t + \kappa(\theta_t - T_t)\right)\Delta t + \sigma_t \sqrt{\Delta t} \cdot Z_t, \qquad Z_t \overset{iid}{\sim} \mathcal{N}(0,1)$$

figure 11 illustrates this for the winter of 2022/23 in Chicago. On the left, a few simulated scenarios oscillate around the climatic mean θ and bracket the temperature actually observed. On the right, by accumulating the HDD of each scenario over the contract window, one obtains over $N = 10000$ draws the distribution of the index. It is this passage from the temperature trajectory to the distribution of the climatic index that will be at the heart of the valuation in 6.

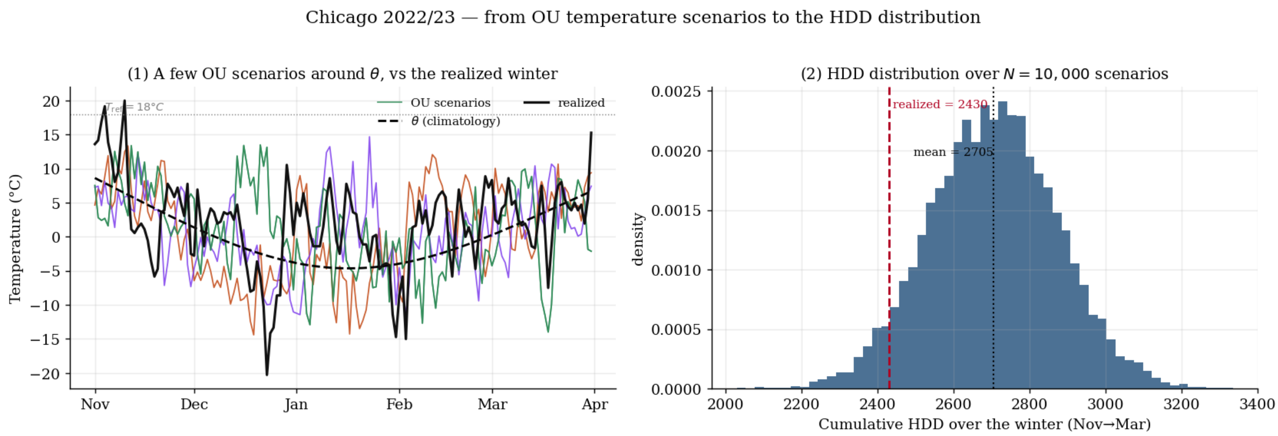

Figure 11: Chicago, winter 2022/23. On the left, three temperature scenarios simulated by the OU (exact AR(1) scheme) and the climatic mean θ, with the realised temperature. On the right, distribution of the cumulative HDD over the contract window ($N = 10\,000$ scenarios), with the mean and the realised HDD.

5 Statistical-Learning Temperature Forecasting

To complement the stochastic approach, we assess whether statistical-learning methods manage to capture the nonlinear relationships that govern temperature.

Rather than predicting the temperature T_t directly, we take as our target the deseasonalized residual X_t: since the deterministic component $\theta(t)$ is already known and well understood (seasonality, trend), it is on X_t that exploitable nonlinear dependencies may remain. The problem is as follows:

given the past residuals $X_1, \dots, X_t$ and a vector of covariates $x_t \in \mathbb{R}^d$ (d features), predict the next-day residual X_{t+1}.

We thus reduce the problem to a supervised-learning problem on the sample $\{(x_i, y_i)\}_{i=1}^N$, where the target is $y_i = X_{t+1}$ and x_i the feature vector built at date t.

5.1 Data Cleaning & Feature Engineering

When a temperature is missing, removing it would be a mistake: since the series is strongly correlated, doing so would break the continuity of the lags. We therefore reconstruct the missing value by combining climatology (same day in past years) and local dynamics (neighboring days), taking the average of these 2 dynamics:

$$\widehat{T}_t = \tfrac{1}{2}\left(\overline{T}_t^{\text{an}} + \overline{T}_t^{\text{jour}}\right), \qquad \overline{T}_t^{\text{an}} = \frac{1}{N}\sum_{\text{an}} T_{t,\text{an}}, \qquad \overline{T}_t^{\text{jour}} = \frac{1}{14}\sum_{i=1}^{7}\left(T_{t-i} + T_{t+i}\right), \tag{11}$$

that is, the average between the value of the same day over past years ($\overline{T}_t^{\text{an}}$) and a sliding window of ±7 days ($\overline{T}_t^{\text{jour}}$).

Next comes the construction of the features. We build explanatory variables that inform the model about seasonality, recent dynamics and the level of volatility, while avoiding *data leakage* (the day-of-year statistics are computed only on the training set). Lacking any source other than temperatures, these $d = 50$ features are built by hand. table 4 groups them by block.

Table 4: The 50 features, grouped by block (target: X_{t+1}).

Block	#	Variables and role
Day index	2	d_t, d_{t+1}: position in the year.
Recent lags	7	$X_{t-1}, \dots, X_{t-7}$: short-term dynamics.
Previous years	23	$X_{t-k\,\text{an}}$ (k=1..5), values around the same calendar day last year (±7 days) and of the target day (k=2..5 years)
Day climatology	2	doy_mean, doy_var: average level and volatility of the day (train only).
Differences	15	$\Delta X(k)$ (k=1..7 days), annual ΔX (k=1..5 years) and their min/max/mean: local trend.
Deterministic component	1	θ_t: seasonality and trend.

As expected, some of these variables are strongly correlated, some being near-linear combinations of the others: for example the differences $\Delta X(k)$ and their mean form a highly correlated block, as do last year's values around the same calendar day. The correlation matrix, in which these collinear blocks clearly stand out, is reported in figure J.1.

Other features could also have been considered, some of which were the weather forecasts provided by NOAA (revised forecasts, recomputed from 2000 onward with more accurate models). This was not favored in our study, however, because of the lack of data that only starts in 2000, and this despite several tests. An analysis of these forecasts was nonetheless carried out in the notebook `4_bis_NOAA_Forecast.ipynb`.

We therefore work on an autocorrelated series $X_1, \dots, X_t$, where each residual is described by 50 features that are themselves largely correlated. Here we seek to avoid *data leakage* and to tune the hyperparameters correctly.

5.2 ML models selection

For each city, we have about 25 000 daily observations and around fifty features, most of them strongly correlated with one another, while the target itself, the residual X_{t+1}, forms a highly autocorrelated series. This twofold correlation, between the features and within the series, constrains the choice of models and requires favoring estimators that are regularized or little sensitive to variable redundancy. Owing to my limited computational resources and the framing of the problem, we test 3 classic supervised-learning models, each with its own strengths and weaknesses.

Generally speaking, all the models considered fall under regularized empirical risk minimization (Vapnik 1998) and differ only in their hypothesis class, their loss and their regularization term. Each produces a predictor $\widehat{f}$ solution of

$$\widehat{f} \in \operatorname*{arg\,min}_{f \in \mathscr{F}} \frac{1}{N} \sum_{i=1}^{N} \ell\big(y_i, f(x_i)\big) + \Omega(f), \tag{12}$$

where $\mathscr{F}$ is the hypothesis class, ℓ the loss function (here quadratic) and Ω the regularization term. It is the choice of this triple that distinguishes the three retained models.

We retain three of them, complementary. First, a linear model such as the classical OLS could have been used, but the strong collinearity of the features makes the feature matrix $X^\top X$ (not to be confused with our target $y_t = X_t$) very unstable, which causes prediction blow-ups, also observed in the recent work of Barnor, Kampouridis, and Kanellopoulos (2026). We therefore prefer a **Ridge** regression, which corrects this flaw thanks to an L^2 penalty on the coefficients.

Next, the problem lends itself well to ensemble methods: they capture nonlinearities without any parametric assumption, remain robust to feature redundancy, and require no prior scaling. We test its two main paradigms. *Bagging* trains many trees in parallel, each on a resampling of the data, and averages their predictions, which reduces variance. *Boosting* adds the trees sequentially, each correcting the residual errors of the previous ones, which reduces bias.

Random Forest introduces nonlinearities by aggregating trees (*bagging*). Its column subsampling decorrelates the trees and makes it robust to correlated features. **XGBoost** complements it with *boosting* (bias reduction, whereas the forest reduces variance), with explicit regularization.

Other ensemble methods could have been considered, such as LightGBM or AdaBoost, but were not suited here (a relatively high number of features makes AdaBoost of little interest, and LightGBM of little relevance in the face of highly correlated features), nor CatBoost, for lack of categorical variables.

The mathematical formulations of the three models are recalled in appendix K.

5.3 Nested time series cross-validation

Tuning the hyperparameters of a model and then judging it on the same splits gives an overly optimistic measure, and above all risks overfitting these hyperparameters to the particularities of a city or a fold.

A classical cross-validation, which randomly distributes the observations into folds, would moreover be misleading: our test would end up surrounded by nearly identical training points, and the model would indirectly see what it is supposed to predict, which artificially inflates performance. We must therefore respect the chronological order and work with an expanding window (*walk-forward*), where the training set always precedes the reserved portion, forbidding any leakage from the future into the past.

To tune the hyperparameters without reintroducing this bias, a nested cross validation is recommended. We nest two loops: an inner loop selects the hyperparameters, while an outer loop reserves data that this tuning has never seen.

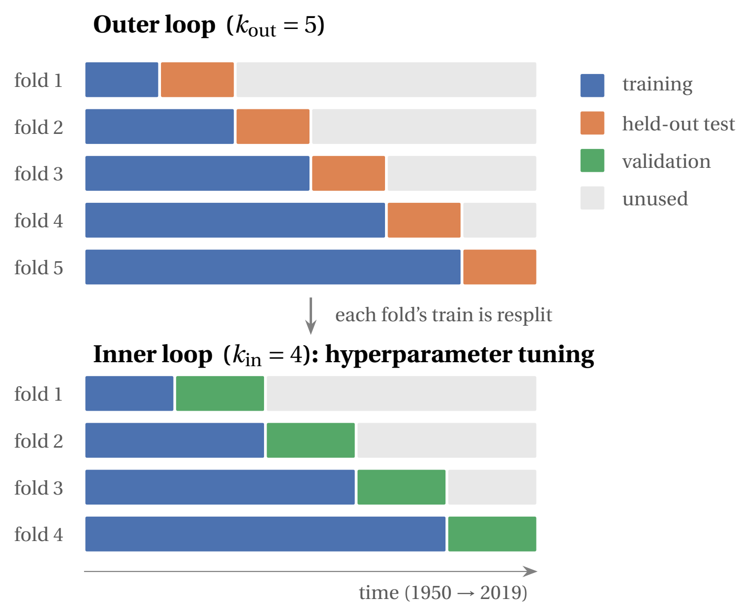

Figure 12: Nested cross-validation with an expanding window (*walk-forward*): outer loop ($k_{out} = 5$) on top, inner loop ($k_{in} = 4$) for hyperparameter tuning at the bottom, over 1950–2019

Concretely, the training history (1950–2019) is split into $k_{out} = 5$ outer folds and for each one, the training set is itself resplit into $k_{in} = 4$ inner folds, on which a grid search selects the hyperparameters (the search grids for the best hyperparameters appear in appendix L). As final values we retain the median of the choices obtained over the outer folds, more robust than a single fold, then the model is retrained one last time over the entire 1950–2019 window.

The criterion guiding the selection of hyperparameters is the mean absolute percentage error (MAPE). We evaluate it on the reconstructed temperature $\widehat{T} = \widehat{X} + \theta$ and express it in Kelvin:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|T_i - \widehat{T}_i|}{T_i + 273.15}, \qquad \text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|T_i - \widehat{T}_i|. \tag{13}$$

Two reasons motivate this choice. On the one hand, the denominator in Kelvin always remains largely positive and avoids the divisions by zero that would make the MAPE explosive around 0 °C, where a relative error measured in Celsius would make no sense. On the other hand, the expanding window makes the sample grow from one outer fold to the next, so that the folds do not cover directly comparable scales. We therefore prefer a relative error (the MAPE, dimensionless) to an absolute error (the MAE, in °C), so as to bring all folds and all cities onto a common basis.

table 5 reports this MAPE (average over the outer folds) for the three models and the seven cities.

Table 5: Internal validation score (MAPE in Kelvin, %, average of the outer folds), by city and by model.

City	Ridge	Random Forest	XGBoost
Chicago	0.918	0.937	0.932
New York	0.771	0.784	0.779
Boston	0.892	0.909	0.901
Minneapolis	0.934	0.954	0.946
Dallas	0.801	0.814	0.810
Los Angeles	0.388	0.394	0.394
Seattle	0.530	0.535	0.536

5.4 Out-of-sample evaluation

Once training is done, we measure the performance of the models on the 2020–2025 test period, touched for the first time. Daily accuracy is not an end in itself: a weather derivative pays the *cumulative* HDD over the contract window. The quantity of interest is therefore the error on this index, and this across the different horizons that a desk must cover. We retain the absolute percentage error (APE), the relative error on what the contract actually pays out:

$$\mathrm{APE}_H = \frac{|H_n - \widehat{H}_n|}{H_n}, \tag{14}$$

where H_n is the realized cumulative HDD (equation (2)) and $\widehat{H}_n$ its forecast. The fidelity of the temperature trajectory and two variants of the HDD error (absolute, and normalized by the climatological standard deviation to compare horizons with one another) are reported in appendix N. To explain what the burn is in the context of temperature forecasting, it is a reference method that uses the historical climatology to estimate future temperatures: it relies on the means and variances of past temperatures to predict future values, without accounting for recent weather conditions or short-term trends.

The decisive subtlety concerns the horizon. A temperature derivative is sold for an upcoming season, bought several months in advance: a gas company or a ski resort buys its cover in October for the following winter, and thus cares about a seasonal window located far ahead, not the next thirty days. We evaluate the models at horizons measured from the first day of each test winter (30, 90 and 365 days) and at a horizon *with lead time*, ISL (*one season lead*), which covers a season starting a quarter later.

To estimate temperatures over an entire period while we only predict X_{t+1}, we use a recursive approach: once $\widehat{X}_{t+1}$ is predicted thanks to the d features, we reconstruct some features from this new value, which provides the d features allowing $\widehat{X}_{t+2}$ to be predicted, and so on.

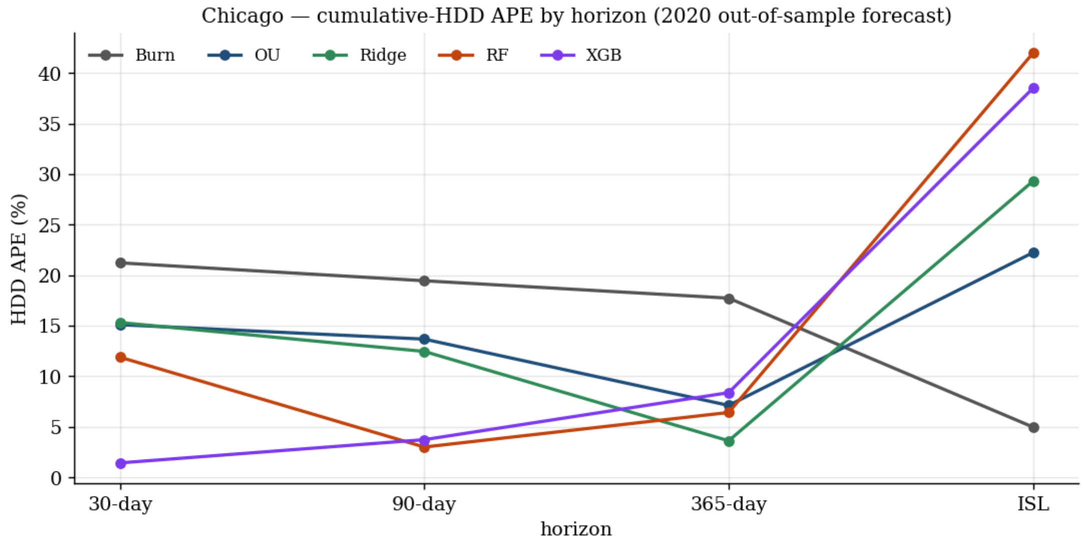

Figure 13: Chicago: APE on cumulative HDD by horizon, for the climatological burn and the four models (out-of-sample forecast from January 1st, 2020, mid-summer excluded).

At near heating horizons, the temperature models clearly outperform the climatological *burn* (figure 13): in Chicago, over the 90-day window, the Random Forest brings the HDD APE down to 3% against 19% for the burn, by exploiting the recent information that the burn ignores. But the advantage reverses at the one-season lead time (ISL): the recursive forecast has lost all conditional information and drifts (APE of 22 to 42%), while the burn, pure climatology, once again becomes by far the most accurate (5%). This is the central result: a temperature model only helps as long as there is information left to exploit (HDD in winter and CDD in summer among others in our case).

Two precautions accompany this reading. On the one hand, the APE inflates when the HDD of the window is low, hence the exclusion of mid-summer. appendix N confirms, through an absolute error and an error normalized by σ_H, that the ranking does not depend on it, and that it holds over the five test winters. By far the most important and the most difficult, since the HDD is an asymmetric index (appendix M), an overestimation and an underestimation of the temperature of the same magnitude do not weigh on it identically.

5.5 Generating temperature scenarios with ML models

The previous evaluation dealt only with a *single* trajectory, the deterministic recursive forecast. But valuing a derivative requires a *distribution* of trajectories, comparable to the one naturally generated by the OU of Part 4. Yet the three learning models are deterministic: with fixed features, $\widehat{X}_{t+1}$ is unique. We must therefore attach a dispersion generator to them, which we do by resampling the out-of-sample errors (bootstrap). The issue here is that computing the price of a financial option requires knowing the distribution of all future probabilities. Thus, from these deterministic predictions we seek to create plausible scenarios that represent the empirical distribution of these predictions.

Methodology After training, we collect the errors of the recursive forecast over the entire test period, $e_t = X_t - \widehat{X}_t = T_t - \widehat{T}_t$. These errors are neither Gaussian nor homogeneous over time: they are asymmetric, heavy-tailed, and slightly more dispersed in winter (figure 14). We therefore resample them with replacement, stratifying by month to respect this seasonality, and add them to the central trajectory:

$$\widetilde{T}_{t_i} = \widehat{T}_{t_i} + \widetilde{e}_i, \qquad \widetilde{e}_i \sim \widehat{F}_{\mathrm{mois}(t_i)}, \tag{15}$$

where $\widehat{F}_{\text{mois}}$ is the empirical distribution of the out-of-sample errors of the month considered (the evaluated window being removed from the pool). The empirical resampling thus restores the asymmetry and the tail thickness of the real errors, which Gaussian noise would not reproduce.

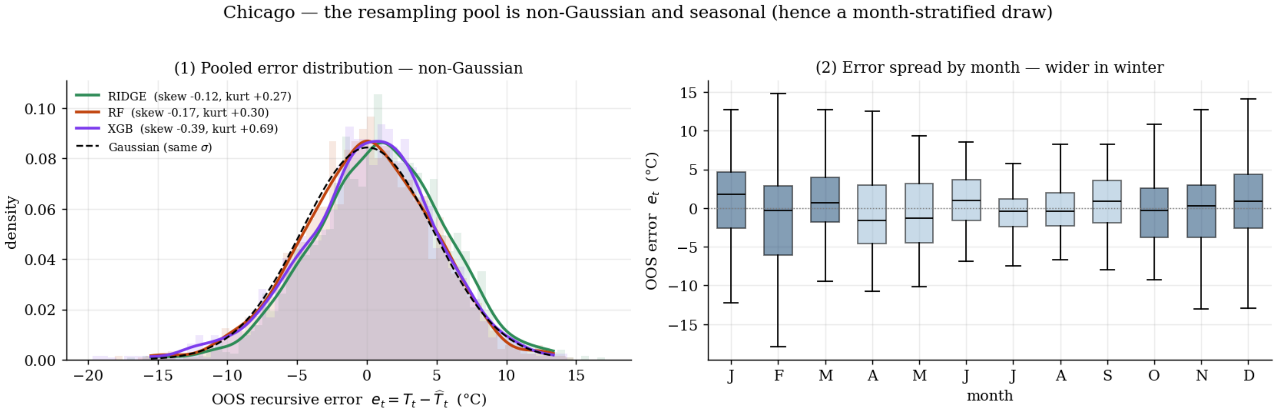

Figure 14: Chicago: on the left, the distribution of the out-of-sample recursive forecast errors of the three models (the resampling pool); on the right, the dispersion of these errors by month.

Repeated N times, this scheme transforms the single forecast into a distribution of trajectories (figure 15), hence into a distribution of the cumulative index.

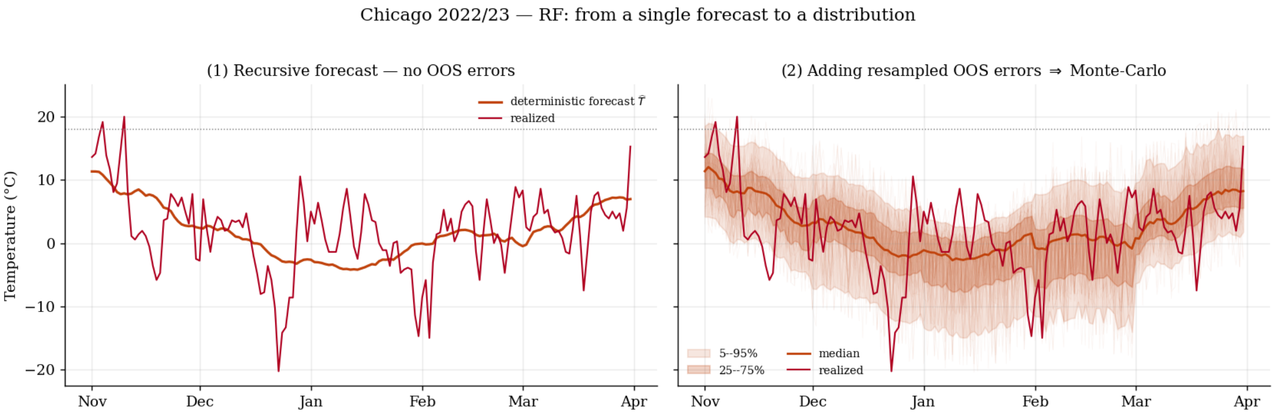

Figure 15: Chicago, winter 2022/23 (Random Forest). On the left, the deterministic recursive forecast; on the right, the Monte Carlo trajectories obtained by adding the resampled out-of-sample errors.

This method, although artificial, allows comparing the historical distributions ("burn") as well as the statistical ones with our OU model. The empirical distribution is rather faithful to the reality of the HDD index outcomes, which allows the asymmetry (skewness) and the heavy tails to be captured.

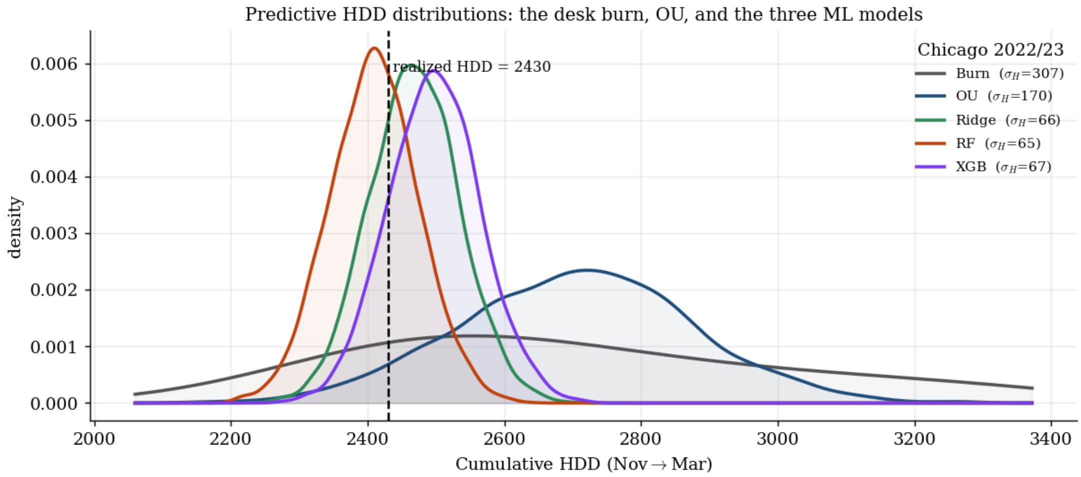

Figure 16: Chicago, winter 2022/23: predictive distribution of cumulative HDD (°C·days) under the desk burn, the OU and the three learning models, with the realized HDD as a dotted line.

The resulting index distribution (figure 16) summarizes everything that matters for valuation: it is this distribution that, confronted with the strike, will set the price. The dispersion σ_H illustrated here for Chicago is by no means universal: across the seven cities, it ranges from about 24 °C·days in Los Angeles (temperate oceanic climate) to 76 in Minneapolis (harsh continental). This gap carries over directly to prices, and it is the average over the seven cities, $\sigma_H \approx 53$, that will appear at the pricing contest (section 7).

6 Pricing Derivatives on an Incomplete Market: the Weather Challenge

This section deals with the valuation of options on weather contracts. We first present the general framework, then develop the *closed-form* valuation under the Ornstein–Uhlenbeck model, whose Gaussian structure lends itself well to it. The learning models, being deterministic, will instead be valued by Monte Carlo simulation.

We take the point of view of a desk (bank, fund, reinsurer) at a valuation date s. The contract is written on a climate index (HDD, CDD, CAT, ...) accumulated over a window that begins at a fixed reference date t_1 (the opening of the heating season) and spans n days, i.e. $t_n = t_1 + n - 1$. The two parameters that vary in our study are therefore the valuation date s and the duration n, subject to the lead-time constraint $s \le t_1$. At date s, the desk has only past information available: we work under the filtration $\mathscr{F}_s$.

Consider a *call* with strike K and tick $\alpha = \$20$ fixed in this part, written on the cumulative HDD index H_n (equation (2)) over the window $[t_1, t_n]$. Under the risk-neutral measure $\mathbb{Q}$, its price at date s is the discounted expectation of the payoff:

$$C_s(n) = e^{-r(t_n - s)}\,\mathbb{E}^{\mathbb{Q}}\Big[\,\alpha\left(H_n - K\right)^+ \Big|\, \mathscr{F}_s\Big]. \tag{16}$$

The date s thus plays a double role: it fixes the discounting horizon TTM $= t_n - s$ and the available information $\mathscr{F}_s$. The corresponding put is denoted $P_s(n)$, with a payoff $\alpha\,(K - H_n)^+$.

The central difficulty is the change of measure. The market here is *incomplete*: the underlying is a weather index and temperature is not a tradable asset (one can neither buy nor sell it to hedge). The risk-neutral measure is therefore not unique, and we introduce a *market price of risk* λ that parameterizes the transition $\mathbb{P} \to \mathbb{Q}$ (in the sense of Girsanov), to be calibrated in the absence of a hedging asset.

A useful simplification, due to Alaton, Djehiche, and Stillberger (2002), rests on a winter hypothesis ($\mathscr{H}$): in the depth of winter, the daily temperature almost surely stays below the reference threshold, i.e. $\mathbb{P}(T_t > T_{\text{ref}}) \approx 0$. We obtain:

$$H_n \overset{(\mathscr{H})}{=} \sum_{t=t_1}^{t_n} \left(T_{\text{ref}} - T_t\right) = n\,T_{\text{ref}} - \sum_{t=t_1}^{t_n} T_t. \tag{17}$$

From the descriptive analysis, the daily temperature is Gaussian. Although $T_{t_1}, \dots, T_{t_n}$ are correlated, their linear combination H_n is Gaussian as well. Under ($\mathscr{H}$), we therefore have

$$H_n \mid \mathscr{F}_s \sim \mathscr{N}\big(\mu_s(n),\, \Lambda_s(n)^2\big), \qquad \mu_s(n) = \mathbb{E}^{\mathbb{Q}}[H_n \mid \mathscr{F}_s], \quad \Lambda_s(n)^2 = \operatorname{Var}(H_n \mid \mathscr{F}_s),$$

which opens the way to a closed-form formula, the Gaussian analog of the Black & Scholes one. This hypothesis must nonetheless be verified empirically: it holds well for cold continental cities but becomes fragile for mild cities.

6.1 Conditional moments of the index under the OU model

Under the OU model, the Girsanov change of measure (price of risk λ) allows the conditional moments of temperature to be computed explicitly. The derivations are deferred to appendix I. The conditional mean under $\mathbb{Q}$ is written

$$\boxed{\mathbb{E}^{\mathbb{Q}}[T_t \mid \mathscr{F}_s] = \theta(t) + \big(T_s - \theta(s)\big)e^{-\kappa(t-s)} - \frac{\lambda\,(1 - e^{-\kappa})}{\kappa} \sum_{j=s+1}^{t} \sigma_X(j)\, e^{-\kappa(t-j)}} \tag{18}$$

and, for $s < t_i < t_j$, the conditional variance and autocovariance are written

$$\boxed{V(t,s) \equiv \mathrm{Var}(T_t \mid \mathscr{F}_s) = \frac{1-e^{-2\kappa}}{2\kappa} \sum_{j=s+1}^{t} \sigma_X^2(j)\, e^{-2\kappa(t-j)}, \qquad \mathrm{Cov}(T_{t_i}, T_{t_j} \mid \mathscr{F}_s) = e^{-\kappa(t_j-t_i)}\, V(t_i,s).} \tag{19}$$

We deduce from this the two Gaussian parameters of the index:

$$\mu_s(n) = n\,T_{\mathrm{ref}} - \sum_{i=1}^{n} \mathbb{E}^{\mathbb{Q}}[T_{t_i} \mid \mathscr{F}_s], \qquad \Lambda_s(n)^2 = \sum_{i=1}^{n} V(t_i,s)\left[1 + 2\,e^{-\kappa}\,\frac{1-e^{-\kappa(n-i)}}{1-e^{-\kappa}}\right]. \tag{20}$$

6.2 Closed-form call price

Since H_n is Gaussian, the call price admits a closed form (the Gaussian analog of Black & Scholes). Setting $\eta_s(n) = \dfrac{K-\mu_s(n)}{\Lambda_s(n)}$:

$$\boxed{C_s(n) = \alpha\, e^{-r(t_n-s)} \Big[\big(\mu_s(n)-K\big)\,\Phi(-\eta_s(n)) + \Lambda_s(n)\,\varphi(\eta_s(n))\Big]} \tag{21}$$

where Φ and φ are the cumulative distribution function and the density of the $\mathcal{N}(0,1)$ law. The put $P_s(n)$ is obtained analogously (equation (I.1)). The full derivation is given in appendix I.

6.3 Price by Monte Carlo simulation

The preceding closed form relies on the Gaussian structure of the OU model and on the hypothesis $(\mathscr{H})$. Monte Carlo simulation offers a general alternative, valid whether the price comes from a statistical method (OU) or from the learning models, and thus allows the price of the same contract to be estimated under each of the models.

We generate N temperature trajectories under the model dynamics, from which we extract the cumulative HDD $H_n^{(i)}$ of each trajectory i. The call price is the discounted empirical average of the payoffs:

$$\boxed{\widehat{C}_s(n) = \alpha\, e^{-r(t_n-s)}\, \frac{1}{N} \sum_{i=1}^{N} \big(H_n^{(i)} - K\big)^{+} \xrightarrow[N\to\infty]{} C_s(n)} \tag{22}$$

By the law of large numbers, this estimator converges almost surely to the price equation (16) as the number of trajectories N grows. The Monte Carlo is fed by the HDD distributions already constructed: that of the OU, simulated under $\mathbb{Q}$, and those of the three learning models, obtained by recursive forecasting and resampling of the out-of-sample errors (section 5.5). For each model, we draw N trajectories, accumulate the HDD over the contract window, then apply the estimator equation (22). Under the OU, the price obtained coincides with the closed form equation (21), which validates the implementation.

figure 17 illustrates this valuation: the call price as a function of the strike, under three trajectory generators. The more dispersed the index distribution (burn, OU), the more value the call retains far out of the money. The narrow law of the Random Forest, on the contrary, makes the price collapse as soon as the strike exceeds its mean. The choice of model therefore weighs above all in the out-of-the-money tail, which the pricing contest (section 7) will quantify in dollars.

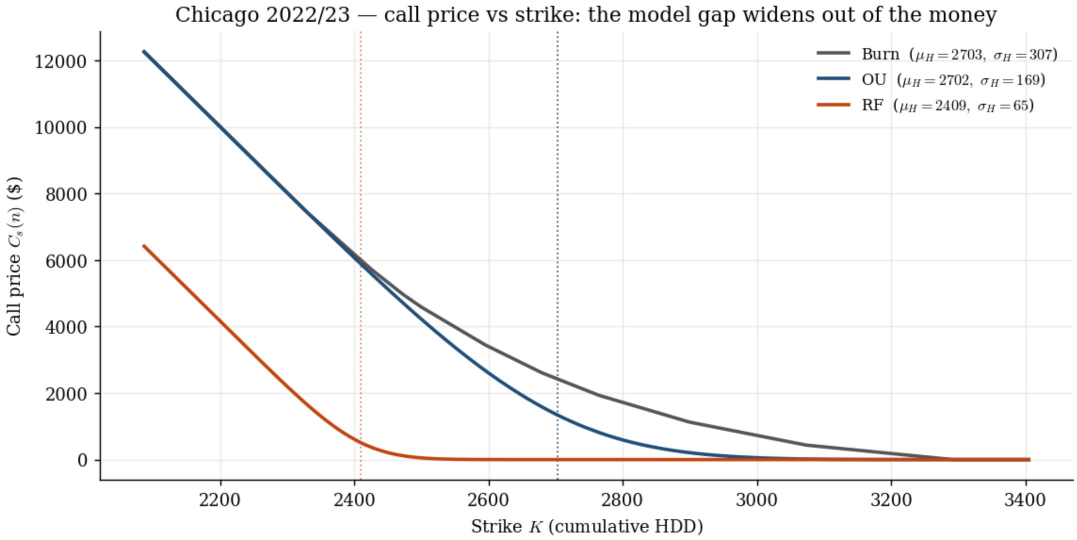

Figure 17: Chicago 2022/23: HDD call price $C_s(n)$ as a function of the strike K, under the burn, the OU, and the Random Forest (mean μ_H and dispersion σ_H of each index distribution in the legend, dotted lines at μ_H).

7 Financial Aspects : a pricing contest

As stated at the beginning of this report, one of the applications of climate-index forecasts is the pricing of weather derivatives. The real test of a temperature model is not its statistical accuracy but the price it assigns to a contract. We therefore pit the six models against one another on their common purpose: valuing the same HDD call, and then comparing that price to the **payoff actually paid out**. This section formalizes the contest and presents its results, before criticizing its design: it is naive, and the true difficulty, capturing the HDD distribution of a given winter, remains out of reach.

Thanks to our various models, which we will denote $m \in \{\text{Climatological}, \text{Burn}, \text{OU}, \text{Ridge}, \text{Random Forest}, \text{XGB}\}$, we can estimate the financial contribution of statistical and learning approaches to the pricing of structured products. Other uses could have emerged, notably coupling with natural-gas volatility to hedge or speculate. Here we focus on the pricing of winter HDD calls to illustrate one such use.

We consider the case where, each year, a desk must protect itself against weather events over winter w and, to do so, estimate the cumulative HDD over the period $[t_1, t_n]$, from November to March. As a naive benchmark, this desk takes as its strike the *burn*, that is, the mean of the cumulative HDD realized over prior winters. The realized index is measured once the winter has elapsed:

$$K_{c,w} = \frac{1}{\#\{w' < w\}} \sum_{w'<w} H^{\text{real}}_{c,w'}, \qquad H^{\text{real}}_{c,w} = \sum_{t=t_1}^{t_n} \max\big(T_{\text{ref}} - T_t, 0\big). \tag{23}$$

Six competitors face off. The *burn* replays the realized degree days of past winters: it is the desk benchmark. The *climatology* is a naive deterministic analyst (the θ component alone, price equal to the intrinsic value). The *OU* is conditioned on the residual observed on October 1$^{\text{st}}$. Ridge, Random Forest, and XGBoost generate their trajectories by resampling their out-of-sample errors.

For each pair (city c, winter w), all models value the *same* contract: an HDD call from November to March, with tick $\alpha = 20\$$, $r = 1\%$ and $\lambda = 0$, and strike equal to the burn $K_{c,w}$ defined above (equation (23)).

For a fixed model m, we simulate for each (c, w) N trajectories conditional on the pricing date s, from which we derive the cumulative HDD $H^{(m,i)}_{c,w}$ of each trajectory i. The price $\hat{\Pi}_m(c, w)$ is the Monte Carlo estimator equation (22), applied to these trajectories with the burn strike $K_{c,w}$:

$$\hat{\Pi}_m(c, w) = \alpha\, e^{-r(t_n - s)} \frac{1}{N} \sum_{i=1}^{N} \big(H^{(m,i)}_{c,w} - K_{c,w}\big)^+ \xrightarrow[N\to\infty]{} \Pi_m(c, w). \tag{24}$$

Once the winter has passed, the contract pays out its realized payoff $Y_{c,w} = \alpha\, e^{-r\,\text{TTM}} \max(H^{\text{real}}_{c,w} - K_{c,w}, 0)$. This $Y_{c,w}$ is the payoff of the single trajectory actually realized by winter w: it is the only observable draw, the one against which we confront the price quoted by each model. The gap between the quoted price and this payoff defines the pricing error, aggregated over the $C \times W = 7 \times 5 = 35$ contracts:

$$\boxed{\text{RMSE}_m = \sqrt{\frac{1}{35} \sum_{c,w} e_m(c, w)^2}, \qquad e_m(c, w) = \Pi_m(c, w) - Y_{c,w}} \tag{25}$$

Since most of the 2020–2025 test winters were mild, the realized HDD often stays below the strike: 22 of the 35 contracts expire at zero. The realized payoffs $Y_{c,w}$ are reported in figure O.2, and the corresponding burn strikes $K_{c,w}$ in figure O.1.

Each model assigns a different price $\hat{\Pi}_m(c, w)$ to each contract, and these prices order themselves right away by their dispersion. The burn shows the largest standard deviation σ_H (figure 16) and therefore quotes the highest. The OU, conditional but endowed with seasonal persistence, likewise remains dispersed. The learning models produce a much tighter HDD distribution and quote the lowest. The climatology, being deterministic, has no dispersion at all: its price reduces to the intrinsic value.

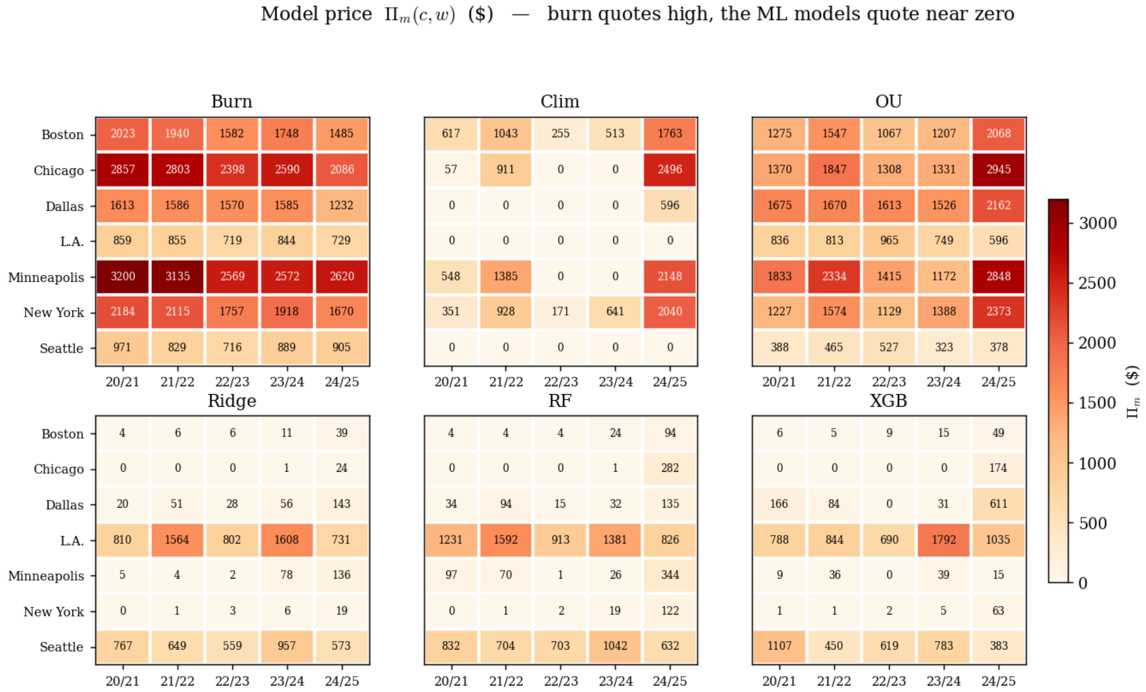

Figure 18: Price $\hat{\Pi}_m(c, w)$ of each model (in $), by city and test winter.

Let us now confront these prices with the realized payoff. The matrix of signed errors $e_m(c, w)$ (figure 19) shows that the burn and the OU systematically overprice (red cells): they price a dispersion that the mild winters of 2020–2025 did not realize. The learning models, on the other hand, stay close to zero almost everywhere, their underpricing coinciding with realized payoffs that are themselves null.

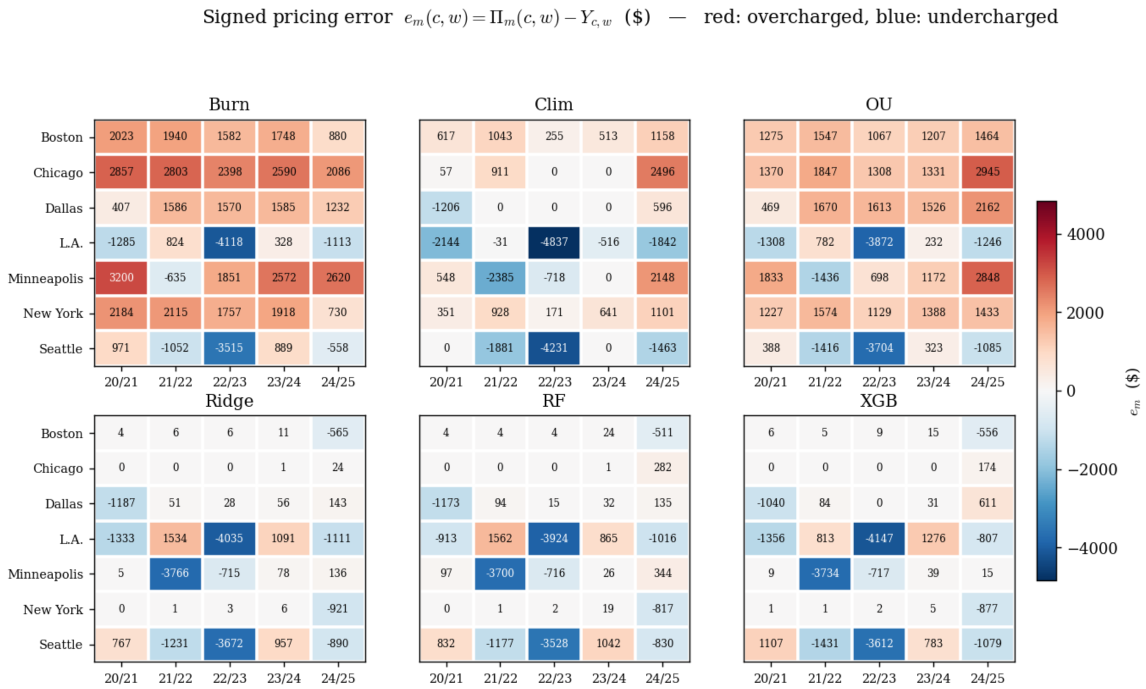

Figure 19: Signed pricing error $e_m(c, w) = \Pi_m(c, w) - Y_{c,w}$ (in $), by city and test winter; in red, the overpriced contracts.

Once aggregated, the error places the three learning models at the top, far ahead of the burn (figure 20 and table 6). The Random Forest achieves the lowest RMSE (1236$ per contract, versus 1974$ for the burn, a decrease of roughly 37%), followed by XGBoost and Ridge. The climatology and the OU rank next, with the burn closing out the field.

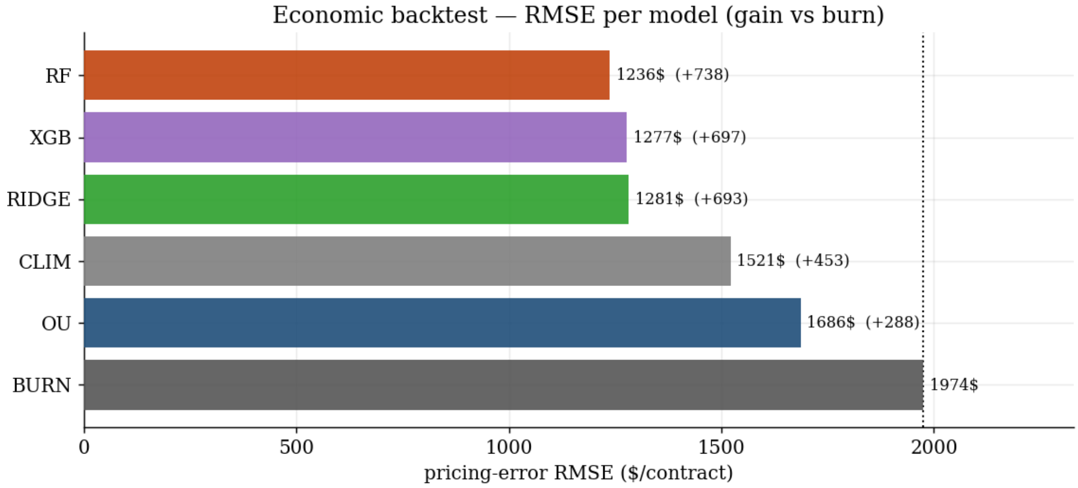

Figure 20: Aggregated pricing error (RMSE, in $) by model and dollar gain relative to the burn.

Table 6: HDD call pricing error (RMSE) and simulated dispersion σ_H, aggregated over 5 winters × 7 cities (35 contracts).

Model	**RMSE ($/contract)**	$\boldsymbol{\sigma_H}$
Random Forest	1236	53
XGBoost	1277	53
Ridge	1281	53
Climatology	1521	—
OU	1686	131
Burn (desk)	1974	228

This ranking holds across the range of economically relevant strikes (figure 21). At the money and below ($K/\text{ATM} \leq 1$), the three learning models clearly dominate the burn, the OU, and the climatology, with the Random Forest and XGBoost sharing the lead. Beyond the money, however, the call becomes worthless under every conditional distribution, which all converge near zero; only the burn, being more dispersed, still charges a premium there and remains everywhere the most costly in error. The hierarchy is therefore decided only where dispersion matters, that is, around and below the strike.

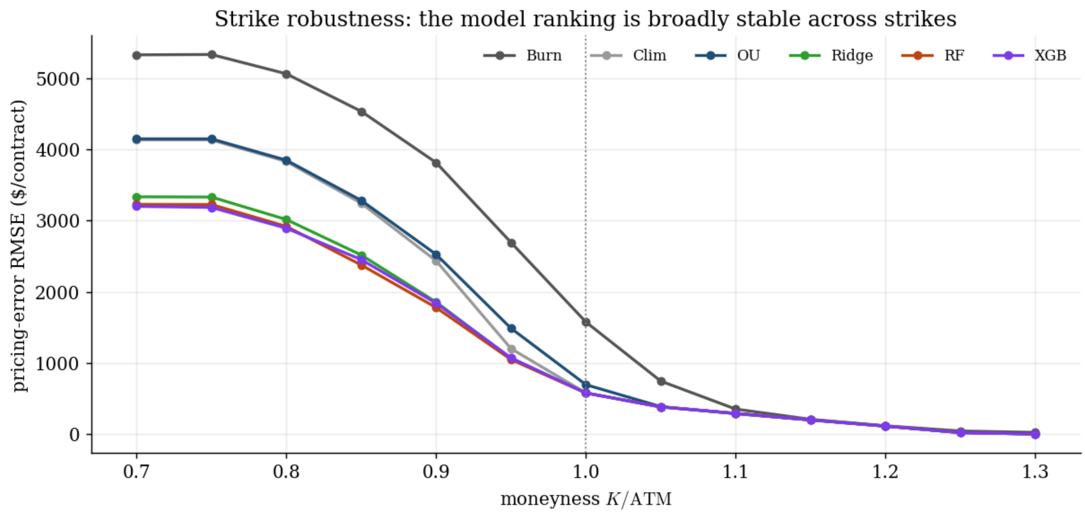

Figure 21: Pricing RMSE (in $) by model as a function of moneyness K/ATM, swept from 0.70 to 1.30 in steps of 0.05.

7.1 Limits of the economic approach

7.2 Limits of the experiment

This ranking, however, rests on a naive design that must now be criticized, starting with the way the strike K is chosen and with the naive benchmark that is the burn.

First, the burn takes as the HDD distribution the set of all past winters. Yet the climate is warming: older, colder winters overestimate the future HDD and set too high a strike (figure 22). This is precisely the bias corrected by the climatological model, which uses only the current trend $\theta(t)$. A first improvement would be to detrend the burn in order to reduce its standard deviation σ_H. However, as figure 22 shows, this is not enough.

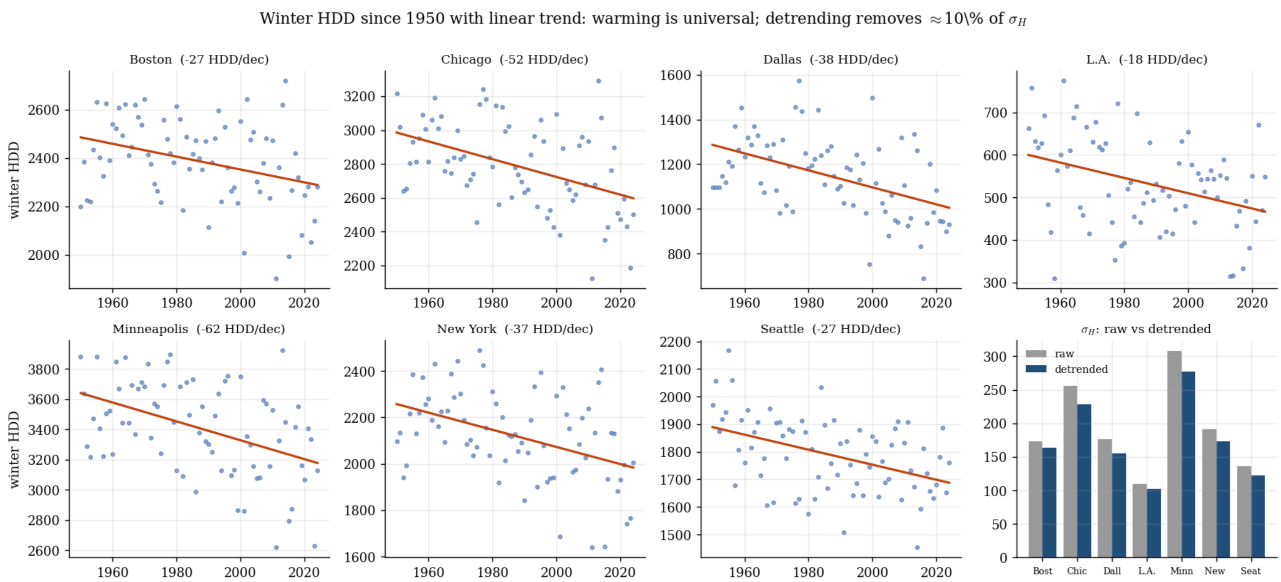

Figure 22: Winter HDD by city since 1950, linear trend, and σ_H before/after detrending (right-hand bars).

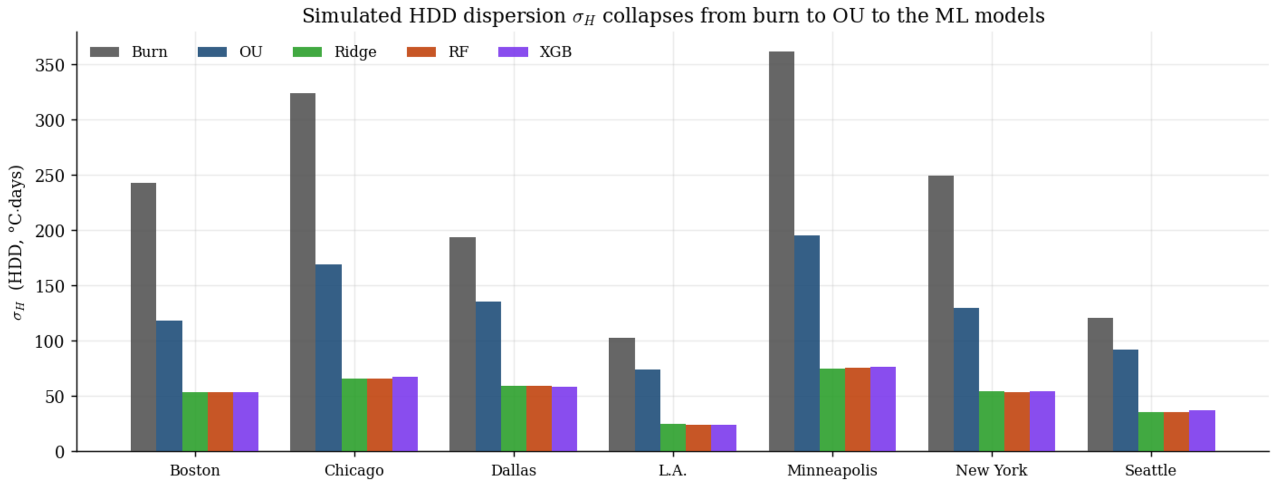

Figure 23: Dispersion σ_H of the cumulative HDD by city, for the burn, the OU, and the three learning models.

The difficulty is ultimately deeper. What we would like to know is not the distribution of the HDD *across* winters, which the burn and its detrending reconstruct, but the distribution of the HDD *of a given winter*, the conditional law that actually governs this contract. The two objects differ: the first blends the variability from one winter to the next, the second is specific to the year under consideration. Yet this conditional law is fundamentally unobservable: for each winter, a single temperature trajectory is realized, hence a single HDD. We never have a sample of the winter itself. This is the true challenge of weather-derivative pricing, and the reason why none of my models can claim to recover the "correct" distribution: **there is no observable ground truth against which to compare it**.

The central issue of this project therefore comes down to the following question: given the past distributions, how can we know that the distribution returned by a model is the correct one? Several avenues are conceivable, for example combining the models and retaining the mean of their distributions, some underestimating the risk and others overestimating it, in the hope that their errors offset one another.

8 Boreas: an interactive pricing desk

The entire chain described so far, from the estimation of the temperature models to the valuation of the contracts, has been gathered into an interactive application, **Boreas**[1], built with Streamlit and deployed publicly on Hugging Face Spaces. The guiding idea is a strict separation between heavy computation and usage: the models are trained offline and serialized into artifacts (`joblib` files), so that the application only performs cheap vector operations on pre-computed scenarios. A quotation is thus rendered almost instantaneously, which makes the tool usable as a genuine trading desk.

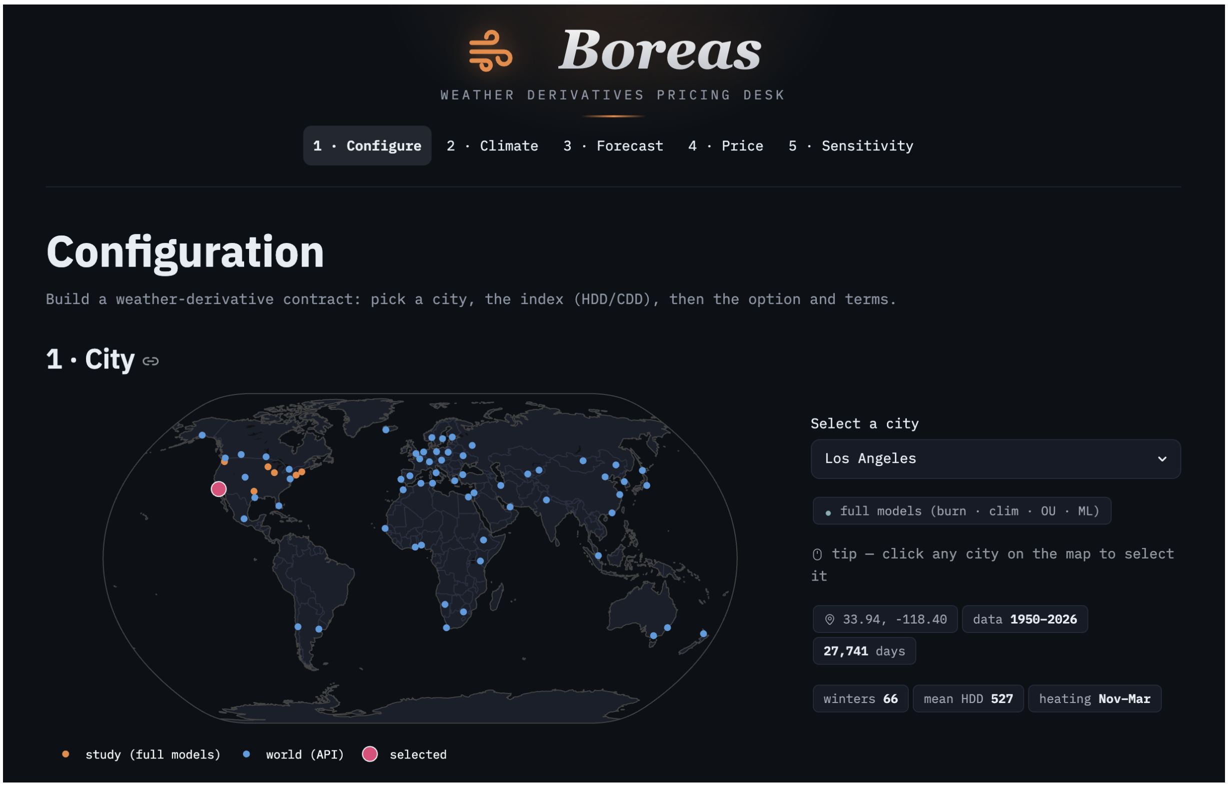

Figure 24: Interface of the Boreas application.

The application follows the natural workflow of an operator and is organized into five pages:

- **Configuration**: choice of the city (one of the seven study cities with trained models, or any major city in the world via the Open-Meteo API, in which case the quotation is limited to the burn) and of the contract parameters (coverage window, valuation date, strike, tick).
- **Descriptive**: the temperature climate of the selected city, seasonality and distribution.
- **Forecast**: the trajectories simulated under the OU and the learning models over the contract window, as a fan chart, against the naive climatology.
- **Pricing**: the quotation of the same HDD/CDD call under all models side by side (burn, climatology, OU, Ridge, Random Forest, XGBoost), together with the index distribution and the risk metrics.
- **Sensitivity**: the response of the price and its greeks to the strike K (Bachelier closed form) and to the market factors, a temperature shift ΔT, the volatility σ and, for the OU, the market price of risk λ (re-simulated under $\mathbb{Q}$ by Girsanov).

[1] Named after Boreas, the Greek god of the north wind.

Here one finds again, in dollars, the collapse of the dispersion σ_H from the burn toward the learning models that structures the whole of section 7: Boreas thus makes the contribution of the project directly explorable, by re-pricing a given contract under each model to materialize the valuation gap implied by the choice of the temperature model.

9 Conclusion and Future Work

This project, undertaken on my own initiative, allowed me to enter the world of weather forecasting and its broad applications in finance. Beyond the pricing of *weather contracts*, the uses of this type of modeling remain varied. In particular, I carried out a parallel study on the statistical link between temperature anomalies and the volatility of Henry Hub natural gas. These experiments, available on GitHub, will soon be complemented by a companion report.

Regarding possible improvements to the modeling, the main limitation stems from the choice to restrict ourselves to freely available temperature data, whereas other meteorological factors also influence the phenomenon. Moreover, the available computing resources did not allow us to train more cities: the Boreas application models any city through the OU model, but only the seven cities in this report are equipped with the learning models.

Many models could emerge in the coming years. One can imagine global training models in which the location of cities would enter explicitly: two nearby cities would share a common climatic dynamic, whereas two distant cities would be treated as nearly independent. The range of possibilities remains wide and open.

At the end of this work, one lesson stands out. At the seasonal horizon that governs weather derivatives, the choice of temperature model offers only a limited pricing advantage over the desk's simple burn, with the dispersion of the index mattering more than the accuracy of the mean forecast. The value of a model reveals itself above all in forecasting weather indices at shorter horizons, and it is the coupling between temperature dynamics and energy price volatility that seems to us the most promising avenue for turning a statistical gain into an economic one. This is the direction we intend to explore further.

References

Alaton, Peter, Boualem Djehiche, and David Stillberger (2002). "On modelling and pricing weather derivatives". In: *Applied Mathematical Finance* 9.1, pp. 1–20. DOI: 10.1080/13504860210132897.

Alfonsi, Aurélien and Nerea Vadillo (2023). "A stochastic volatility model for the valuation of temperature derivatives". In: *arXiv preprint arXiv:2209.05918*. AXA Climate / CERMICS ; détails à revérifier.

Barnor, Nee, Michael Kampouridis, and Panagiotis Kanellopoulos (2026). *Machine learning for temperature forecasting in weather derivatives*. détails à revérifier. DOI: 10.1007/s10462-026-11539-0.

Benth, Fred Espen and Jurate Šaltytė Benth (2007). "The volatility of temperature and pricing of weather derivatives". In: *Quantitative Finance* 7.5. détails à revérifier, pp. 553–561. DOI: 10.1080/14697680601155334.

Cao, Melanie and Jason Wei (2004). "Weather derivatives valuation and market price of weather risk". In: *Journal of Futures Markets* 24.11, pp. 1065–1089. DOI: 10.1002/fut.20122.

Cheng, Tuoyuan, Saikiran Reddy Poreddy, and Kan Chen (2025). "Tail risk in weather derivatives". In: *Commodities* 4.2, p. 11. DOI: 10.3390/commodities4020011.

CME Group (2024). *Weather Derivatives Grow as Risks Intensify*. CME Group OpenMarkets. URL: https://www.cmegroup.com/openmarkets/energy/2024/Weather-Derivatives-Grow-as-Risks-Intensify.html (visited on 07/01/2026).

Kuppelwieser, Thomas and David Wozabal (2023). "Intraday power trading: toward an arms race in weather forecasting?" In: *OR Spectrum* 45, pp. 57–83. DOI: 10.1007/s00291-022-00698-5.

Schwartz, Eduardo and James E. Smith (2000). "Short-term variations and long-term dynamics in commodity prices". In: *Management Science* 46.7, pp. 893–911. DOI: 10.1287/mnsc.46.7.893.12034.

Vapnik, Vladimir N. (1998). *Statistical Learning Theory*. New York: Wiley.

Appendices

Appendix A Cities covered by CME weather contracts

Table A.1: The eighteen cities worldwide covered by CME weather contracts.

Region	City	Country
North America	Atlanta	United States
	Boston	United States
	Burbank	United States
	Chicago	United States
	Cincinnati	United States
	Dallas	United States
	Houston	United States
	Las Vegas	United States
	Minneapolis	United States
	New York	United States
	Philadelphia	United States
	Portland	United States
	Sacramento	United States
Europe	Amsterdam	Netherlands
	Essen	Germany
	London	United Kingdom
	Paris	France
Asia	Tokyo	Japan

Appendix B Peak-winter and peak-summer distributions

figure B.1 reproduces the seasonal distributions of figure 3, but restricting winter to the months of December through February and summer to the months of June through August, that is, to the core of each season. The only notable effect is a better separation of the two regimes, since the transition months have been discarded. The conclusions drawn from figure 3 otherwise remain unchanged.

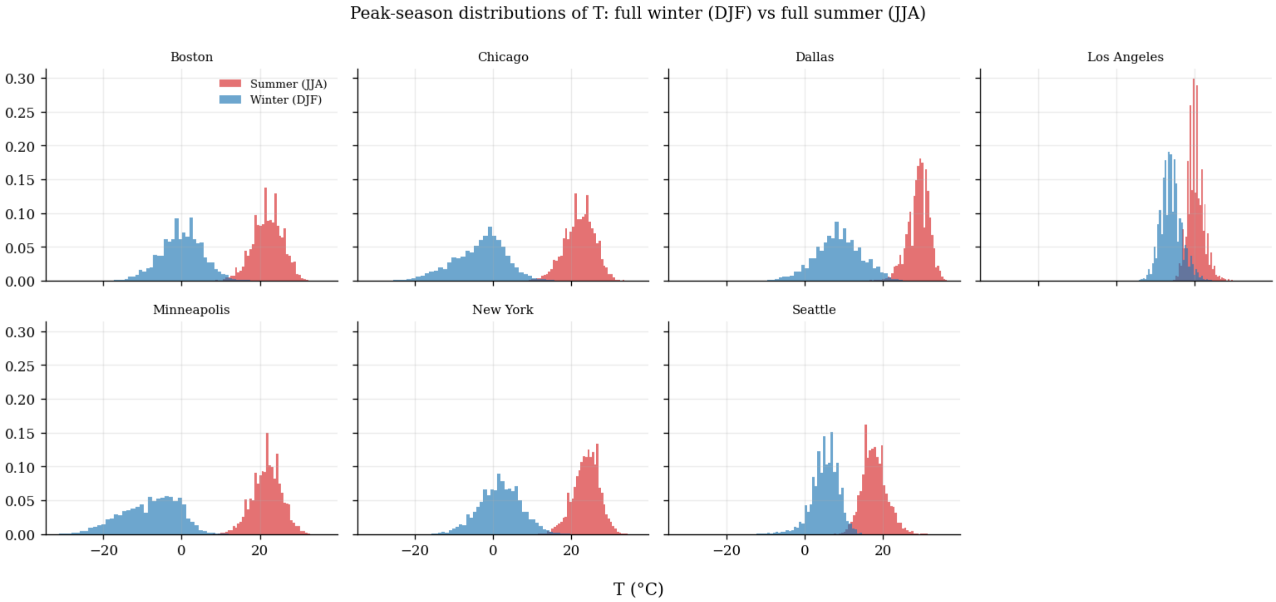

Figure B.1: Distributions of the temperature T in peak winter (December through February) and peak summer (June through August)

Appendix C Seasonal volatility of the additional cities

figure C.1 complements figure 5 for the three cities added outside the CME list. Dallas exhibits the same cyclical behavior as the continental cities of the main study, with pronounced winter variance peaks. Los Angeles and Seattle, with their temperate oceanic climate, display on the contrary a much lower and markedly less cyclical variance, their volatility remaining roughly stable from one season to the next.

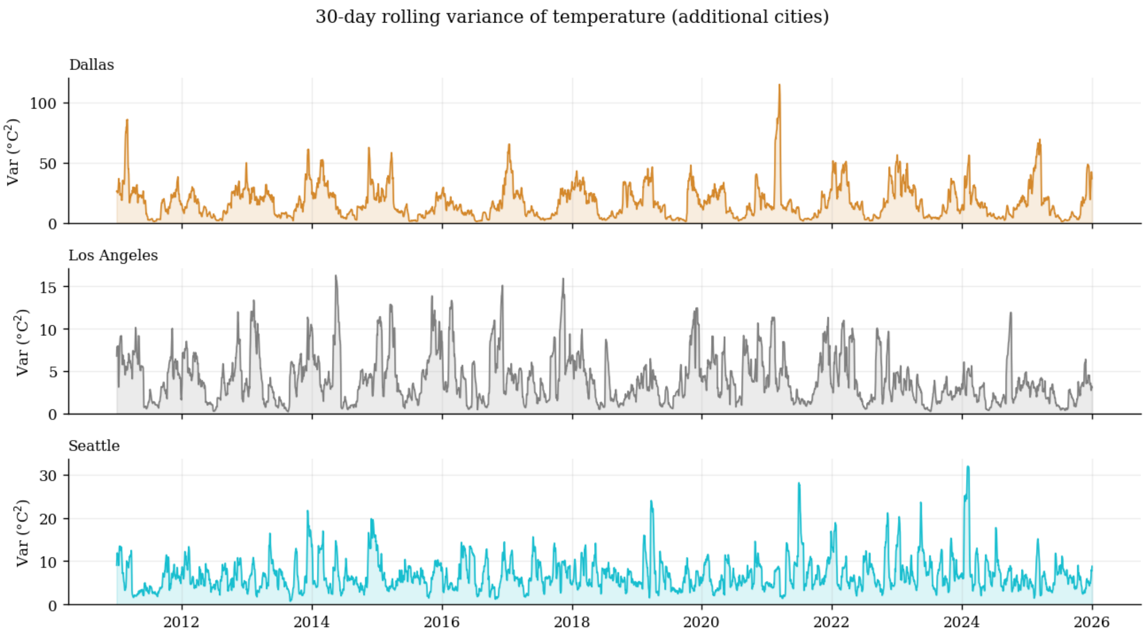

Figure C.1: 30-day rolling variance of temperature, over the last fifteen years, for Dallas, Los Angeles and Seattle.

Appendix D Variance Stability under Warming

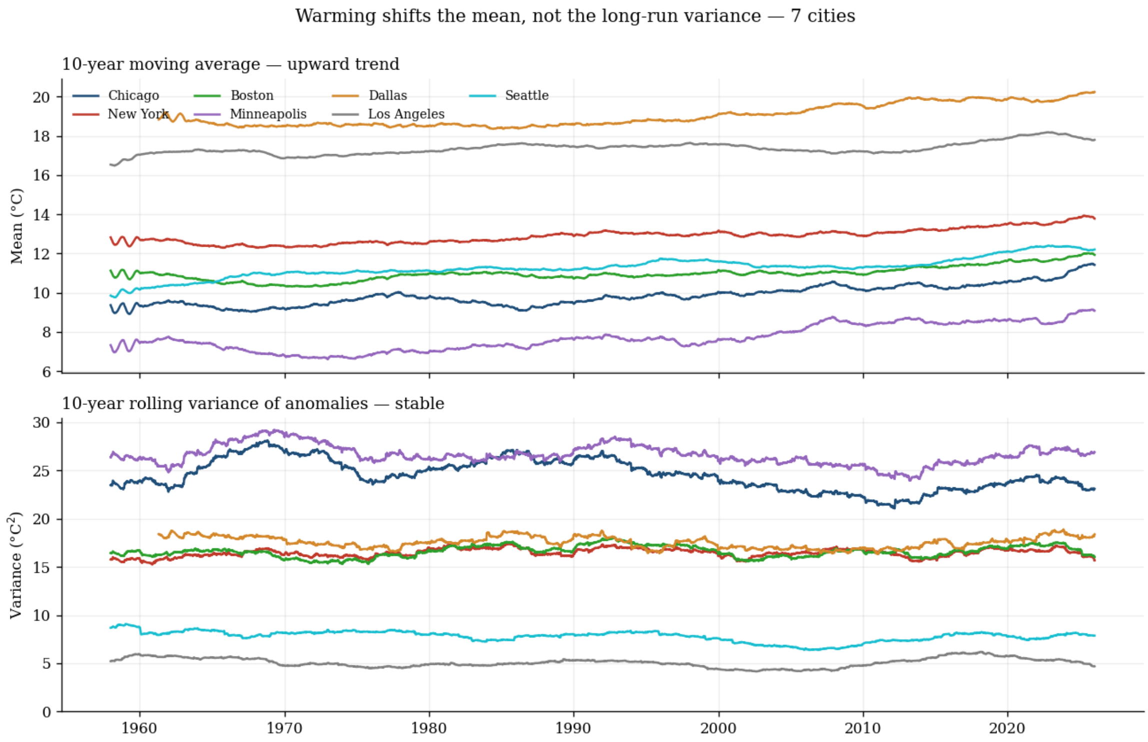

Figure D.1: Ten-year rolling mean (top) and ten-year rolling variance of the anomalies (bottom), for the seven cities.

Appendix E Maximum likelihood estimation of the residual variance

We detail here the computation, summarized in section 4, of the maximum likelihood estimator of the residual variance of the deterministic component. Let $X_1, \ldots, X_n$ be the observed residuals ($X_t = T_t - \theta(t)$), assumed independent and distributed as $\mathcal{N}(0, \sigma^2)$, with density:

$$\forall t \in \{1, \ldots, n\}, \quad f(X_t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{X_t^2}{2\sigma^2}}.$$

The likelihood reads:

$$L_n(X_1, \ldots, X_n; \sigma) = \prod_{t=1}^{n} f(X_t) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{t=1}^{n} X_t^2\right).$$

Passing to the log-likelihood $\ell_n(\sigma) = \log L_n$:

$$\ell_n(\sigma) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{n} X_t^2.$$

The first-order condition $\partial \ell_n / \partial \sigma = 0$ yields:

$$-\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{t=1}^{n} X_t^2 = 0 \quad \Longrightarrow \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n} X_t^2.$$

Substituting $\hat{\sigma}^2$ back into ℓ_n, the constant terms (independent of the order N) group together, which leads to the form used for the AIC and the BIC:

$$-2\log(\hat{L}_n) = n\log(\hat{\sigma}^2) + \text{constant}.$$

Appendix F Fourier order selection: AIC/BIC grid by city

Table F.1: AIC and BIC of the deterministic component $\theta(t)$ by city and by Fourier-series order N (estimation window 1950–2019), with the minimum of each row in bold.

City	Criterion	N=0	1	2	3	4	5	6	7	8
Chicago	AIC	123 172	82 638	82 253	82 178	82 175	82 168	82 138	82 139	**82 126**
	BIC	123 196	82 679	82 311	**82 251**	82 265	82 274	82 260	82 277	82 281
New York	AIC	115 713	72 104	72 062	71 975	71 978	71 966	71 956	71 955	**71 942**
	BIC	115 737	72 145	72 119	**72 048**	72 068	72 072	72 078	72 093	72 096
Boston	AIC	115 540	72 390	72 390	72 238	72 241	72 227	72 225	72 225	**72 218**
	BIC	115 564	72 431	72 447	**72 312**	72 331	72 333	72 347	72 364	72 372
Minneapolis	AIC	130 185	85 050	84 343	84 268	84 264	84 257	84 226	84 211	**84 208**
	BIC	130 209	85 091	84 400	**84 342**	84 353	84 363	84 348	84 349	84 363
Dallas	AIC	112 496	73 254	72 482	72 444	72 430	**72 429**	72 430	72 434	72 436
	BIC	112 521	73 294	72 539	**72 517**	72 519	72 535	72 552	72 572	72 591
Los Angeles	AIC	65 357	43 364	42 274	42 185	42 083	42 062	42 060	**42 043**	42 046
	BIC	65 381	43 405	42 331	42 259	42 173	**42 167**	42 182	42 182	42 200
Seattle	AIC	89 705	55 084	53 225	53 217	53 179	53 174	53 158	53 144	**53 139**
	BIC	89 730	55 125	53 282	53 291	**53 268**	53 279	53 280	53 282	53 294

Appendix G Estimated coefficients of the deterministic component $\theta(t)$

Here are the coefficients of the deterministic component $\theta(t)$ (equation (4)) estimated by ordinary least squares over the estimation window (1950–2019), at the order N selected by the BIC. The constant A and the harmonic amplitudes a_k, b_k are expressed in °C, and the trend slope B is in °C per year (the average annual warming). The empty cells ($k > N$) correspond to the harmonics that were not retained.

Table G.1: Estimated coefficients of $\theta(t)$ by city, at the order N selected by the BIC (1950–2019). A and a_k, b_k in °C, B in °C per year.

	Chicago	New York	Boston	Minneapolis	Dallas	Los Angeles	Seattle
N (BIC)	3	3	3	3	3	5	4
A	9.04	12.39	10.59	6.71	18.36	16.80	10.30
B	0.0206	0.0135	0.0106	0.0258	0.0183	0.0145	0.0257
a_1	-4.78	-4.67	-4.98	-4.91	-3.39	-2.43	-2.53
b_1	-13.22	-11.40	-11.19	-15.71	-10.81	-3.00	-6.58
a_2	-0.05	-0.01	0.03	-0.46	0.65	0.46	1.10
b_2	-0.87	-0.25	-0.07	-1.15	-0.79	-0.49	-0.09
a_3	-0.32	-0.34	-0.41	-0.33	-0.14	0.12	0.07
b_3	-0.23	-0.04	-0.19	-0.23	-0.20	-0.16	0.04
a_4						0.21	0.10
b_4						-0.03	-0.13
a_5						0.03	
b_5						-0.10	

Appendix H Solving the Ornstein–Uhlenbeck Equation

We detail here the solution of equation (6), $\mathrm{d}X_t = -\kappa X_t \,\mathrm{d}t + \sigma_X(t)\,\mathrm{d}W_t$, summarized in section 4. Multiply by the integrating factor $e^{\kappa t}$:

$$e^{\kappa t}\,\mathrm{d}X_t = -\kappa\, e^{\kappa t} X_t \,\mathrm{d}t + e^{\kappa t}\sigma_X(t)\,\mathrm{d}W_t.$$

On the other hand, the differential of the product (with $e^{\kappa t}$ being of finite variation) reads:

$$\mathrm{d}\big(e^{\kappa t} X_t\big) = \kappa\, e^{\kappa t} X_t \,\mathrm{d}t + e^{\kappa t}\,\mathrm{d}X_t.$$

Substituting the first expression, the terms in $\kappa\, e^{\kappa t} X_t \,\mathrm{d}t$ cancel, in integration variable u:

$$\mathrm{d}\big(e^{\kappa u} X_u\big) = e^{\kappa u}\sigma_X(u)\,\mathrm{d}W_u.$$

Integrating over $[s,t]$ with $s < t$:

$$e^{\kappa t} X_t - e^{\kappa s} X_s = \int_s^t e^{\kappa u}\sigma_X(u)\,\mathrm{d}W_u,$$

whence, rearranging the terms, the solution equation (7):

$$X_t = e^{-\kappa(t-s)} X_s + \int_s^t e^{-\kappa(t-u)}\sigma_X(u)\,\mathrm{d}W_u.$$

The same manipulation applied to $T_t = \theta(t) + X_t$, whose dynamics are $\mathrm{d}T_t = \big(\theta'(t) + \kappa(\theta(t) - T_t)\big)\mathrm{d}t + \sigma_X(t)\,\mathrm{d}W_t$, yields the solution equation (8):

$$T_t = \theta(t) + \big(T_s - \theta(s)\big)e^{-\kappa(t-s)} + \int_s^t e^{-\kappa(t-u)}\sigma_X(u)\,\mathrm{d}W_u.$$

Appendix I Closed-form price of the HDD call under the OU model

We detail the closed-form valuation of the HDD call under the OU model, summarized in section 6. The desk values at date s over the window $[t_1, t_n]$, with $s < t_1 < \cdots < t_n$. Φ and φ denote the cumulative distribution function and the density of the $\mathcal{N}(0,1)$ distribution.

Dynamics under $\mathbb{P}$ and under $\mathbb{Q}$. Solving the SDE (appendix H) gives, under $\mathbb{P}$,

$$T_t = \theta(t) + \big(T_s - \theta(s)\big)e^{-\kappa(t-s)} + \int_s^t e^{-\kappa(t-u)}\sigma_X(u)\,\mathrm{d}W_u.$$

Since the market is incomplete, we introduce the market price of risk λ: $V_t = W_t + \lambda t$ defines a Brownian motion under $\mathbb{Q}$, so that $\mathrm{d}W_t = \mathrm{d}V_t - \lambda\,\mathrm{d}t$. The temperature dynamics thus acquire an additional drift $-\lambda\sigma_X(t)\,\mathrm{d}t$,

$$\mathrm{d}T_t = \big(\theta'(t) + \kappa(\theta(t) - T_t) - \lambda\sigma_X(t)\big)\mathrm{d}t + \sigma_X(t)\,\mathrm{d}V_t,$$

and the same integrating-factor method as in appendix H yields the solution under $\mathbb{Q}$:

$$T_t = \theta(t) + \big(T_s - \theta(s)\big)e^{-\kappa(t-s)} + \int_s^t e^{-\kappa(t-u)}\sigma_X(u)\,\mathrm{d}W_u - \lambda\int_s^t e^{-\kappa(t-u)}\sigma_X(u)\,\mathrm{d}u.$$

Conditional mean. The stochastic integral has zero expectation (deterministic integrand) and $T_s - \theta(s)$ is $\mathscr{F}_s$-measurable. Since σ_X is assumed constant over each day, we split the drift into daily integrals and compute, for $u \in [j-1, j]$,

$$\int_{j-1}^{j} e^{-\kappa(t-u)}\,\mathrm{d}u = e^{-\kappa t}\int_{j-1}^{j} e^{\kappa u}\,\mathrm{d}u = \frac{e^{-\kappa t}}{\kappa}\left(e^{\kappa j} - e^{\kappa(j-1)}\right) = \frac{1-e^{-\kappa}}{\kappa}\,e^{-\kappa(t-j)}.$$

The drift term therefore equals $\dfrac{\lambda(1-e^{-\kappa})}{\kappa}\sum_{j=s+1}^{t}\sigma_X(j)\,e^{-\kappa(t-j)}$, which establishes the conditional mean equation (18).

Conditional variance. The deterministic part does not contribute to the variance, by Itô's isometry (under $\mathbb{Q}$, the stochastic integral is unchanged), and with σ_X^2 constant per day,

$$\operatorname{Var}(T_t \mid \mathscr{F}_s) = \int_s^t e^{-2\kappa(t-u)}\sigma_X^2(u)\,\mathrm{d}u = \sum_{j=s+1}^{t}\sigma_X^2(j)\int_{j-1}^{j} e^{-2\kappa(t-u)}\,\mathrm{d}u,$$

where, by the same computation as above,

$$\int_{j-1}^{j} e^{-2\kappa(t-u)}\,\mathrm{d}u = \frac{e^{-2\kappa t}}{2\kappa}\left(e^{2\kappa j} - e^{2\kappa(j-1)}\right) = \frac{1-e^{-2\kappa}}{2\kappa}\,e^{-2\kappa(t-j)},$$

which establishes the first part of equation (19).

Conditional covariance. For $s < t_i < t_j$, let $\widetilde{T}_t = \int_s^t e^{-\kappa(t-u)}\sigma_X(u)\,\mathrm{d}W_u$ denote the centered stochastic part. Decomposing $T_{t_i} = \mathbb{E}^{\mathbb{Q}}[T_{t_i} \mid \mathscr{F}_s] + \widetilde{T}_{t_i}$, the expansion of $\operatorname{Var}(T_{t_i} \mid \mathscr{F}_s) = \mathbb{E}^{\mathbb{Q}}[(\mathbb{E}^{\mathbb{Q}}[T_{t_i} \mid \mathscr{F}_s] + \widetilde{T}_{t_i})^2 \mid \mathscr{F}_s] - \mathbb{E}^{\mathbb{Q}}[T_{t_i} \mid \mathscr{F}_s]^2$ makes the cross term vanish (since $\mathbb{E}^{\mathbb{Q}}[\widetilde{T}_{t_i} \mid \mathscr{F}_s] = 0$), whence $\mathbb{E}^{\mathbb{Q}}[\widetilde{T}_{t_i}^2 \mid \mathscr{F}_s] = V(t_i, s)$ and $\operatorname{Cov}(T_{t_i}, T_{t_j} \mid \mathscr{F}_s) = \mathbb{E}^{\mathbb{Q}}[\widetilde{T}_{t_i}\widetilde{T}_{t_j} \mid \mathscr{F}_s]$. Splitting the integral at t_i,

$$\widetilde{T}_{t_j} = e^{-\kappa(t_j-t_i)}\,\widetilde{T}_{t_i} + I_{i,j}, \qquad I_{i,j} = \int_{t_i}^{t_j} e^{-\kappa(t_j-u)}\sigma_X(u)\,\mathrm{d}W_u.$$

The increment $I_{i,j}$ is independent of $\mathscr{F}_{t_i}$ and has zero expectation, so by the tower property,

$$\mathbb{E}^{\mathbb{Q}}[\widetilde{T}_{t_i}\, I_{i,j} \mid \mathscr{F}_s] = \mathbb{E}^{\mathbb{Q}}\big[\widetilde{T}_{t_i}\,\mathbb{E}^{\mathbb{Q}}[I_{i,j} \mid \mathscr{F}_{t_i}] \mid \mathscr{F}_s\big] = 0.$$

There remains $\operatorname{Cov}(T_{t_i}, T_{t_j} \mid \mathscr{F}_s) = e^{-\kappa(t_j-t_i)}\,\mathbb{E}^{\mathbb{Q}}[\widetilde{T}_{t_i}^2 \mid \mathscr{F}_s] = e^{-\kappa(t_j-t_i)}\,V(t_i, s)$, that is, the second part of equation (19).

Gaussian parameters of the index. Under $(\mathscr{H})$, $H_n = n\,T_{\mathrm{ref}} - \sum_{i=1}^n T_{t_i}$, whence $\mu_s(n) = n\,T_{\mathrm{ref}} - \sum_{i=1}^n \mathbb{E}^{\mathbb{Q}}[T_{t_i} \mid \mathscr{F}_s]$. For the variance,

$$\operatorname{Var}\Big(\sum_{i=1}^{n} T_{t_i}\,\Big|\,\mathscr{F}_s\Big) = \sum_{i=1}^{n} V(t_i, s) + 2\sum_{i<j}\operatorname{Cov}(T_{t_i}, T_{t_j} \mid \mathscr{F}_s) = \sum_{i=1}^{n} V(t_i, s)\Big[1 + 2\sum_{j=i+1}^{n} e^{-\kappa(t_j-t_i)}\Big].$$

Since the days are spaced one step apart ($t_j - t_i = j - i$), the inner sum is geometric:

$$\sum_{j=i+1}^{n} e^{-\kappa(t_j-t_i)} = \sum_{k=1}^{n-i} e^{-\kappa k} = e^{-\kappa}\,\frac{1-e^{-\kappa(n-i)}}{1-e^{-\kappa}},$$

which yields the form equation (20).

Price of the Gaussian call. With $\tau = t_n - s$ and $f_{H_n}(x) = \dfrac{1}{\Lambda_s(n)}\,\varphi\Big(\dfrac{x-\mu_s(n)}{\Lambda_s(n)}\Big)$, the payoff equals $x-K$ only for $x \geq K$:

$$C_s(n) = \alpha\, e^{-r\tau}\int_K^{+\infty}(x-K)\,f_{H_n}(x)\,\mathrm{d}x = \alpha\, e^{-r\tau}\Big[\underbrace{\int_K^{+\infty} x\, f_{H_n}(x)\,\mathrm{d}x}_{(1)} - K\underbrace{\int_K^{+\infty} f_{H_n}(x)\,\mathrm{d}x}_{(2)}\Big].$$

Let $\eta_s(n) = \dfrac{K-\mu_s(n)}{\Lambda_s(n)}$. Term (2) is a tail probability: $(2) = \mathbb{Q}(H_n \geq K) = \Phi(-\eta_s(n))$. For (1), the change of variable $z = \dfrac{x-\mu_s(n)}{\Lambda_s(n)}$ (lower bound $z = \eta_s(n)$) gives

$$(1) = \Lambda_s(n)\int_{\eta_s(n)}^{+\infty} z\,\varphi(z)\,\mathrm{d}z + \mu_s(n)\int_{\eta_s(n)}^{+\infty}\varphi(z)\,\mathrm{d}z.$$

The identity $\varphi'(z) = -z\varphi(z)$ gives $\int_{\eta_s(n)}^{+\infty} z\varphi(z)\,\mathrm{d}z = \big[-\varphi(z)\big]_{\eta_s(n)}^{+\infty} = \varphi(\eta_s(n))$, and $\int_{\eta_s(n)}^{+\infty}\varphi(z)\,\mathrm{d}z = \Phi(-\eta_s(n))$, whence $(1) = \Lambda_s(n)\,\varphi(\eta_s(n)) + \mu_s(n)\,\Phi(-\eta_s(n))$. Collecting $(1) - K\,(2)$:

$$C_s(n) = \alpha\, e^{-r\tau}\big[(\mu_s(n)-K)\,\Phi(-\eta_s(n)) + \Lambda_s(n)\,\varphi(\eta_s(n))\big],$$

that is, the closed-form price equation (21).

Under the same hypothesis $(\mathscr{H})$, the put has payoff $\alpha\,(K-H_n)^+$. A cumulative HDD index is non-negative, so its price integrates the density from 0 to K:

$$P_s(n) = \alpha\, e^{-r\tau}\int_0^K (K-x)\,f_{H_n}(x)\,\mathrm{d}x.$$

The same change of variable now runs from the lower bound $z = -\mu_s(n)/\Lambda_s(n)$ (image of $x = 0$) up to $\eta_s(n)$, which leaves a boundary term at 0:

$$\boxed{P_s(n) = \alpha\, e^{-r\tau}\Big[(K-\mu_s(n))\big(\Phi(\eta_s(n)) - \Phi(-\mu_s(n)/\Lambda_s(n))\big) + \Lambda_s(n)\big(\varphi(\eta_s(n)) - \varphi(\mu_s(n)/\Lambda_s(n))\big)\Big]} \tag{I.1}$$

Under $(\mathscr{H})$ the index is almost surely positive ($\mu_s(n) \gg \Lambda_s(n)$), so the two terms at 0 are negligible and the put reduces to the symmetric form $(K-\mu_s(n))\,\Phi(\eta_s(n)) + \Lambda_s(n)\,\varphi(\eta_s(n))$.

Appendix J Feature Correlation Matrix

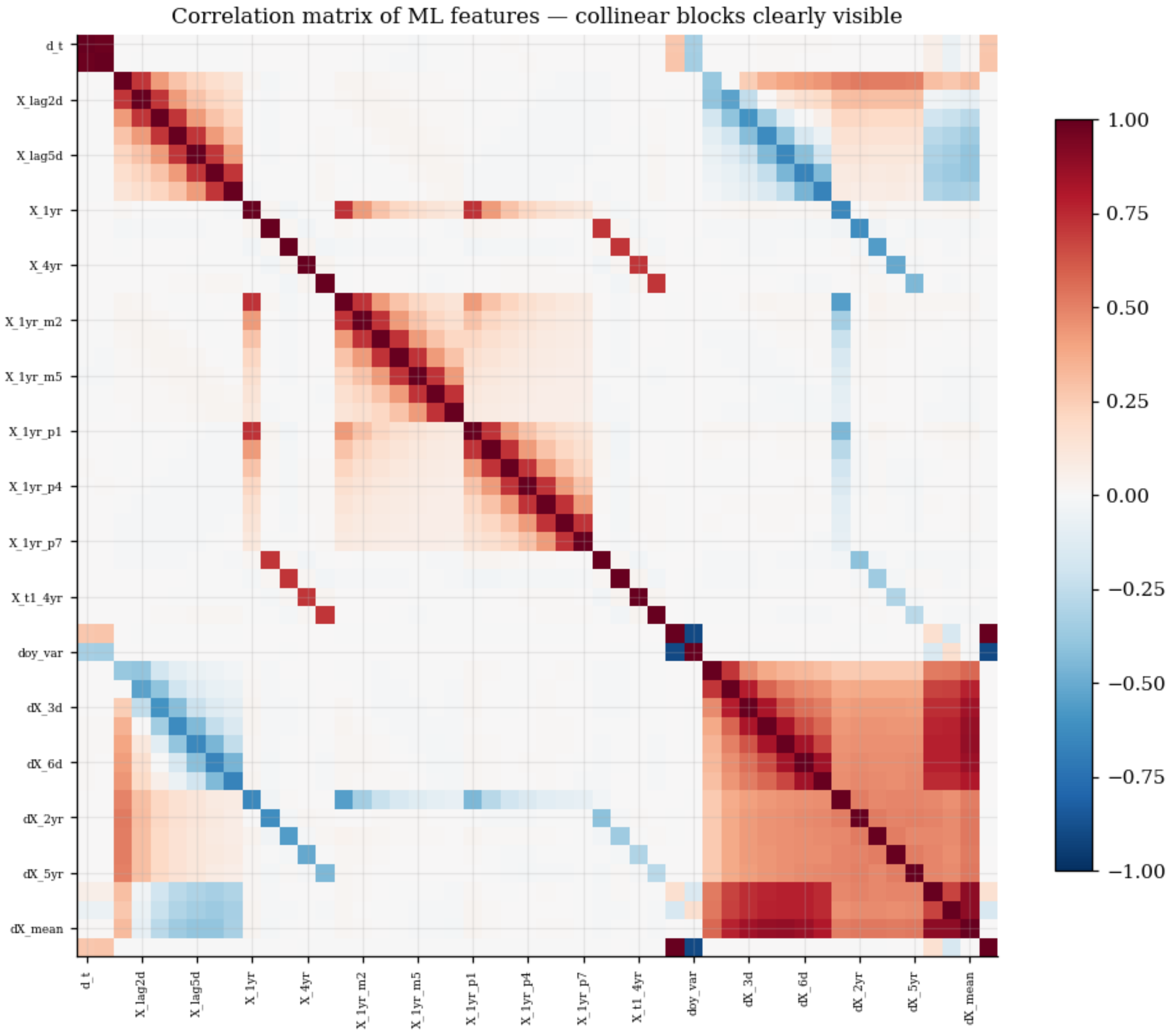

Figure J.1: Correlation matrix of the 50 learning features (recent lags, previous-year values, differences ΔX).

Appendix K Mathematical background of the learning models

We recall here the formulations of the three selected models (section 5). All of them minimize a regularized empirical risk over the sample $\{(x_i, y_i)\}_{i=1}^{N}$ $(y_i = X_{t+1})$ and differ only in their hypothesis class and their regularization.

Ridge. Ridge regression minimizes the L^2-penalized squared error, which admits an explicit solution:

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 = \left(X^\top X + \lambda I\right)^{-1} X^\top y.$$

The term λI makes $X^\top X + \lambda I$ well conditioned even when $X^\top X$ is nearly singular (collinear features): it bounds the coefficients and reduces the variance of the estimator, at the cost of a slight bias. It is the canonical remedy for the multicollinearity that makes OLS diverge.

Random Forest. A random forest aggregates B regression trees, each trained on a bootstrap sample and considering, at every split, only a random subset of $m \le d$ features. The prediction is the average $\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$. If the trees have variance σ^2 and pairwise correlation ρ, the variance of their average is

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2.$$

Reducing ρ is therefore essential: this is precisely the role of column subsampling, which decorrelates the trees and makes the forest robust to strongly correlated features.

XGBoost. *Gradient boosting* builds an additive model $\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$ through successive additions of trees. At step k, one minimizes the regularized objective

$$\sum_i \ell\big(y_i,\ \hat{y}_i^{(k-1)} + f_k(x_i)\big) + \Omega(f_k), \qquad \Omega(f) = \gamma T + \tfrac{1}{2}\lambda\|w\|_2^2,$$

where T is the number of leaves of the tree and w their weights. A second-order Taylor expansion (gradients g_i and Hessians h_i of the loss) yields the optimal weight per leaf,

$$w_j^\star = -\frac{\sum_{i\in j} g_i}{\sum_{i\in j} h_i + \lambda},$$

along with a gain criterion that guides the choice of splits. Boosting reduces the bias (where the forest reduces the variance), and the explicit regularization Ω controls the complexity, which explains its strong performance on tabular data.

Appendix L Nested Cross-Validation

We detail here the training protocol summarized in section 5. Using a single cross-validation both to choose the hyperparameters and to estimate performance leads to an optimistically biased estimate: the folds used to select the model have also been used to judge it. Nested cross-validation avoids this pitfall by separating the two roles. The outer loop splits the data into k_{out} folds. For each of them, the remainder serves as the training set and the fold itself as the held-out block. On each outer training set, an inner loop of k_{in} folds performs a grid search to select the hyperparameters, which therefore never "see" the corresponding held-out block.

One cannot randomly permute the observations of a time series without leaking future information into the past. The folds therefore follow chronological order, in an expanding window (*forward chaining*): fold i trains on the entire history up to a given date, then validates on the subsequent temporal block, which preserves causality (figure 12). Since the hyperparameters may vary from one outer fold to another, we retain their median (or their mode for discrete parameters) over the k_{out} folds, which is more robust than a choice drawn from a single fold. The final model is then retrained on the entire training window with these values.

Tuning uses as its metric the MAPE in Kelvin ($T + 273.15$), and all the preprocessing (imputation, standardization, target transformation) is learned within each fold in order to avoid data leakage. Table L.1 gives the grids explored by the inner search, identical for the seven cities.

Table L.1: Hyperparameter grids explored in the inner loop.

Model	Search grid
Ridge	λ: 300 log-spaced values from 10^{-3} to 10^4.
Random Forest	`n_estimators` $\in \{100, 200, 400, 700\}$; `max_depth` $\in \{2, 4, 10, 20\}$; `min_samples_leaf` $\in \{1, 5, 20, 50\}$; `max_features` $\in$ {all, $\sqrt{\cdot}$, $\log_2$, 0.3} (256 combinations).
XGBoost	`n_estimators` $\in \{50, 100, 200, 400\}$; `max_depth` $\in \{2, \ldots, 6\}$; `learning_rate` $\in \{0.05; 0.1; 0.2\}$; `subsample` $\in \{0.8; 1\}$; `colsample_bytree` $\in \{0.8; 1\}$ (240 combinations).

Appendix M HDD Distribution by City and Asymmetry of the Evaluation

To complement the evaluation of the HDD forecast (figure 13), figure M.1 shows the distribution of the daily HDD $\text{HDD}_t = \max(T_{\text{ref}} - T_t, 0)$ for the seven cities. The floor at zero, which nullifies any day warmer than the reference threshold, makes the index markedly asymmetric: to the mass concentrated at 0 (the mild days) is added a right tail (the cold spells), which is all the more pronounced as the city is temperate (skewness from +0.3 in Seattle up to +1.6 in Dallas).

This asymmetry has a direct consequence for the evaluation. Since the HDD is a *convex and truncated* function of temperature, an overestimation and an underestimation of T_t of the same magnitude do not carry over to the index in the same way: around the threshold T_{ref}, forecasting too warm shaves the HDD down to zero, whereas forecasting too cold increases it linearly. A temperature error therefore does not carry the same weight depending on its sign, and the error on the paid index (APE) inherits this. This is one more reason to judge the models on the index itself rather than on the temperature trajectory alone.

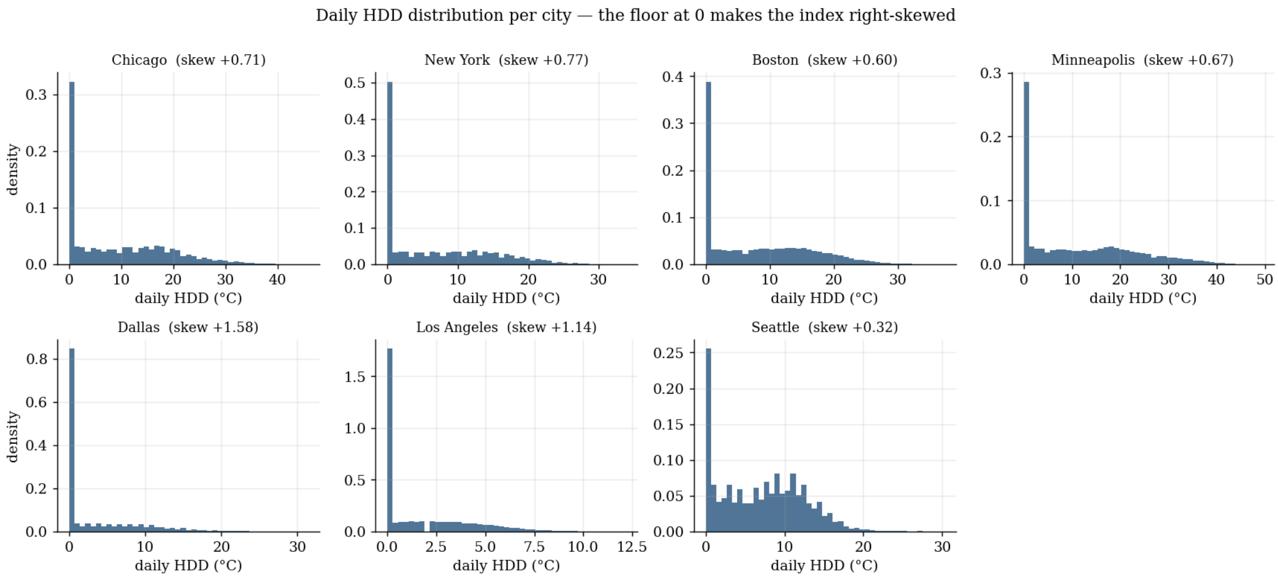

Figure M.1: Distribution of the daily HDD $\max(T_{\text{ref}} - T_t, 0)$ by city (base 18 °C, 1950–2025), with the skewness coefficient for each city.

Appendix N Out-of-sample evaluation: complementary metrics

This appendix complements figure 13 (HDD APE for Chicago) from three angles: the per-city detail, the fidelity of the upstream temperature trajectory, and two variants of the HDD error that control the robustness of the ranking.

The metrics used here are defined as follows. The MAE (mean absolute error), applied both to the daily temperature and to the cumulative HDD, and the MAPE in Kelvin, which serves only to tune the hyperparameters (equation (13)), are written

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y}_i|, \qquad \mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|T_i - \widehat{T}_i|}{T_i + 273.15}, \tag{N.1}$$

where y denotes the quantity being evaluated (temperature in °C, cumulative HDD in °C·days). They complement the HDD APE defined in the body of the text (equation (14)).

Figure N.1 reproduces the HDD APE by horizon for the seven cities. The Chicago profile generalizes (the temperature models ahead of the burn at close horizons, the burn becoming competitive or even better at a one-season lead time), with amplitudes that vary from one city to another.

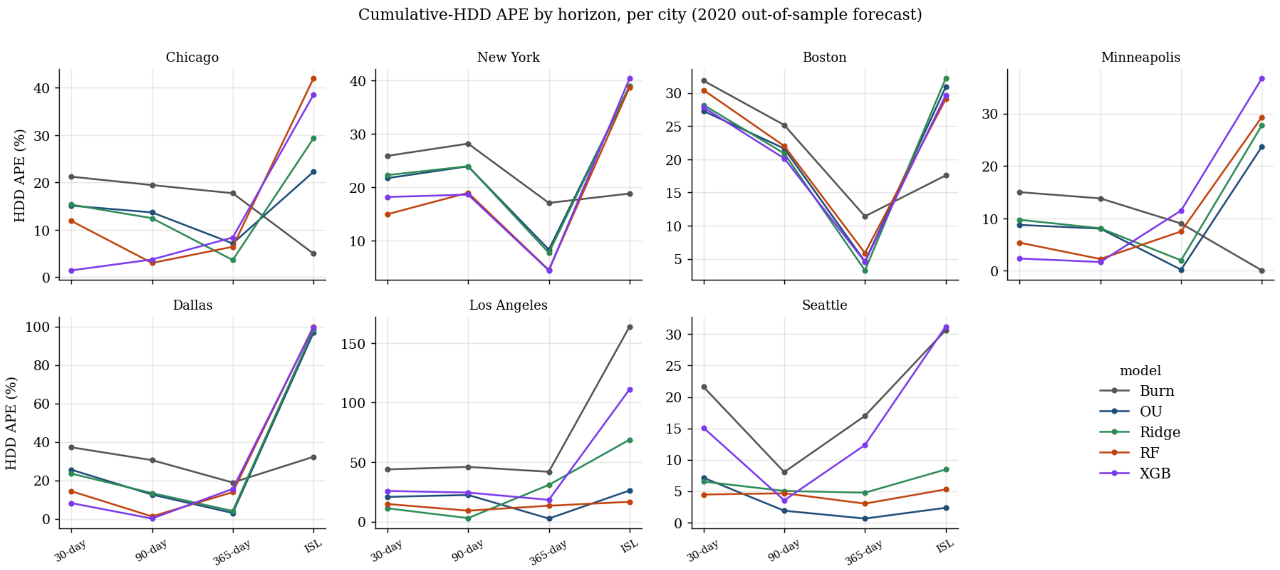

Figure N.1: APE of the cumulative HDD by horizon and by city (out-of-sample forecast from January 1, 2020), for the climatological burn and the four models.

Upstream of the index, figure N.2 gives the mean absolute error (MAE) of the daily *temperature*, averaged over the five test winters. The models lead climatology only by about 0.3 °C in the short term, a gap erased at long horizons. The level depends mostly on the city (difficult continental versus predictable oceanic), far more than on the model.

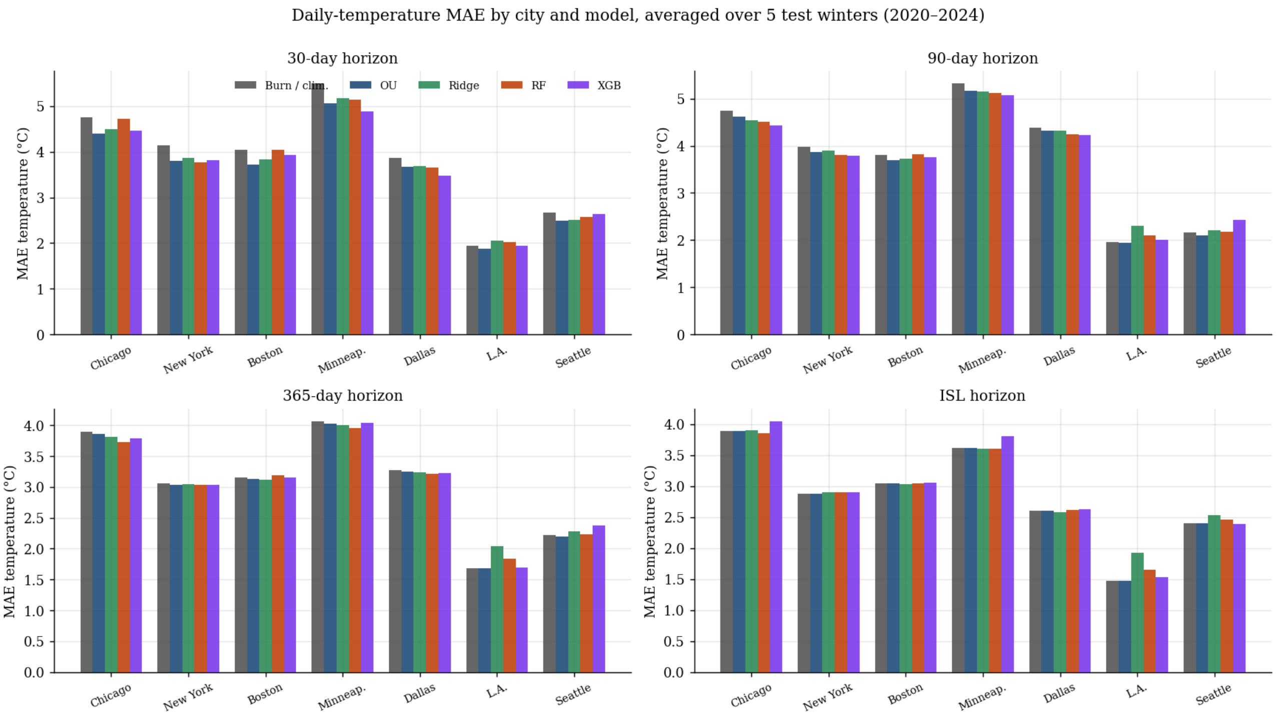

Figure N.2: MAE of the daily temperature (°C), by city and by model, averaged over the five test winters, at four horizons. Here the burn corresponds to the climatological temperature θ.

Since the APE inflates when the HDD of the window is low, two variants confirm that the ranking does not depend on it. The MAE of the cumulative HDD (figure N.3, in °C·days, averaged over five winters) removes any denominator. Conversely, the error normalized by the climatological standard deviation σ_H (figure N.4) makes the horizons comparable with one another. Both give the same reading as the APE: the models win at close horizons, the burn becomes competitive or even better at a one-season lead time, where the recursive forecast has lost all conditional information.

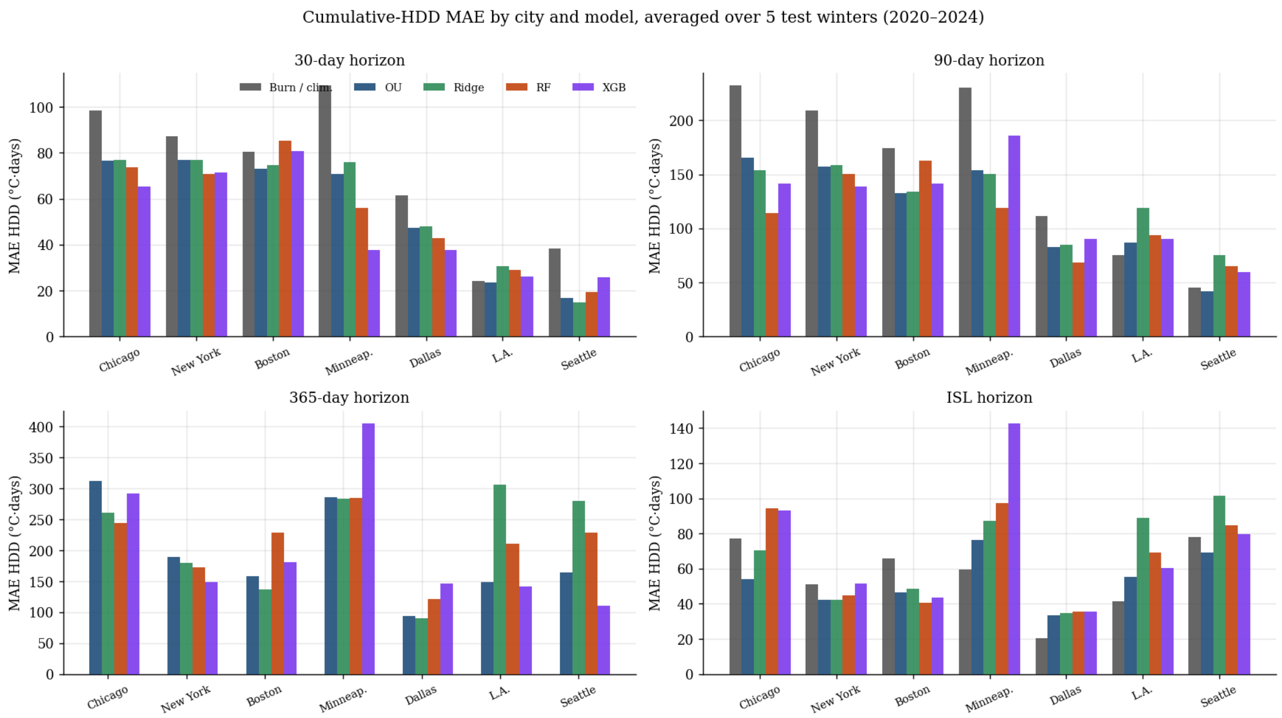

Figure N.3: MAE of the cumulative HDD (°C·days), by city and by model, averaged over the five test winters, at four horizons. The burn is the mean HDD of the previous winters.

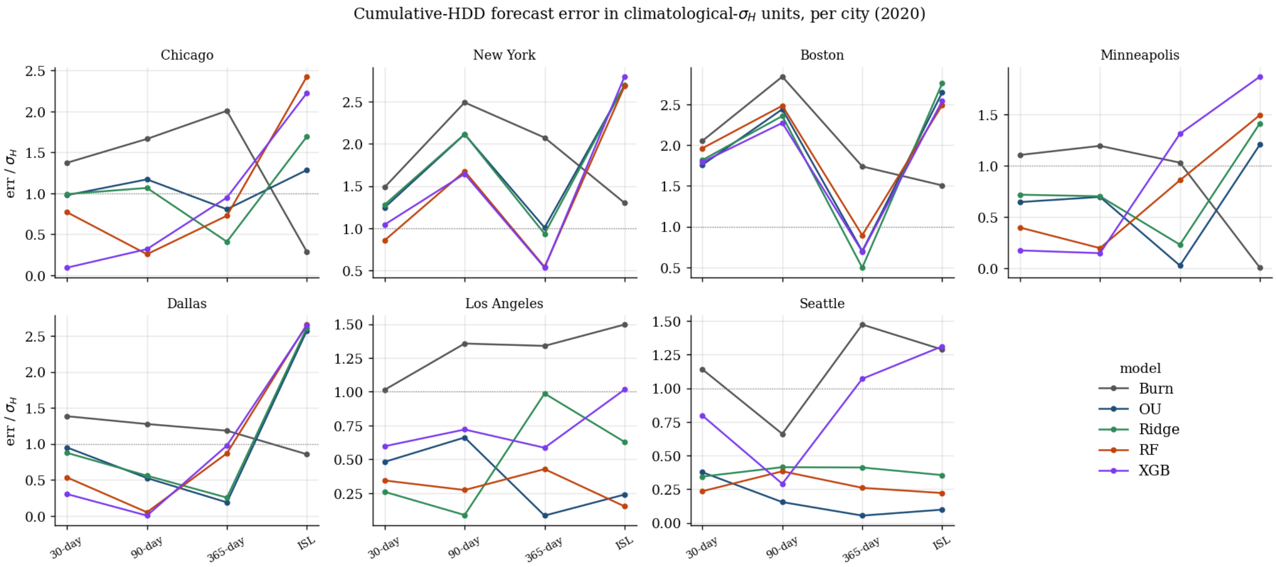

Figure N.4: Forecast error of the cumulative HDD by horizon and by city, normalized by the climatological standard deviation σ_H (out-of-sample forecast from 2020), for the burn and the four models.

Appendix O Contest strikes and payoffs

The strike of each contract is the desk burn, $K_{c,w}$, the average cumulative HDD of the preceding winters (equation (23)). figure O.1 gives its values by city and by test winter. Since HDD has no common scale from one city to another (Los Angeles hovers around 440 degree-days, Minneapolis above 3200), we present them as a plain table, without a color gradient.

Burn strike $K_{c,w}$ (HDD)

	20/21	21/22	22/23	23/24	24/25
Boston	2317	2294	2332	2301	2253
Chicago	2706	2660	2706	2682	2571
Dallas	1023	1021	1023	1021	978
L.A.	442	442	428	440	456
Minneapolis	3264	3218	3297	3298	3168
New York	2050	2019	2055	2015	1957
Seattle	1701	1686	1674	1689	1687

Figure O.1: Burn strike $K_{c,w}$ (cumulative HDD, base 18 °C) by city and by test winter.

Once the winter is over, each contract pays out its realized payoff $Y_{c,w}$ (figure O.2). It is this quantity, denominated in dollars and therefore comparable from one city to another, that serves as the benchmark for the contest: the pricing error $e_m(c, w) = \Pi_m(c, w) - Y_{c,w}$ measures the deviation of each model from this payoff. Since most of the 2020–2025 test winters were mild, the realized HDD stays below the strike and 22 of the 35 contracts expire at zero.

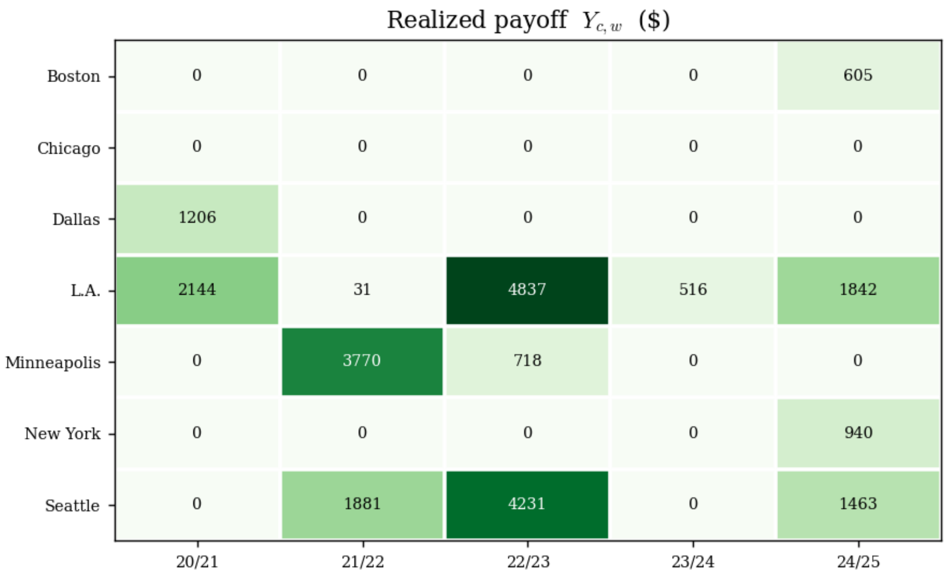

Figure O.2: Realized payoff $Y_{c,w}$ (in $) by city and by test winter.