Artificial Intelligence Review (2026) 59:144
https://doi.org/10.1007/s10462-026-11539-0

Machine learning for temperature forecasting in weather derivatives

Nee Barnor[1] **· Michael Kampouridis**[1] **· Panagiotis Kanellopoulos**[1]

Received: 11 May 2025 / Accepted: 12 March 2026 / Published online: 25 March 2026

Abstract
Weather derivatives enable businesses to hedge weather related fluctuations in volume and revenue. Pricing these contracts requires forecasting the underlying weather variable over the relevant risk window. Temperature represents the largest share of exchange traded weather derivatives, yet despite more than two decades of research, there remains little consensus on the most effective modelling approach across diverse locations, with most studies focusing on location specific performance. We evaluate whether machine learning methods can provide a more generalisable solution for temperature forecasting in the context of weather derivative pricing. Our study conducts an extensive comparison of 17 models across 65 global locations, including 12 machine learning approaches (linear models, simple tree-based models, sequential and randomised ensemble tree models, neural network models and support vector machines), 2 Ornstein–Uhlenbeck benchmarks, and 3 classical time series models. We also assess performance over the specific pricing windows used in real world contracts, offering practical value to traders and investors. Model accuracy is evaluated using absolute percentage error, along with root mean square error, mean absolute error, and weighted absolute percentage error. Our results show that machine learning models statistically and significantly outperform benchmark approaches at forecast horizons aligned with typical hedging needs; for example, Random Forest reduces cumulative seasonal temperature forecasting error predicted 4.5 months ahead more than any alternative model. We further document systematic geographical variation in model performance and demonstrate economic relevance through a pricing experiment.

Keywords Weather risk · Temperature derivatives · Ornstein Uhlenbeck · Machine learning

Michael Kampouridis and Panagiotis Kanellopoulos have contributed equally to this work.

Extended author information available on the last page of the article

1 Introduction

Variability in the weather presents a volume risk to businesses in industries such as agriculture, energy, tourism and transportation (Alexandridis and Zapranis 2013b) that either depend directly on the weather for production or the use of their services is significantly influenced by the weather. This weather exposure is anticipated to increase as, according to the report of climate-related market risk by the US Commodity Futures Trading Commission (CFT) and (2020), rising global average temperatures could impair the productive capacity of economies and erode their ability to generate employment, income, and opportunity.

One avenue available to businesses for hedging weather risk is conventional weather insurance. This, however, does not provide cover against business volume risk since proof of business loss is required before a pay-out is made. Weather derivatives, which are risk management financial products written on an underlying weather index, have arisen to cater for this gap; they do not require proof of business loss before pay-out and thus present a useful hedge against volume risk for businesses.

Pricing weather derivatives can be broadly grouped into two approaches; historical methods including Actuarial (Jewson 2004) and Burn Analysis (Dischel 1999) that price the derivative directly and statistical methods that model the underlying dynamics of the weather variable itself to predict its future paths and then use the future predictions to price the derivative.

Research in the statistical approach to pricing weather derivatives poses two key challenges. The first is how to model the behaviour of the underlying weather variable, on which the derivative contracts are priced, in order to predict its future values whereas the second is how to then convert the predictions to financial contract prices. Temperature derivatives are the most popular by traded volume and accurately predicting the future evolution of temperature is critical to pricing this type of weather derivative. Temperature is however, notoriously volatile, difficult to predict and, with climate change, undergoing fundamental changes (Alexandridis et al. 2017). The weather derivatives market is, furthermore, incomplete because traders cannot physically hold or trade the underlying asset (temperature) underpinning the financial instrument (Alexandridis et al. 2017). This makes the development of temperature derivative prices a challenge and has contributed to its limited availability worldwide, leaving numerous locations and businesses exposed to weather risk bearing the costs of weather variation themselves. The difficulty in predicting temperatures and the subsequent pricing of weather derivatives from the predicted temperatures has driven years of research in the field without a consensus on the best temperature derivative pricing method that can be applied universally. Several recent researchers continue to point to this lack of consensus and call for further research to help improve the adoption of weather derivatives (see also Sect. 2.2.5 where we elaborate on this).

1.1 Our contribution

The statistical approach to pricing weather derivatives has been found to be more accurate (Alexandridis and Zapranis 2013b) and in this paper, we establish that machine learning algorithms can do equally well, if not better, and, importantly, can generalise better and scale more efficiently to a greater number of locations than the current methods available

under the statistical approach including the stochastic process models that dominate the literature (see Sect. 2.2 where we go into greater detail).

Our work advances the field in the following ways: (i) it evaluates the performance of a set of 12 commonly used machine learning algorithms, while most existing studies typically consider two-three; (ii) it compares the performance of these algorithms with five state-of-the-art approaches, encompassing most of the clusters of research on temperature prediction; (iii) it analyses model performance on discrete forecasting horizons, reflecting the various risk profiles of potential participants in the temperature derivatives market; (iv) it tailors the analysis on business needs, by including forecasting horizons that better represent the pricing window for which temperature derivatives are sold, rather than adopting traditional forecasting horizons that are seldom used for pricing weather derivatives; and (v) it explores how model performance is influenced by the geographic location of the weather station.

The rest of this paper is organised as follows, Sect. 2 gives a background to temperature derivatives and summarises past research, Sect. 3 explains the methodology we followed including our experimental set up, Sect. 4 explains our results and provides a discussion on key findings after which we summarise our conclusions and offer suggestions for future research in Sect. 5.

2 Background and related work

In this section, we provide the necessary background, including definitions and key concepts, in weather derivatives together with related work on the topic. We begin with an introduction to the weather derivatives market, detail the main methods cited in the literature for modelling temperature for the purpose of pricing weather derivatives and, finally, we conclude with arguments in favour of further research in the field.

2.1 The weather derivatives market

Weather derivatives started mainly in the United States as energy companies, whose business volumes are highly correlated with the weather, needed to hedge this volume risk and reduce the volatility of their earnings. Energy companies already had tools for hedging against price risk, but the deregulation of the energy market exposed them to volume risks that highly correlated with the weather hence the creation of weather derivatives (Müller and Grandi 2000). Today, the weather derivatives market is dominated by contracts written against temperature risk and the Chicago Mercantile Exchange (CME) offers temperature products based on three temperature indices, Heating Degree Days (HDD), Cooling Degree Days (CDD) and Cumulative Average Temperature (CAT). A HDD is the number of degrees by which the daily temperature is below a base temperature, and a CDD is the number of degrees by which the daily temperature is above the base temperature. The base temperature is usually 65 (°F) in the USA and 18 (°C) in Europe and Japan. HDDs and CDDs cannot be negative, and usually, they are accumulated over a month or over a season (Alexandridis and Zapranis 2013). The CAT index is the sum of Daily Average Temperatures (DATs) over the contract period and given a time interval $[\tau_1, \tau_2]$ can be expressed as:

$$\text{CAT} = \int_{\tau_1}^{\tau_2} T_s \, ds, \tag{1}$$

where T_s is the daily temperature on a given day s. The daily heating and cooling degree days are expressed as:

$$\text{Daily HDD} = \max(0, \text{base temperature} - \text{daily average temperature}),$$

$$\text{Daily CDD} = \max(0, \text{daily average temperature} - \text{base temperature}).$$

For this study we focused on modelling and forecasting the CAT and the HDD metrics to cover both cumulative and reference based temperatures used in temperature derivative pricing. Researchers generally select one of either the cumulative or reference-based metrics when undertaking temperature pricing research; see (Alaton et al. 2002; Benth 2003; Šaltyte Benth and Benth 2012).

2.2 Temperature modelling

Weather derivatives are priced on a weather index thus before prices can be computed, the underlying weather variable has to be modelled and predicted. Temperature derivatives dominate the weather market, hence this study focuses on modelling temperature for the purpose of pricing temperature derivatives to be better applicable to as many areas that need to hedge weather risk as possible. Temperature derivatives can be priced on the temperature indices directly but modelling and forecasting the daily temperature before computing the index for pricing can in principle be more accurate, as a lot of information can be lost when calculating the indices (Jewson 2004). Deriving an accurate daily model is however not simple and Jewson and Brix (2005) explain how small mis-specifications can lead to large mispricing in weather contracts. Daily models also allow for the use of the same model for the pricing of different contracts whilst a new model must be built if one wants to price different indices with an index model approach.

Weather derivatives can also be susceptible to geographical basis risk, arising from the difference (in weather) between the weather measuring station and the location of the business being hedged. Early researchers, including Geyser (2004), noted the restriction that basis risk placed on the adoption of weather derivatives especially in business like agriculture. Geyser (2004) highlighted the preference of farmers to have weather derivatives written on weather as close to their farms as possible. This understanding has driven a lot of research not only into basis risk itself but weather modelling methods that can better generalise to other locations. Yang et al. (2009) in their work empirically analysed the hedging efficiency of weather derivatives with respect to basis risk. They analysed 76 locations in the US and encouraged further research in Europe and other areas to validate their findings. Elias et al. (2014) also compared four variations of regime switching models in temperature forecasting for the pricing of weather derivatives and noted how for the same location the models performed differently highlighting further the need for methods with an ambition of wide adoption to also have cross location generalisability. Zong and Ender (2016) directly tried to answer the question of location generalisability to increase the risk management efficiency of weather derivatives in their work by building a city specific model and a climatic-

zone model covering several locations and compared their risk hedging performance. Their results favoured the climatic model for the Chinese cities they looked at but their research also highlighted the importance of models that can be used for several locations especially when introducing weather derivatives to new markets.

Interest in temperature models that can perform well across different locations has persisted in recent research where more recently Oettli et al. (2022) in their attempt to forecast monthly temperature in the Kanto region of Japan employed a hybrid approach involving an ensemble ML forecast combined with a dynamical meteorological model and found this outperformed the pure dynamical models for two-month ahead predictions but also highlighted the location-specific limitations in weather forecasting. Oettli et al. (2022) included data from 102 stations in the Kanto region in their work. Wenjie et al. (2024) proposed a composite weather index to try and solve the location generalisation challenge. They selected 9 weather stations to cover all the sub-climatic zones in China and created a composite temperature index from which they proposed highly flexible and adaptable prices for hedging weather risk in China. Tallarico and Olivares (2024) admit the need for models to generalise in their work where they build a model for Hong Kong but then validate this with data from both Toronto and Chicago in a bid to directly tackle the challenge of cross-location generalisation of weather derivative models. Their study found neural networks better at adaptive learning of weather dynamics across the different locations compared to static models. Tallarico and Olivares (2024) encouraged further testing of their formulation in more regions as well as extending the forecast horizon beyond one month as future improvements to their work. Most recently Cheng et al. (2025) analysed tail risk in weather derivatives using data from 13 US cities and highlighted the need for modelling and pricing approaches that can generalise more easily to other locations by suggesting extensions of the approach to other non-US locations as an improvement to their work.

We also remark that temperature modelling for the purposes of pricing derivatives differs from short term meteorological forecasts that focus on short term daily or intra day temperature predictions and primarily rely on physical models that require a lot of computational power to run. In contrast, statistical models are more popular in the temperature derivative literature and include the Ornstein-uhlenback methods that model the daily temperature as a stochastic process but also time series models including classical time series methods as well as deep machine learning models.

In the following sections we will briefly describe an evolution focused view of the various methods that have been employed in modelling and forecasting temperature for the purpose of pricing weather derivatives since their inception while giving greater attention to the more popular benchmarks in literature.

2.2.1 Classical time series models

Before interest in temperature modelling arose for the purpose of pricing weather derivatives, researchers had been using classical time series techniques to model and predict future values of temperature. Tol (1996) argued that changes in meteorological variables exhibit regular variations after analysing 30 years of historical daily mean winter and summer temperatures for De Bilt in the Netherlands. They proposed a Generalised Autoregressive Conditional Heteroskedastic (GARCH) model to capture the conditional variance and an Auto-Regressive (AR_2) model for the conditional mean. Subsequently Franses et al.

(2001) also proposed a non linear GARCH model to capture the volatility in Dutch temperature and found the imposed asymmetry in the dynamics led to better out-of-sample predictions. Campbell and Diebold (2005) extend this approach to model temperature for pricing weather derivatives after observing seasonality and trend in American temperature data. They model and forecast daily average temperatures in US cities by employing a low-order Fourier series for the temperature seasonality and then using a low-order polynomial deterministic trend for the trend component. The model by Campbell and Diebold (2005) performed equally as well as the Earth Satellite Corporation (EarthSat) forecasts after the 8-day horizon but worse for predictions before that horizon. Nevertheless, such a simple time series formulation proved useful, considering especially the expiration time of most weather derivative contracts, and set one of the early benchmarks for temperature models in the field of weather derivative pricing.

Some early researchers including Carmona (1999) and Moreno (2000) modelled temperature as a discrete $AR(p)$ process but subsequently three forms of the Autoregressive Integrated Moving Average (ARIMA) model (Box and Jenkins 1976) were used more frequently in temperature derivative research. These are ARIMA(lag,1,lag), ARIMA(lag,0,lag) (also termed ARMA) and AutoARIMA (lag,auto,lag), where in all three forms an appropriate lag length needs to be calculated in order to fit the model to the observed historical daily average temperature before the resulting model parameters can be used to forecast future values. Most recently, Murat et al. (2018) found ARIMA models captured the time series dynamics of daily temperatures in four European sites, namely Jokioinen, Dikopshof, Lleida and Lublin, well enough to produce sensible forecasts. Zhu and Li (2023) also modelled daily Berkeley (United States) temperatures using an ARIMA (12,1,5) model in order to forecast long-term temperature trends whilst Yu et al. (2023) found ARIMA models outperformed Long Short Term Memory (LSTM) models in predicting long-term global monthly temperature forecasts. In contrast, De Saa and Ranathunga (2020) found LSTM outperformed an ARIMA model in forecasting short-term (12) monthly forecasts in Szeged, Hungary.

Other researchers have experimented with extensions of the classical AR models including seasonal ARIMA (SARIMA) and SARIMA with exogenous variables (SARIMAX). Singh et al. (2019) compared a SARIMA model with an SVM, ANN, and RNN models in forecasting monthly Ahmedabad, India temperatures over a ten-year horizon and found the SARIMA was best based on RMSE scores. Dimri et al. (2020) also in their attempt to overcome overfitting and misspecification challenges of an ARIMA model employed a SARIMA model to forecast monthly temperatures in Uttarakhand, India up to 20 years ahead but found that even though the model fitted the data well it tended to overpredict the actual temperature. Sharma and Singh (2024) compared both the SARIMA and SARIMAX models with an LSTM in their attempt to find a model that generalises well across India. They included eight Indian cities in their work and found that LSTM captured long-term dependencies in the temperature data better than conventional time series models. Yet other researchers have attempted hybrid time series models in their work. Elshewey et al. (2022) compared Wavelet Decomposition and Seasonal Auto-Regressive Integrated Moving Average with Exogenous Variables (WD_SARIMAX) with 5 regression ML models in predicting Delhi weather and found the hybrid model performed better than the ML algorithms.

Given the location and model inconsistency in results, a lot of researchers still advocate further research, especially to cover more areas when concluding their work.

2.2.2 Ornstein-Uhlenbeck models

Similar to the classical time series models some early researchers including Alaton et al. (2002) modelled temperature as a nested Ornstein-Uhlenbeck (OU) (Ornstein 1930) process after observing two main components of the daily temperature dynamics: a deterministic component, made up of seasonality and trend, and a stochastic one. They modelled the seasonal dependency with a sine function and added a linear trend to capture the deterministic aspects of temperature. For the stochastic part, they found that a Wiener process provided a good fit whilst also being mathematically tractable. Alaton et al. (2002) also include a mean reversion component in the stochastic element and combine all these together in an OU framework.

Brody et al. (2002) also propose a dynamical model for temperature evolution, where the OU process is driven by a fractional Brownian motion. Their approach is motivated by empirical observations of daily temperature in central England exhibiting long range temporal dependencies with normally distributed fluctuations. Benth and Šaltyte Benth (2005) suggest an OU model with seasonal mean and volatility, where the residuals are generated by a Levy process rather than a Brownian motion; they use non-Gaussian OU process because the normal hypothesis was rejected for the Norwegian data they studied. They also use a Fourier series representation for capturing possible seasonality in the volatility of temperature. The model by Benth and Šaltyte Benth (2005) fit Norwegian temperature data quite successfully and in particular, explained the seasonality, heavy tails and skewness observed in the data.

Subsequent research has largely focused on improving various characterisations of the OU process driving temperature and this is evident in the various flavours of OU models that dominate temperature derivative research. A large part of related research focuses on location-specific modifications to the OU model; this leaves open the question of non OU-based research on the topic. Oetomo and Stevenson (2005) model temperature under an OU framework but capture the stochastic element with a Levy process to account for temperature jumps. Oetomo and Stevenson (2005) hypothesise that this will allow the model to be flexible and applicable to different locations and periods. They however concede that jumps are not ubiquitous and including them may result in better estimation of the temperature index (HDD and CDD) while deteriorating the fit of the model to actual temperatures.

We now proceed to define in more detail the models used in (Alaton et al. 2002) and (Benth and Šaltyte Benth 2005), two of the most referenced works in temperature derivative modelling. Alaton et al. (2002) express the deterministic part of temperature, T, in Eq. 2 below as:

$$T_t^m = A + B_t + C\sin(\omega t + \varphi), \tag{2}$$

where T_t^m is the mean daily temperature on day t, $A + B_t$ is capturing the weak linear trend observed by Alaton et al. (2002) and $C\sin(\omega t + \varphi)$ captures the seasonality in daily average temperature. The reader is referred to Section 3.4 of the paper by Alaton et al. (2002) for greater detail of the estimation of $\omega = 2\pi/365$ and A, B_t, C and φ which we implemented in our study. Alaton et al. (2002) also assume a Wiener process as driving the noise of their temperature innovations and after accounting for mean reversion, arrive at a stochastic differential equation whose solution is shown in Eq. 3 as:

$$T_t = (x - T_s^m)e^{-a(t-s)} + T_t^m + \int_s^t e^{-a(t-\tau)}\sigma_\tau dW_\tau, \tag{3}$$

where T is daily temperature, T^m is the mean daily temperature of the respective day, s is the current date, t a future date whose daily average temperature is to be predicted, x is the actual current date's average daily temperature, e is Euler's number, a is the temperature mean reversion parameter, σ is the daily temperature volatility estimated from the data and dW is the Wiener noise driving process.

The model by Benth and Šaltyte Benth (2005) differs from the one by Alaton et al. (2002) in using a Levy rather than Brownian process to model the noise dynamics of daily temperature after removing trend, seasonality and auto-correlation. Benth and Šaltyte Benth (2005) formally define their model in Eq. 4 as follows:

$$dT_t = ds_t + k(T_t - s_t)dt + \sigma_t dL_t, \tag{4}$$

with an explicit solution given in Eq. 5 as:

$$T_t = s_t + (T_0 - s_0)e^{kt} + \int_0^t \sigma(u)e^{k(t-u)}dL_u, \tag{5}$$

where again T is the daily temperature, t is the future day whose temperature is being predicted, s is the day-of-year mean temperature, 0 is the start day or day with the known temperature, e is Euler's number, k is the speed of mean reversion, σ is the volatility of daily average temperature estimated from the data, dL is the Levy noise process with u the mean of that noise distribution.

Most recently, Gyamfi (2024) model daily temperature in the Bono region of Ghana with a modified OU process with Brownian motion residuals, while Eggen et al. (2022) found OU model with inverse Gaussian residuals worked well for stratospheric polar temperature modelling. Alfonsi and Vadillo (2024) also recently extended the OU concept further with a stochastic volatility model where the daily average temperature in eight European cities is allowed to return to a seasonal trend with a deterministic time dependent volatility.

2.2.3 Machine learning models

The majority of weather derivative research within the statistical branch has been in stochastic process models followed by the auto-regressive models with fewer researchers employing methods outside these main two. In recent years there has been a lot of interest in using machine learning models in weather forecasting but these have mainly focused on ML models as an alternative to the physical models that are the state of the art for short term/meteorological forecasts.

Liu et al. (2014) employed a deep neural network model to learn important temperature features from huge volumes of weather data and predict changes in Hong Kong weather 24 h ahead. Hossain et al. (2015) also compared standard neural network architecture with a deep learning network for predicting hourly temperature data in Nevada (USA) and found the deep neural network to be better. They also observed that the inclusion of air pressure, humidity and wind speed data in the training process improved the prediction of air tem-

perature at such short horizons. Karevan and Suykens (2018) employed a two-layer stacked spatiotemporal LSTM to predict temperature in the European cities of Brussels, Antwerp, Liege, Amsterdam and Eindhoven. The first LSTM layer was a model of each location which became an input for the second layer. Karevan and Suykens (2018) also included 18 other meteorological variables as covariates and found that the stacked LSTM outperformed the single state LSTM when predicting temperature up to six days ahead. Dueben and Bauer (2018) appraised the potential of NN models to replace the physical models used in meteorological forecasts and concluded that they are yet to see evidence that NN models can produce competitive forecasts compared to conventional physical models for horizons beyond a day and especially on a global level across the wide range of parameters provided by the current state of the art physical models. Frnda et al. (2019) also proposed a neural network architecture to improve the forecast provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) and achieved an improvement in RMSE for the three-hour-ahead forecasts of nine cities in the Czech Republic and Slovakia. Frnda et al. (2019) acknowledged in their work that local weather models do provide better forecasts compared to global or less localised models even though they require greater computational power.

Research interest in applying machine learning models to temperature prediction is still ongoing with more recent work. Kreuzer et al. (2020) compared a 2D-convolutional LSTM (convLSTM) to a single layer LSTM and a SARIMA model in predicting hourly temperature data up to 24 h ahead for the German cities of Bremerhaven, Essen, Kassel, Ulm, and Kempten; they found the convLSTM outperformed the others for horizons beyond a few hours and argued that this was worth the extra computational expense of the convolutional LSTM. Kreuzer et al. (2020) in their work also acknowledge the generalizability challenge of temperature forecasting and adopt a shuffling mechanism during training of the LSTM network to enable the model optimiser to receive different seasons for the first year of training and help it generalise better. Haque et al. (2021) compared four deep learning architectures (simple recurrent neural network, gated recurrent unit, LSTM and CNN) and two hybrid models (CNN-LSTM and GRU-LSTM) in predicting Beijing hourly temperature up to six hours ahead using high resolution/hourly historical data. They also ran performance consistency analysis using Toronto, Las Vegas, Seattle and Dallas. They found the GRU-LSTM returned the lowest RMSE for Beijing (the main location) but the GRU returned the most consistent result across all locations, which highlights the need for models that can generalise across numerous locations even for ultra short-term predictions. Ren et al. (2021) in their survey of deep learning based methods in weather forecasting evidenced a wealth of research done in replacing or augmenting (short-term) meteorological forecasts with deep learning methods as a means of saving on the computational costs of the alternative physical models. Ben Bouallègue et al. (2024) investigated the viability of machine learning algorithms as a replacement for the computationally expensive physical models that dominate meteorological weather forecasting. Their comparative tests involving the prediction of hourly temperatures up to 10 days ahead highlighted the inability of ML models to predict extremes not seen in the training set, as well as their lack of interpretability, even though they show promise in numerical weather prediction. Ben Bouallègue et al. (2024) concluded that further research is needed to demonstrate the value that ML brings to weather forecasting. In the area of general weather forecasting Ham et al. (2019) were able to forecast El Niño-Southern Oscillation (ENSO) values for lead times up to a year and half using convolutional neural networks and surmount the challenge of multi-year ENSO forecasting.

Far fewer researchers however, have tried to solve the specific problem of temperature prediction for weather derivative pricing using machine learning algorithms. Early on, Zapranis and Alexandridis (2008) suggested a paradigm shift in temperature modelling for derivative pricing by proposing the use of non-linear, non-parametric neural networks to capture the mean reversion rate of daily average temperature. Alexandridis et al. (2017) noted that a fundamental problem with the linear models specified in previous literature was their inability to capture asymmetric features as well as outliers which occur in real temperature data and suggested machine learning algorithms as non-linear alternatives. They successfully show the superiority of non-parametric models across different locations and forecast horizons. Alexandridis et al. (2017) use wavelet networks (WNs) and genetic programming (GP) to successfully model temperature. In particular, they observed that WNs exhibit superior performance compared to earlier linear approaches, while GPs show promise for achieving similar performance with fitness functions more suited to temperature modelling.

To conclude, ML algorithms have seen increased use in short term meteorological temperature forecasting as noted by Zhang et al. (2025) in their recent survey of literature in the field. There have also been recent research in rainfall derivatives by (He et al. 2022) and (Dehvari et al. 2025) using ML models whilst others including Price et al. (2025) and Supto (2025) have even used Generative AI models in weather forecasting but these have also been limited to replacing the short term forecasts provided by the physical general circulation weather models. Machine learning algorithms relax the parametric assumptions made by the auto regressive and OU models whilst being able to capture non linear dynamics in historical temperature making them suitable for forecasting temperature across a wide range of locations for the long term horizons needed in derivative pricing but there is still a dearth in research using machine learning algorithms to model long-term temperatures needed for the pricing of weather derivatives. This leaves open the possibility for machine learning approaches to bring improvements in temperature modelling for the specific purpose of pricing weather derivatives. To our knowledge, few researchers have attempted to apply machine learning algorithms across a very wide range of locations to not only solve the temperature prediction problem but also address the location generalisation issue. We are also the first to compare the performance of temperature forecasting models across the specific forecasting windows used in pricing temperature derivatives.

2.2.4 Statistical models mentioned less frequently in literature

Outside the main methods mentioned above that dominate the literature there have been other less often used methods including the following;

(a) Regime Switching Models: other researchers have observed the importance of jumps or shifts in temperature data (Li et al. 2025) and employed regime switching models to capture this dynamic. Türkvatan et al. (2020) proposed a Markov chain regime switching framework for modelling daily Chicago, US, temperature for the purpose of pricing derivatives and concluded that the regime switching presents a better representation of temperature dynamics as well as being at par or better in forecasting compared to other methods. Li et al. (2025) also employed a regime switching OU model to predict temperatures in Toronto, Canada. They argued that a regime switching model was more

universal and therefore more adaptable than single process models but advised prior investigation of the data for any location before its inclusion in the model.

(b) Spline Models: cubic splines have been used to smooth out the volatility in derivatives prices but researchers like Gyamfi et al. (2025) have included it directly in the temperature model itself to capture the non-constantnt volatility of daily temperature within the framework. This follows earlier work by Schiller et al. (2012) who applied splines to de-seasonalise and de-trend temperature data from 35 stations in the United States in their work to price temperature derivative contracts.

(c) Hybrid Models: some researchers have combined models from the above-mentioned groups to form hybrids aimed at improving these single models without losing their benefits. Recent examples include Yan et al. (2020) who employed an empirical mode decomposition-temporal convolutional network (EEMD TCN) to model future El Niño-Southern Oscillation (ENSO) values. They decomposed a Nino index into flat components, used a TCN to predict each component and finally combined these to predict ENSO outperforming pure TCN and LSTM models. Grover et al. (2015) set out to overcome the shortcomings in weather prediction by explicitly modelling spatio temporal dependencies in weather data via a hybrid that combines discriminatingly trained predictive models with a deep neural network modelling the joint statistics of a set of weather-related variables for 60 locations in the United Sates showing that their hybrid approach can outperform NOAA benchmarks for short term predictions up to 12 h ahead. Haque et al. (2021) also trialled hybrid (LSTM) models in their attempt to forecast (short-term) temperature but found them to be highly inconsistent in their performance despite their higher computational cost.

2.2.5 Lack of consensus in the field

There is a lack of global adoption of weather derivatives beyond the United States and a few other cities in other countries. The CME, for example, does not offer index based weather derivatives for developing economies where risk exposure is higher and there is greater need for weather risk hedging. This is partly due to lack of consensus on the best method for pricing temperature derivatives but also, a lack of research covering more extensive locations.

Berhane et al. (n.d.) find performance inconsistencies when comparing different models across temperature indexes, forecast horizons and locations. After examining six different types of temperature models, they find that no model consistently outperforms the rest. They also find that ARMA models provide better fit compared to Monte Carlo simulations, but this superiority is not replicated in out-of-sample tests. Zong and Ender (2016) argue that, after years of research, temperature derivative pricing is still missing a standardised model to capture weather dynamics as well as an effective valuation method for weather derivatives. They fit an ARFIMA seasonal GARCH model that allows for long memory effects and other important temperature properties to six US cities as an attempt to move the field towards this consensus. Tong et al. (2020) compare the performance of seven models that include some or all of the following features: varying levels of seasonal cycle in mean, global warming, cyclical components in mean, long memory, seasonal cycles in volatility and volatility clustering. They remark that models that account for more features perform better and they also observe a ranking of importance within the features for modelling tem-

perature. They suggest continued research to help build consensus but also advise comprehensive performance valuation within future research. Cabrales et al. (2022) highlight the lack of coverage and location inconsistency of current temperature models when they attempt to price Equatorial temperature derivatives. They note that there are characteristic changes in temperature specific to Equatorial regions and explore several combinations of different deterministic functions and stochastic processes to model Bogota in Colombia's temperature. They try four Fourier specifications of the seasonal component of the deterministic temperature process and conclude that a fourth-order truncated Fourier sum plus a mean-reverting process is the best model for fitting Bogota temperature data. Oetomo and Stevenson (2005) also study six models including both autoregressive and OU models and conclude that models performing better in temperature forecasting do not necessarily lead to more accurate pricing.

To conclude, most previous researchers call for further research to improve adoption of weather derivatives as a solution to the problem of weather risk. The main obstacles they have identified are model consistency, global coverage, and lack of standardisation. This paper proposes machine learning as an answer to these challenges. Given their flexibility and adaptability, machine learning algorithms can improve the accuracy and generalisability of temperature forecasts across different locations, which will ultimately support the broader adoption of weather derivates in markets that are most vulnerable to climate risk.

3 Methodology

The main research question we intend to answer with this study is to find out whether machine learning algorithms can be used to model and predict future temperature values across a wide range of locations as well as the state-of-the-art alternatives. This will encourage wider adoption of temperature derivatives by market participants looking to hedge their businesses against temperature risk. We aim to achieve this by setting up a framework in Python (Van Rossum and Drake 2009) to fit 12 machine learning algorithms to predict daily average temperature data from 65 locations around the world, select the best set of parameters for each model-data set pair and use this model to forecast future daily average temperatures. Secondly, we will model and forecast daily average temperature using the current state-of-the-art models and, finally, compare the performance of our machine learning algorithms with the state-of-the-art alternatives over specific forecast horizons for which temperature derivatives are priced. The performance of each algorithm will be mainly measured by the forecast CAT and HDD Absolute Percentage Error (APE) formulated by Alexandridis et al. (2021) specifically for weather derivatives as well as some other metrics used in the wider forecasting literature; see also Sect. 3.2.4.

The Python machine learning models we used follow the empirical risk minimisation (ERM) principle as formalised in statistical learning theory where the goal is to learn a predictive function f that maps inputs $x \epsilon R^d$ to outputs y. Vapnik (1998) formalised this through the concept of risk but because the true distributions of the input and target variables are unknown, he proposed the ERM principle to replace the expectation with the empirical average over the observed training data $\{(x_i, y_i)\}_{i=1}^{N}$:

$$R_{\text{emp}}(f) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}\big(f(x_i),\, y_i\big)$$

where N is the number of training samples, $x_i \in \mathbb{R}^d$ is the feature vector for sample i, y_i is the observed target value for sample i, $R_{\text{emp}}(f)$ is the empirical risk (average loss on the training set). In the Python Scikit Learn libraries we used, the ERM estimator is defined to incorporate a regularisation to control model complexity leading to a regularised ERM objective defined as:

$$\hat{f} = \arg\min_{f\in\mathcal{F}}\left\{\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}\big(f(x_i),\, y_i\big) + \lambda\,\Omega(f)\right\}$$

with $\mathcal{F}$ as the hypothesis class (e.g., linear models, trees, neural networks), $\hat{f}$ as the predictor which minimises empirical risk, $\arg\min$ is the operator returning the minimising function, $\Omega(f)$ is the regularisation term penalising model complexity and $\lambda \geq 0$ as the regularisation strength controlling the trade off between fit and complexity.

In Scikit Learn, `LinearRegression` for example minimises empirical squared-error loss without explicit regularisation, `DecisionTreeRegressor` selects splits that minimise empirical mean squared error impurity at each node, and `MLPRegressor` minimises empirical squared error loss with L2 regularisation on the network weights. They are all specific instances of the ERM framework above, differing only in the choice of hypothesis class $\mathcal{F}$, loss function $\mathcal{L}$, and regulariser Ω. We refer the reader to the appendix of this study for explicit mathematical formulations of the ML models used in our experiments.

3.1 Temperature data

We aim to identify machine learning algorithms that can be used to model temperature for pricing derivatives across a wide range of locations. Therefore, we selected 65 locations across Europe, Northern America and Australia (see Fig. 1). The locations spanned a broad band of latitudes from a maximum of 69.6767 in the Northern Hemisphere to −33.8595 in the south. The locations were selected to include cities where exchange traded temperature derivatives are currently available as well as those included in previous published temperature derivatives research.

Historical daily maximum and minimum temperatures for each of the 65 locations were downloaded from the National Oceanography and Atmospheric Administration (NOAA) website (Noaa 2024). The datasets for different locations start and end on different days of the year, enabling us to identify models robust to different forecasting periods of the year. The NOAA temperatures are recorded in degrees Fahrenheit (°F) and the different locations show great variations in the differences between minimum and maximum daily temperatures reported over the sample period. The summary descriptive statistics of the temperatures from the 65 locations are shown in Table 1.

Previous work has observed wide variations in the statistical properties of Daily Average Temperatures (DATs) over different locations which influence the choice of models and DAT stylised facts accounted for in modelling. For example, temperature variation around

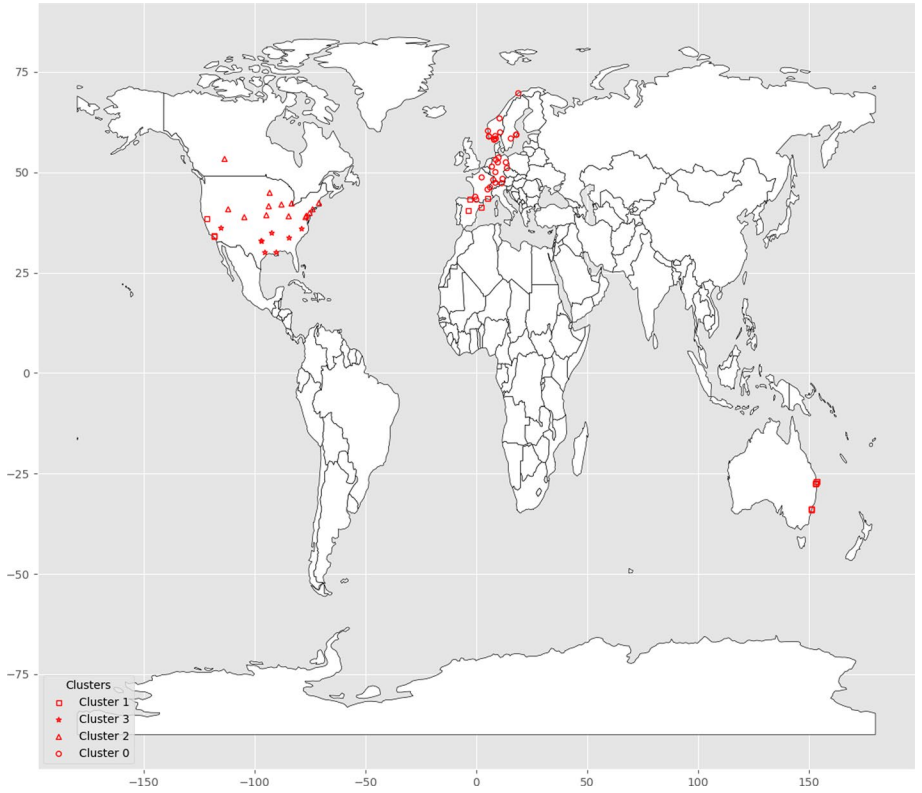


Fig. 1 Locations of weather derivatives stations used in this study. These locations are clustered by temperature descriptive statistics - see Sect. 4.5 for more information on how clustering was performed

the mean of winter was observed to be much higher than that of summer variations leading to assumptions of heteroskedasticity. The historical DATs also exhibit wide variations in their descriptive statistics requiring models robust to different types of temperature stylised facts. The mean DAT of the locations range from a low of 35.9(°F) in Edmonton to 70.5(°F) in Cape Moreton. The other DAT attributes also exhibit a wide spread across the different locations; Edmonton recorded the lowest minimum temperature of −41.5(°F) with Cape Moreton recording the highest DAT minimum of 50.5(°F). Tromso-Langnes recorded the lowest maximum DAT (72.5(°F)) and McCarran in Las Vegas recorded the highest maximum DAT of 106(°F). The standard deviations of the recorded DATs ranged from as low as 6 in Cape Moreton to as high as 23 for Minneapolis. Most of the locations also exhibited negative skewness and kurtosis.

Let us consider the Boston station as an example; see Fig. 2. The DAT distribution shows clear seasonality in the daily average temperature with a winter mean around 40(°F) and a summer mean around 75(°F), and it also shows the asymmetrical nature of daily average temperatures with a negative skewness as observed from the longer tail on the left of the DAT histogram.

Table 1 Temperature descriptive statistics: these are mean, standard deviation, minimum, maximum, median, skewness and kurtosis

Location	Mean	Std	Min	Max	Med	Skew	Kurt
Archerfield Airport	69.2	7.8	48.0	94.0	69.5	−0.0978	−0.8581
Atlanta Hartsfield-Jackson	62.4	15.0	5.0	92.0	64.0	−0.4107	−0.7756
Baltimore-Washington Intl	55.9	17.3	0.0	93.5	56.5	−0.1894	−0.9821
Bankstown Airport Aws	64.2	8.8	42.5	94.0	64.0	0.1868	−0.7055
Barcelona Aeropuerto	60.7	10.5	25.0	90.0	59.5	0.0881	−0.9588
Bergen Flesland	45.1	10.3	5.5	76.0	45.0	−0.0970	−0.4513
Berlin-Dahlem	49.0	13.9	−1.0	84.0	49.5	−0.1761	−0.5866
Bilbao Aeropuerto	59.3	9.6	31.5	89.5	59.5	−0.0165	−0.5974
Boston	51.7	17.3	−4.0	92.5	52.0	−0.1489	−0.8650
Bremen-Seefahrtschule	48.9	12.5	2.5	85.5	49.5	−0.1954	−0.4385
Brisbane Aero	68.9	7.2	48.5	89.0	69.5	−0.2235	−0.8675
Burbank Valley Pump Plt	64.5	9.2	38.5	93.5	64.0	0.0880	−0.6532
Cape Moreton Lighthouse	70.5	6.0	50.5	86.0	71.0	−0.2539	−0.7617
Chicago Ohare Intl Ap	49.8	20.1	−18.0	92.5	51.0	−0.3096	−0.7371
Cincinnati Northern Kentu	54.1	18.4	−12.5	90.0	56.0	−0.3855	−0.7492
Colorado Springs Muni Ap	49.3	17.2	−16.0	84.5	50.0	−0.3076	−0.6341
Dal-Ftw Wscmo Ap	66.3	16.4	8.0	97.5	68.0	−0.3905	−0.7426
Dallas Faa Ap	66.9	16.4	9.0	99.5	68.5	−0.3953	−0.7045
Des Moines Intl Ap	50.5	21.6	−17.5	92.5	52.5	−0.3805	−0.7616
Detroit Metro Ap	49.8	19.2	−12.0	89.5	50.5	−0.2377	−0.9165
Dresden-Strehlen	51.4	14.1	4.5	87.0	51.5	−0.1641	−0.5284
Edmonton Int'L A	35.9	22.2	−41.5	76.5	39.5	−0.6475	−0.2629
Essen-Bredeney	50.3	12.3	3.5	88.0	50.5	−0.1210	−0.4861
Geneve Cointrin	50.8	13.2	3.0	86.5	50.7	−0.0101	−0.8272
Hamburg Fuhlsbuettel	48.2	12.6	−5.0	85.5	48.5	−0.1863	−0.4820
Hannover	48.8	12.9	−2.0	84.5	49.5	−0.2735	−0.3523
Houston Intercontinental	69.5	13.5	17.5	95.0	72.0	−0.5770	−0.5501
Innsbruck	49.3	14.8	−6.0	85.0	50.5	−0.2685	−0.6905
Kansas City Intl Ap	54.6	19.9	−15.5	93.5	56.5	−0.3813	−0.6640
Kjevik	44.9	12.8	−9.0	76.5	45.0	−0.3360	−0.3708
Laguardia Ap	55.3	17.4	2.5	94.5	55.5	−0.1694	−0.9333
Linkoeping-Malmslaett	44.0	14.6	−16.5	80.5	43.5	−0.2092	−0.5767
Little Rock	62.5	16.9	3.0	98.5	64.0	−0.3932	−0.7364
Los Angeles Intl Ap	63.0	6.5	39.5	94.0	63.0	0.0707	−0.1632
Lyon - St Exupery	53.4	13.3	2.5	90.5	53.5	−0.1345	−0.5642
Madrid Barajas	58.0	13.5	19.0	91.0	56.5	0.1978	−1.0254
Marseilles-Marignane	59.3	12.3	13.5	91.5	59.0	−0.0563	−0.8651
Mccarran Intl Ap	67.9	17.2	19.5	106.0	67.0	0.0212	−1.1792
Minneapolis-St Paul Intl	45.8	23.0	−24.5	91.0	48.0	−0.3469	−0.8167
Mont-De-Marsan	56.0	11.5	9.0	89.0	56.0	−0.1344	−0.4973
New Orleans Ap	69.2	12.7	19.0	92.5	72.0	−0.6236	−0.5086
Oksoey Fyr	46.4	11.3	2.5	80.5	46.5	−0.1877	−0.6318
Orly	52.6	11.8	7.5	91.5	52.5	−0.0864	−0.5263
Oslo Blindern	43.7	15.2	−8.0	80.0	43.5	−0.2039	−0.6919
Phila Intl Ap	55.5	17.6	0.5	92.5	56.0	−0.1924	−0.9656
Raleigh Ap	60.1	15.7	4.0	91.0	61.5	−0.3267	−0.9002
Rhein Main	50.5	13.6	1.0	90.5	50.5	−0.1211	−0.6235

Table 1 (continued)

Location	Mean	Std	Min	Max	Med	Skew	Kurt
Sacramento Ap Asos	61.1	11.8	27.5	94.5	61.0	0.0140	−0.9383
Salt Lake City Intl Ap	52.7	18.7	−7.0	93.0	51.5	−0.0297	−0.9466
Schwaigermoos	47.8	13.5	−7.5	82.0	48.1	−0.2469	−0.4375
Sola	46.1	10.7	3.5	77.0	46.0	−0.2152	−0.3144
Stavanger - Valand	47.9	10.4	16.0	79.0	47.5	−0.0128	−0.5676
Stockholm	44.3	14.7	−10.5	83.5	43.5	−0.0824	−0.7106
Stockholm-Bromma	44.2	14.9	−11.5	80.5	44.0	−0.2047	−0.5925
Stockholm A	46.9	14.5	1.0	81.0	46.0	−0.0350	−0.7886
Sydney (Observatory Hill)	64.0	7.7	43.5	92.5	64.0	0.0945	−0.6997
Sydney Airport Amo	64.5	8.2	44.0	94.5	64.5	0.1396	−0.6506
Tarbes - Ossun	54.1	11.0	9.0	87.0	54.0	−0.0853	−0.5339
Torungen Fyr	46.2	11.8	1.0	77.5	46.0	−0.2107	−0.6770
Tromso - Langnes	37.8	11.9	1.5	72.5	37.0	0.0139	−0.6308
Trondheim - Voll	43.2	13.0	−2.0	78.5	42.5	−0.1219	−0.3681
Tveitsund	42.3	14.0	−18.0	74.5	42.5	−0.3292	−0.3682
Vogtsburg-Oberrotweil	51.1	13.5	−0.5	84.0	51.5	−0.2363	−0.5643
Washington Reagan Natl Ap	58.2	17.1	2.0	93.5	59.0	−0.2135	−0.9885
Zuerich Fluntern	49.5	13.8	−4.5	83.0	50.0	−0.1309	−0.7699

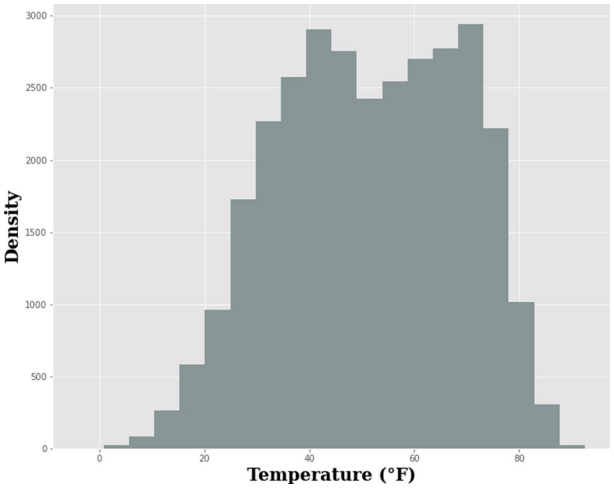


Fig. 2 Bi-modal temperature distribution for Boston illustrating the two dominant seasonal regimes (winter and summer)

3.1.1 Data preprocessing

The historical NOAA data consist of daily minimum and maximum temperatures and we defined the daily average temperature (DAT) for each location as the average of the minimum and maximum daily temperatures. We removed February 29th from the dataset used for machine learning models to maintain equal observations per year, following standard practice in temperature forecasting literature (Šaltyte Benth and Benth 2012). However, we retained all dates for the ARIMA-based models we implemented using the Python Sktime module (Löning et al. 2019), as these models explicitly encode the temporal dimension of

the data and require complete daily frequency without gaps. This difference in preprocessing does not bias performance comparisons because: (1) February 29th represents only 0.27% of the data, (2) all models are evaluated on the same test period which excludes leap days, and (3) the ML models treat observations independently without temporal encoding, making the removal methodologically appropriate for those architectures. We then proceeded to replace missing data using the technique from Alexandridis and Zapranis (2013b) where, if for example the 1st of January were missing, we calculate the average temperature for all other 1st of January as T_{Avy}, then a 14-day average of the seven days before and the seven days after the missing value as T_{Avd}, and finally set the missing DAT (T_{miss}) as the average of these two prior averages:

$$T_{Avy,t} = \frac{1}{N} \sum_{yr=1}^{N} T_{t,yr}, \tag{6}$$

$$T_{Avd,t} = \frac{\sum_{i=1}^{7} T_{t-i} + \sum_{i=1}^{7} T_{t+i}}{14}, \tag{7}$$

$$T_{miss,t} = \frac{T_{Avy,t} + T_{Avd,t}}{2}. \tag{8}$$

Daily average temperature over long periods exhibits a positive trend due to urbanisation and global warming (Alaton et al. 2002). Similar to Alaton et al. (2002) we assume this trend effect is linear and extract it from the daily average temperature with a simple linear equation using ordinary least squares (Johnston and DiNardo 2007). Daily average temperature also shows clear seasonality in temperate locations (Benth 2003). To model the seasonal component, we use Fourier decomposition (Berhane et al. 2021), in which a set of sine and cosine basis functions with annual and sub-annual frequencies captures the smooth, repeating seasonal cycle. After removing the trend and seasonality, the de-trended, de-seasonalised DAT as well as all the other engineered temperature features were standardised by removing the mean and then scaling to unit variance using Scikit-learn StandardScaler before training the machine learning models. The models in (Alaton et al. 2002) and (Benth and Šaltyte Benth 2005) already had these steps in their formulation so in replicating their work for our locations, just the raw daily average temperatures were used as input for those respective models. A similar testing regime was implemented for all the methods; all of the historical temperature data apart from the latest year was used to train each model and estimate its parameters per location, and then the latest year was used to assess the model's predictive performance.

3.1.2 Feature engineering

To improve the information on daily temperature that the ML algorithms can learn from, we created features to mimic stylised DAT facts included in non machine learning temperature models by previous researchers. We computed day-of-year averages to mimic seasonal behaviour, lagged features to capture auto-correlations and features based on differencing and variance to capture the volatility in daily temperature fluctuations. The features are

listed in Table 2 together with a short explanation and a running example. In particular, t is the date of reference (e.g., today), d_t is the index of this date (e.g., first day of year), T_t is the temperature on the date of reference, T_{t+1} corresponds to a forecast temperature for the next day, T_{t-1} is the previous temperature, $T_{t-1\text{yr}}$ is the temperature one year ago, and $\Delta T(1)$ is the change in temperature from the previous one. These features are extended to cover other dates as well and Table 2 provides the full list.

3.2 Experimental set up

All data processing, model development, replication of established state of the art benchmarks, and results analysis were carried out in Python. Data preprocessing was performed using standard Python ETL (Extract, Transform, Load) libraries, including Pandas (version 1.5.3) and NumPy (version 1.24.4). Classical machine-learning models were implemented using scikit-learn (version 1.3.2) (Pedregosa et al. 2011), alongside the XGBoost library (version 1.7.6). Autoregressive models were developed using the sktime framework (version 0.24.0) (Löning et al. 2019), which, like scikit-learn, depends on SciPy (version 1.11.3). For the deep neural-network models, we used the Darts library (version 0.35.0), with Optuna (version 4.4.0) providing the nested hyperparameter tuning functionality. To ensure reproducibility, all components of the workflow that required stochastic operations were executed with the random seed fixed at 23.

3.2.1 Machine learning models and other benchmarks

We selected 12 machine learning models with different architectures and complexities to find the best for modelling temperature over a wide range of locations. The models included simple linear models [Linear Regression (lrm), K Nearest Neighbours (knn)], simple tree-based models [Decision Tree Regressor (dtr)], sequential ensemble tree models [Gradient Boosting Regressor (gbr) and Extreme Gradient Boosting Regressor (xgb)], randomised tree ensemble models [Random Forest Regressor (rfr)], neural network models [Multi-Layer Perceptron Regressor (mlpr), Long Short Term Memory (lstm), Temporal Convo-

Table 2 Machine learning features

Features	Example when date (t) is 1/1/2025
d_t, d_{t+1}	Index of dates: $d_t = 1, d_{t+1} = 2$
$T_{t-7}, T_{t-6}, T_{t-5}, T_{t-4}, T_{t-3}, T_{t-2}, T_{t-1}$	Most recent week temperatures: from 25/12/2024 to 31/12/2024
$T_{t-1\text{yr}}, T_{t-2\text{yr}}, T_{t-3\text{yr}}, T_{t-4\text{yr}}, T_{t-5\text{yr}}, T_{t-1-1\text{yr}}, T_{t-2-1\text{yr}}, T_{t-3-1\text{yr}}, T_{t-4-1\text{yr}}, T_{t-5-1\text{yr}}, T_{t-6-1\text{yr}}, T_{t-7-1\text{yr}}, T_{t+1-1\text{yr}}, T_{t+2-1\text{yr}}, T_{t+3-1\text{yr}}, T_{t+4-1\text{yr}}, T_{t+5-1\text{yr}}, T_{t+6-1\text{yr}}, T_{t+7-1\text{yr}}, T_{t+1-2\text{yr}}, T_{t+1-3\text{yr}}, T_{t+1-4\text{yr}}, T_{t+1-5\text{yr}}$	Previous years temperatures: we collect the five previous yearly temperatures for t and $t+1$ and a fortnight around t one year ago. $T_{t-1\text{yr}}$ is temperature on 01/01/2024, $T_{t-1-1\text{yr}}$ is temperature on 31/12/2023, and so on
$(avg[T_1], avg[T_2], \ldots, avg[T_{365}])$, $(var[T_1], var[T_2], \ldots, var[T_{365}])$	Average temperature and variance for each day over all available data
$\Delta T(1), \Delta T(2), \Delta T(3), \Delta T(4), \Delta T(5), \Delta T(6), \Delta T(7), \Delta T(1\text{yr}), \Delta T(2\text{yr}), \Delta T(3\text{yr}), \Delta T(4\text{yr}), \Delta T(5\text{yr})$	Change in temperature: $\Delta T(x) = T_t - T_{t-x}$ while $\Delta T(x\text{yr}) = T_t - t_{t-x\text{yr}}$
$\min_{i\in\{1,\ldots,7\}}\{\Delta T(i)\}$, $\max_{i\in\{1,\ldots,7\}}\{\Delta T(i)\}$, $\text{avg}_{i\in\{1,\ldots,7\}}\{\Delta T(i)\}$	Minimum, maximum and average change of 1/1/2025 to previous week's temperatures (25/12/2024 - 31/12/2024)

lutional Network (tcn)] and vector machines [Support Vector Machine (svr) and Linear Support Vector Regressor (lsvr)] as well as computational simplifications of boosted tree models [Histogram Gradient Boosting Regressor (hgbr)].

We set up a framework in Python to test and select the best specification of each of the machine learning models. After data preprocessing and feature engineering, the list of features per location was recursively passed through this framework with a saved version of the best model as an output; the best model was then used to forecast future daily average temperatures for the various locations across the different horizons of interest.

Apart from setting up the machine learning framework to model temperature, we also implemented the two most cited state-of-the-art Ornstein-Uhlenbeck models by Alaton et al. (2002) and Benth and Šaltyte Benth (2005) as well as three auto-regressive models [ARIMA (arima), ARMA (arma) and AutoARIMA (autoar)] (Hyndman and Athanasopoulos 2018) that have proved successful for some locations in previous work. We computed the parameters of the solutions for both models by Alaton et al. (2002) and Benth and Šaltyte Benth (2005) for each location in our data set and used those parameters to forecast future temperatures across the different horizons of interest. For the AR models, Bayesian information criteria was used to select the optimal lag length for each location. Both models by Alaton et al. (2002) and Benth and Šaltyte Benth (2005) were formulated for specific locations to capture the temperature characteristics of those locations and our hypothesis is that allowing machine learning algorithms to learn which features are more important for a given location will enable them to generalise better and perform well for most locations without the need to know the specific temperature stylised facts of a location a priori.

3.2.2 Nested time series cross-validation

To ensure our experiments were robust to station specific parameter over fitting, we employed a nested cross-validation strategy to select the optimal hyperparameters for each model. Each training data set was first split into three equal outer non overlapping sections. Within each of these sections, five cross-validation blocks were created using the TimeSeriesSplit available in the Python Scikit-learn module to split each outer block into five subsequent walk-forward train and validate sets. This ensured the five inner cross-validation blocks maintained the temporal integrity needed for time series modelling (Fig. 3).

For each of the five inner training sets, the last year was kept for testing the accuracy of the model and all data prior to that was used to tune the hyperparameters of the machine learning algorithms. This was done via a temporal KFold splitting procedure where in the k-th split the TimeSeriesSplit function returns the first k folds as train set with the next year as the validation set; successive training sets are therefore supersets of those that came before them. The final hyperparameters for each model was then chosen based on the best performing outer block.

To summarise, all historical data except the final year were used as the training period for all models. For the machine learning algorithms specifically, hyperparameters were selected using a nested cross-validation strategy, after which the final model was fitted once using the entire training set. Out-of-sample forecasts for all 17 models were generated using a recursive walk-forward approach, where each predicted day served as an input for the next day prediction. No rolling window or expanding window re-training was performed; the model remained fixed throughout forecasting for the test period.

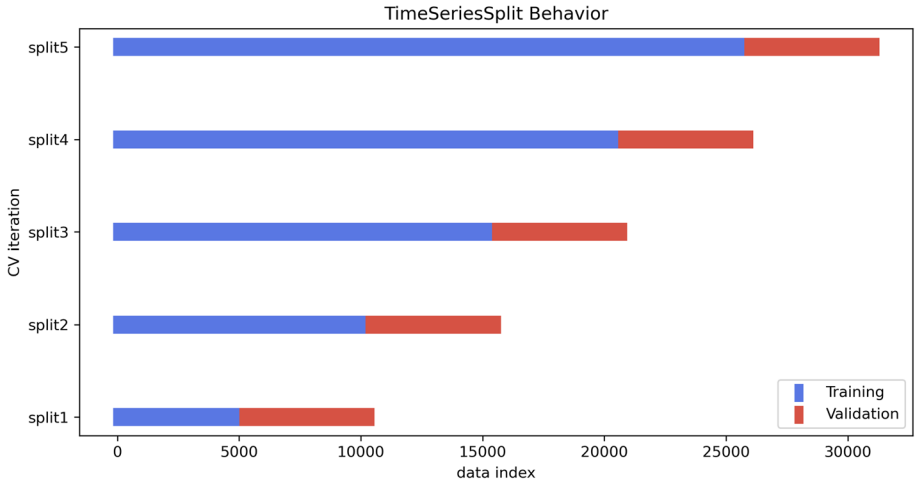


Fig. 3 Walk forward train validation scheme used in the inner loop of the nested cross-validation for ML hyperparameter tuning

3.2.3 Hyper-parameter tuning and temperature forecasting

Using the nested temporal grid search, the hyperparameters for each model were tuned using the nested training data set. Initial values for each hyperparameter were set using heuristics and suggestions from the model documentation over a small subset of the training data set. The final grid search ranges were set from observations of performance from the initial sample grids; grid ranges that did not show promise or were too close together were not included in the final grid search ranges (see Table 3) for each parameter for the nested temporal grid search cross-validation. The best parameters were selected for each data set model combination using Mean Absolute Performance Error (MAPE) as the criterion. The best model was then used for forecasting the unseen data in the test set for each location. After selecting the best model from the model selection framework, we forecast future daily average temperatures across different forecast horizons. We focused on out-of-sample forecasting where the first day of the out-of-sample prediction is made using the last day of the training data and, then, the second day of the out-of-sample prediction is made using the first predicted out-of-sample and so on until the end of the prediction window.

3.2.4 Prediction performance metrics

Our choice of performance metrics was largely influenced by the weather derivatives market. Temperature derivatives cover a period (typically a season) over which the buyer wants to hedge their temperature exposure and their payout is tied to an index accumulated over the contract period for each day that the weather index meets a certain condition (either above or below a certain threshold for example). We therefore adopted the Absolute Percentage Error (APE) used in Alexandridis and Zapranis (2013a) as the main performance metric for this research as it is closest to what the payout for temperature derivatives is based on, and also allows us to compare performance across different horizons which will represent the different temperature risk exposures of different weather market participants.

Table 3 ML models grid search parameter space (parameter definitions from Scikit-learn)

Model	Parameters
Linear regression	Fit_intercept: [True, False]
k nearest neighbor regression	n_neighbors: [10, 15, 50, 100, 500], weights: [‘uniform’, ‘distance’], algorithm: [‘auto’, ‘ball_tree’, ‘kd_tree’], leaf_size: [1, 5, 30, 100, 200]
Random forest regression	n_estimators: [100, 200, 400, 500, 700], criterion: [‘squared_error’, ‘absolute_error’], max_depth: [2, 3, 4, 15, 20], min_samples_split: [2, 5, 10, 50, 100], max_features: [None, ‘sqrt’, ‘log2’, 0.3]
Extreme gradient boosting regression	n_estimators: [5, 50, 80, 100, 150, 200], max_depth: [2, 3, 4, 5, 6], learning_rate: [0.09, 0.1, 0.14, 0.18, 0.2, 0.21]
Gradient boosting regression	Loss: [‘squared_error’, ‘absolute_error’], learning_rate: [0.1, 0.5], n_estimators: [1, 10, 50, 100, 200, 500], subsample: [0.1, 0.5, 1.0], criterion: [‘friedman_mse’, ‘squared_error’], min_samples_split: [2, 3, 4], min_samples_leaf: [2, 3, 4], min_weight_fraction_leaf: [0.0, 0.5], max_depth: [1, 4, 5, 6, 9], max_features: [None, ‘sqrt’, ‘log2’]
Histogram gradient boosting regression	Lloss: [‘squared_error’, ‘absolute_error’], learning_rate: [0.1, 0.5], max_iter: [2000], max_leaf_nodes: [2, 5, 20, 50, 100], max_depth: [1, 4, 5, 6, 9], min_samples_leaf: [2, 3, 4], max_bins: [2, 10, 50, 200]
Linear support vector regression	Epsilon: [0.0, 0.1, 10], tol: [0.0001, 0.0005], regularisation parameter: [1.0], loss: [‘epsilon_insensitive’, ‘squared_epsilon_insensitive’], max_iter: [2000]
Support vector machine regression	Kernel: [‘rbf’, ‘linear’, ‘poly’, ‘sigmoid’], degree: [1, 2, 3], gamma: [‘scale’, ‘auto’], tol: [0.0001, 0.0005], epsilon: [0.1, 0.2, 0.7, 1.0, 10.0], regularisation parameter: [1.0]
Decision tree regression	Criterion: [‘squared_error’, ‘absolute_error’], max_depth: [4, 5, 6, 7, 8], splitter: [‘best’, ‘random’], min_samples_split: [2, 3, 4], min_samples_leaf: [2, 3, 4], max_features: [None, ‘sqrt’, ‘log2’]
Multi layer perceptron regression	Hidden_layer_sizes: [(50), (25), (12), (25, 10), (25, 10,5)], activation: [‘relu’, ‘tanh’], learning_rate: [‘constant’, ‘invscaling’, ‘adaptive’], learning_rate_init: [0.001, 0.005], max_iter: [2000], tol: [0.0001, 0.0005], early_stopping: [True], validation_fraction: [0.1, 0.2]
Temporal convolutional network	Kernel_size: [3, 5, 7, 9], num_filters: [2, 4, 8], weight_norm: [False, True], dilation_base: [2, 4], dropout: [0.0, 0.4]
Long short term memory	Hidden_dim: [4, 8, 16, 32, 64, 128], n_rnn_layers: [1, 2], dropout: [0.0, 0.4], activation: [Relu, tanh]

The APE is defined in Eq. 9, where y is the actual observed temperature and $\hat{y}$ is the forecasted CAT or HDD for the horizon of interest.

$$APE = \left| \frac{y - \hat{y}}{y} \right| \tag{9}$$

In both the CAT and HDD APEs, both y and $\hat{y}$ are cumulative temperatures, calculated from the start of the forecasting period to its end to match derivatives which are priced on cumulative temperature for the pricing window. We computed APE for both CAT forecasts as well as HDD temperature forecasts to account for the two main types of temperature indexes used in pricing weather derivatives. We also report summarised forecast performance for Mean Absolute Percentage Error, Root Mean Squared Error, Mean Absolute Error and Weighted Mean Absolute Percentage Error which have been used in literature for more general model forecast accuracy measurement. These metrics are defined in Eqs. 10 and 11 where n is the number of observations in each horizon, y is the actual temperature i the time step and $\hat{y}$ is the predicted temperature.

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|, \quad \text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{10}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \quad \text{wMAPE} = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{\sum i = 1^n |y_i|}. \tag{11}$$

3.2.5 Defining temperature derivative pricing forecast horizons

Traditional forecast performance horizon definition starts from the first day of the test set and is not usually discretised. Temperature derivatives are primarily sold for a (discrete) season that in practice will need to be priced for a period starting a season or two after the last known temperature and derivative purchase time. As an example, a ski resort would want to buy temperature derivatives in October to cover January - March of the following year. In this instance, the ski resort managers are not concerned about temperature risks from day one of the test set but, instead, from day 120 and then for three months (a ski season) after that; hence, the need is for a model that predicts the temperature within that window more accurately than models that predict the next 30 days in October well.

To ensure we selected the best model for capturing the array of typical risks that weather derivatives will be used to hedge against, we compared the results across horizons measured from day one of the forecast as well as horizons with some lead time. For example, the 30-day horizon looks at how well models can forecast the first 30 days after the training set and the 365-day horizon compares model performance for the first 365 days after the training set. The ISL (One Season Lead) horizon starts one season after the observed temperature and IISL starts two seasons after but both spanning one season (90 days). The mid 90-day horizon (see Fig. 4) compares model performance on a 90-day (season) window that starts 4.5 months after the last day of the training set. The first 4.5 months of forecasts are used to forecast the subsequent three months used in the horizon performance measurement, but they do not directly affect the model's performance scoring for that horizon. This captures the practical scenario when a model is deployed in the real world to price derivatives in the actual weather derivatives market where participants looking to hedge risk in advance will want to buy derivatives that protect them for a given season (90-days) in the future. This is one of the novel contributions of our study; we explicitly look for models that meet the risk profile of most weather derivative market participants.

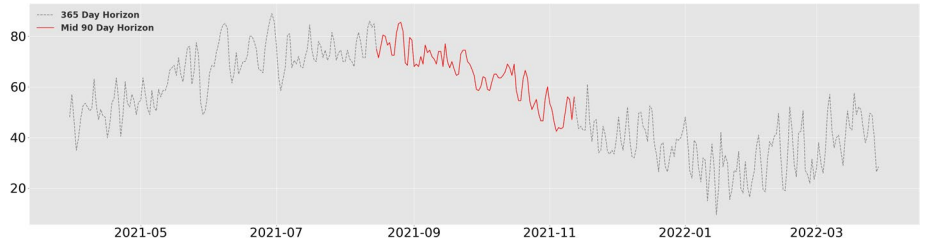


Fig. 4 A visual representation of the mid 90-day horizon; the period of interest is the 90 days starting four and half months after the training set ends and the first forecast begins

Table 4 Model train CoreTimes in hours

Model	$\text{CoreTime}_{(hr)}$
lstm	231
tcn	195
hgbr	172
gbr	16
rfr	7

3.2.6 Running the experiments

All the models were trained on a computer cluster in parallel so that the slower models did not become bottlenecks for completing the entire experiment. Each model was trained on all 65 locations with an allocation of enough CPU cores appropriate for the training length observed from initial modelling tests. Normalised parallelism (number of CPUs x clock time taken) for the top 5 slowest ML models for training BANKSTOWN AIRPORT AWS data are shown in Table 4 for comparison, clearly showing the significant compute costs for some of the more compute intense ML models.

4 Our results

In this section, we compare the out-of-sample predictive performance of the ML algorithms with the OU and autoregressive benchmarks across the 65 locations of our study over different forecast horizons.

4.1 APE results per location

We calculated APE for each combination of model, horizon, and location and in Table 5, we show the 365-day (full year) horizon results for the CAT APE forecasts. We focus on the APE metric (Alexandridis and Zapranis 2013a) because it is the closest to what the payout for temperature derivatives is based on while also allowing us to compare performance across different horizons. We, however, also include summaries of other metrics used in wider forecasting literature for completeness (see Tables 8, 9, 10, 11, 12, 13). Detailed APE performance results are presented in Table 5 for the full prediction year as this gives the most unbiased performance of the models but summaries of all the other horizons are also

Table 5 Full year horizon CAT APE - model with lowest APE per location is in bold font

Name	Arima	Arma	Autoar	Alaton	Benth	dtr	gbr	hgbr	knn	lr	lstm	lsvr	mlp	rfr	svm	tcn	xgb
	(3)	(1)	(2)	(2)	(6)	(6)	(3)	(11)	(3)	(0)	(2)	(3)	(6)	(4)	(7)	(2)	(4)
ARCHERFIELD AIRPORT	0.1	1.48	1.48	1.14	**0.07**	3.12	1.12	1.33	1.56	6.75e+48	0.22	0.37	2.56	1.41	1.79	0.8	3.14
ATLANTA HARTSFIELD-JACKSO	3.84	2.21	1.88	5.26	3.54	1.68	1.56	1.06	1.52	4.14e+41	3.35	**0.01**	1.05	0.39	1.16	4.67	0.2
BALTIMORE-WASHINGTON INTL	4.77	4.37	4.03	5.8	4.98	1.72	4.44	**1.52**	3.84	8.94e+41	4.86	3.53	2.69	3.63	3.42	3.04	2.55
BANKSTOWN AIRPORT AWS	0.75	2.73	2.73	1.66	**0.54**	3.58	2.96	4.83	3.02	1.18e+46	2.58	4.7	3.72	3.03	4.81	1.23	5.19
BARCELONA AEROPUERTO	3.96	0.86	0.86	6.75	3.91	0.29	0.16	3.92	0.88	6.28e+48	2.01	2.89	1.84	0.82	**0.05**	3.09	2.53
BERGEN FLESLAND	4.33	0.93	0.86	7.57	4.13	**0.51**	1.29	3.78	2.45	4.09e+44	3.02	7.22	3.13	2.02	1226.24	4.56	2.63
BERLIN-DAHLEM	2.49	0.31	0.28	6.14	2.84	3.47	NaN	2.41	0.25	2.61e+49	0.31	2	**0.02**	0.79	0.54	1.07	4.39
BILBAO AEROPUERTO	0.68	1.41	1.41	**0.28**	0.74	1.34	NaN	3.78	1.26	2.93e+35	0.93	6.13	3.29	0.54	3.26	1.05	2.76
BOSTON	5.57	4.04	3.52	6.65	5.06	6.21	3.62	2.54	3.55	5.95e+42	3.36	3.17	2.56	4.31	2.92	2	**0.7**
BREMEN-SEEFAHRTSCHULE	2.41	0.45	0.35	4.56	2.59	4.38	0.68	3.39	0.8	1.10e+53	0.14	2.35	**0.13**	0.78	1.43	0.22	2.61
BRISBANE AERO	0.45	0.39	**0.15**	0.9	0.26	0.7	0.32	1.33	0.21	1.38e+49	0.19	0.19	0.92	0.44	0.68	0.67	0.61
BURBANK VALLEY PUMP PLT	0.18	0.62	0.63	1.2	0.48	0.09	NaN	1.6	0.59	2.40e+41	**0.05**	0.71	1.36	1.28	0.71	0.45	1.41
CAPE MORETON LIGHTHOUSE	0.45	1.65	1.65	0.44	**0.41**	0.72	NaN	1.69	2.13	2.61e+45	0.92	3.59	2.35	2.65	2.25	1.28	3.12
CHICAGO OHARE INTL AP	6.01	3.43	3.41	7.24	6.08	1.45	3.59	**0.13**	3.39	3.38e+43	6.37	3.6	2.12	2.92	2.97	4.44	0.22
CINCINNATI NORTHERN KENTU	4.13	3.35	2.8	4.88	3.51	2.56	1.77	**0.04**	1.4	1.76e+44	3.2	1.58	0.89	0.66	0.26	2.95	0.09
COLORADO SPRINGS MUNI AP	6.43	5.05	4.8	6.98	6.4	2.07	NaN	2.21	3.59	1.67e+45	4.66	3.63	4.51	**2.03**	NaN	5.92	2.76
DAL-FTW WSCMO AP	2.11	0.64	1.01	2.73	2.09	0.99	**0.00**	3.69	0.44	5.78e+33	0.6	0.02	1.16	1.16	2.5	1.35	1.71
DALLAS FAA AP	1.57	0.6	**0.46**	3.54	1.51	0.6	NaN	4.16	1.46	1.31e+44	0.59	1.01	1.8	2.5	NaN	0.86	3.18
DES MOINES INTL AP	NaN	NaN	NaN	6.19	3.52	4.25	NaN	1.18	0.38	1.79e+37	1.65	1	0.4	0.17	**0.04**	0.37	0.99
DETROIT METRO AP	4.18	0.56	0.67	7.84	4.16	4.1	0.9	2.28	0.71	1.39e+43	0.94	0.5	**0.11**	1.02	0.4	0.41	3.6
DRESDEN-STREHLEN	1.1	2.72	2.74	**0.13**	0.59	2.59	1.83	0.89	2.52	2.34e+42	0.72	0.96	8.34	2.78	1.63	2.4	4.89
EDMONTON INT'L A	**0.01**	2.41	3.89	0.81	0.93	9.81	3.25	9.45	4.86	8.47e+41	0.25	8.08	11.59	6.02	6.7	4.24	7.54
ESSEN-BREDENEY	1.63	1.16	1.06	4.59	1.79	4.19	0.93	1.45	**0.17**	1.47e+44	0.85	0.91	1.36	0.87	2.5	2.71	4.98
GENEVE COINTRIN	1.72	2.53	2.53	5.77	1.63	NaN	2.15	6.01	3.92	6.76e+53	1.42	1.66	2.2	4.69	**0.34**	1.21	7.07
HAMBURG FUHLSBUETTEL	3.92	2.39	1.85	5.28	3.65	2.51	1.79	2.09	2.19	8.02e+43	3.66	1.37	1.64	3.96	1.17	3.23	**0.5**
HANNOVER	2.96	0.97	0.77	6.65	3.04	1.02	0.66	**0.03**	0.48	3.32e+47	0.52	0.84	0.5	0.27	2.35	3.81	5.24
HOUSTON INTERCONTINENTAL	2.15	0.25	0.28	4.58	2.1	1.91	1.79	1.62	0.43	9.29e+41	**0.03**	1.6	0.81	1.29	NaN	1.55	2.81

Table 5 (continued)

Name	Arima	Arma	Autoar	Alaton	Benth	dtr	gbr	hgbr	knn	lr	lstm	lsvr	mlp	rfr	svm	tcn	xgb
	(3)	(1)	(2)	(2)	(6)	(6)	(3)	(11)	(3)	(0)	(2)	(3)	(6)	(4)	(7)	(2)	(4)
INNSBRUCK	7.4	3.55	3.55	11.17	7.51	2.18	3.52	2.34	1.36	9.98e+43	7.21	4.71	4.88	0.93	4.82	5.98	**0.47**
KANSAS CITY INTL AP	4.14	2.03	2.1	2.96	4.02	0.91	0.23	2.19	0.96	1.75e+44	2.31	4.98	**0.07**	0.07	NaN	3.21	0.93
KJEVIK	4.41	1.14	1.15	7.44	4.4	**0.18**	NaN	1.9	1.45	2.33e+41	1.16	2.42	2.91	0.84	0.65	3.88	2.37
LAGUARDIA AP	5	2.18	2.18	6.03	4.87	0.8	2.34	0.9	2.07	1.43e+44	2.27	2.64	2.95	1.93	**0.72**	5.16	1.19
LINKOEPING-MALMSLAETT	2.45	0.98	1.78	5.83	3.29	2.9	**0.5**	1.66	1.39	6.03e+33	1.95	0.94	3	1.39	0.63	0.68	0.55
LITTLE ROCK	1.59	1.65	1.64	1.02	**0.02**	4.59	2.3	2.43	2.91	2.49e+45	1.81	4.15	2.24	4.48	4.49	0.38	4.11
LOS ANGELES INTL AP	0.79	1.44	1.44	2.82	0.62	0.93	1.33	**0.17**	0.43	1.94e+43	1.15	0.8	0.93	0.91	0.75	0.21	3.98
LYON - ST EXUPERY	4.3	0.63	0.99	7.3	4.18	3.55	**0.13**	1.66	0.43	3.33e+45	0.85	1.78	1.18	0.51	1.97	2.48	2.49
MADRID BARAJAS	1.66	0.98	0.99	4.41	1.73	1.94	1.85	**0.49**	1.46	3.02e+39	0.69	2.1	0.7	2.37	NaN	0.5	4.39
MARSEILLES-MARIGNANE	0.49	3.08	3.08	3.57	**0.26**	4.31	4.35	3.65	3.73	1.67e+54	3.29	3.35	2.98	5.78	610.66	0.58	6.91
MCCARRAN INTL AP	5.46	1.06	1.06	5.94	5.25	1.03	2.63	**0.22**	0.64	4.86e+41	1.12	0.45	2.5	1.69	0.34	5.09	3.01
MINNEAPOLIS-ST PAUL INTL	4.91	3.32	2.79	1.4	4.58	4.83	3.57	0.13	2.18	1.55e+42	4.7	3.46	0.85	1.03	NaN	4.49	**0.02**
MONT-DE-MARSAN	0.4	2.98	2.97	2.3	**0.33**	4.76	3.15	4.04	3.97	2.62e+40	2.7	3.59	3.63	3.89	3.77	5.12	6.07
NEW ORLEANS AP	2.93	0.96	0.97	5.06	3.02	0.76	0.38	1.48	1.03	3.77e+51	0.95	**0.09**	0.22	0.11	0.14	3.47	1.34
OKSOEY FYR	4.81	0.73	0.73	8.31	4.64	1.04	0.88	2.34	1.38	2.74e+40	0.97	1.94	2.13	**0.21**	0.95	3.16	3.17
ORLY	1.52	1.42	1.42	5.22	1.97	5.76	2.21	**0.07**	1.11	7.43e+38	1.44	0.43	0.27	1.02	1.95	2.61	4.72
OSLO BLINDERN	6.33	3.53	3.35	9.46	6.74	**0.21**	3.38	4.17	3.67	3.65e+53	3.38	3.94	4.02	3.56	3.79	9.63	1.05
PHILA INTL AP	5.09	2.27	2.27	7.85	5.15	2.63	0.99	**0.24**	2.43	7.51e+41	2.23	2.3	2.5	1.14	1.85	5.68	1.11
RALEIGH AP	3.48	1.77	2.13	5.52	3.79	0.18	2.17	0.5	1.32	7.21e+43	2.03	1.49	0.89	**0.13**	0.23	3.74	0.54
RHEIN MAIN	1.58	2.86	2.84	5.54	1.4	2.22	3.21	5.28	3.56	2.08e+35	2.31	3.05	1.6	3	2.79	**0.88**	6.92
SACRAMENTO AP ASOS	2.26	**0.61**	0.61	4.13	2.34	2.23	0.8	1.19	0.69	3.56e+45	0.73	1.05	1.48	1.51	1.66	2.89	1.23
SALT LAKE CITY INTL AP	6.03	2.5	2.52	9.5	6.06	4.88	NaN	1.62	0.94	6.46e+47	5.98	1.9	0.4	**0.36**	1.21	3.9	1.51
SCHWAIGERMOOS	**0.14**	2.72	2.72	4.57	0.29	1.77	3.26	5.35	4.28	6.44e+40	1.69	3.36	2.89	4.69	3.28	1.43	5.71
SOLA	5.15	1.66	1.72	8.71	5.18	**1.19**	1.35	4.4	3.22	3.22e+42	5.39	4.8	4.14	1.98	3.17	4.35	2.41
STAVANGER - VALAND	4.24	1.05	1.31	3.5	1.23	1.81	NaN	1.5	**0.16**	4.23e+42	0.84	7.63	0.68	0.6	1.38	8.17	4.27
STOCKHOLM	6.74	1.71	1.72	11.92	6.87	**1.14**	1.43	3.06	1.6	2.91e+40	2.22	3.05	4.3	1.29	2.24	3.56	3.8
STOCKHOLM-BROMMA	3.68	0.77	0.82	8.03	3.74	0.92	NaN	0.74	1.03	1.16e+43	1.04	**0.09**	0.41	3.05	1.59	1.98	5.53

Table 5 (continued)

Name	Arima	Arma	Autoar	Alaton	Benth	dtr	gbr	hgbr	knn	lr	lstm	lsvr	mlp	rfr	svm	tcn	xgb
	(3)	(1)	(2)	(2)	(6)	(6)	(3)	(11)	(3)	(0)	(2)	(3)	(6)	(4)	(7)	(2)	(4)
STOCKHOLM A	1.67	1.73	1.86	5.68	1.21	1.19	3.77	2.63	2.29	2.97e+48	2.13	1.6	0.77	3.74	NaN	**0.65**	5.19
SYDNEY (OBSERVATORY HILL)	4.62	2.11	2.1	7.48	4.5	1.44	1.65	0.46	1.91	1.81e+54	2.6	0.69	2.21	1.41	**0.37**	3.82	0.58
SYDNEY AIRPORT AMO	1.6	1.36	1.16	4.37	1.57	2.6	2.27	3.56	1.62	3.24e+46	1.66	2.48	**1.11**	2.41	4	1.89	5.24
TARBES - OSSUN	3.46	1.12	1.04	5.68	3.49	0.23	0.9	1.89	**0.04**	6.45e+43	3.28	3.32	0.94	0.46	0.17	3.54	2.22
TORUNGEN FYR	5.19	1.89	1.89	9.63	5.58	1.72	2.18	3.26	2.65	6.27e+45	2.07	5.73	3.31	2.56	**0.29**	4.83	1.75
TROMSO - LANGNES	0.91	3.19	2.4	2.97	0.67	6.13	2.15	**0.09**	1.26	1.71e+42	0.18	1.14	0.83	1.09	3.24	0.44	6.19
TRONDHEIM - VOLL	**0.04**	0.2	0.26	0.13	0.13	4.9	0.22	2.45	1.4	4.98e+45	0.19	2.81	2.28	0.21	1.17	4.15	2.1
TVEITSUND	6.8	1.51	1.51	12.84	6.35	3.67	NaN	**1.14**	2.21	4.47e+54	1.36	4.43	4.88	1.4	2.27	8.91	3.33
VOGTSBURG-OBERROTWEIL	3.44	2.07	1.62	4.89	3.36	3.06	1.65	2.27	0.79	1.65e+45	1.69	1.97	1.32	1.39	**0.69**	3.13	1.42
WASHINGTON REAGAN NATL AP	3.71	1.73	2	6.11	3.93	**0.29**	4.08	0.87	1.49	8.72e+39	3.69	1.53	0.41	0.74	1.11	2.95	1.05
ZUERICH FLUNTERN	2.2	1.24	0.78	3.11	1.88	3.03	0.85	3.69	0.92	1.11e+40	2.12	5.52	**0.67**	0.84	1.49	0.83	1.04

presented in subsequent sections. In terms of performance comparison, the full year horizon performance identifies models that work well across both long- and short-term forecasts and over the full four seasons of the year. In contrast, other horizons expose model performance for specific risk windows and allow us to analyse performance evolution and bias over different temperature derivative windows for the different models (see Sect. 4.2).

As noted by other researchers, temperature derivatives can be sensitive to tail events (Cheng et al. 2025), and we therefore examined model behaviour at temperature extremes. Defining a "tail" or "extreme" season is however not straightforward in our setting. Because our forecast performance metric accumulates daily temperature over the full pricing window, isolated extreme days are naturally diluted, making distinct "tail seasons" difficult to detect. We inspected the test set to identify locations exhibiting unusually extreme seasons that might disproportionately influence model performance, but no such cases were observed across the locations we analysed.

To complement this season level assessment, we also evaluated daily forecast errors across temperature quantiles for 4 locations. This quantile based daily error analysis provides an alternate sensitive view of tail behaviour by examining forecast accuracy across the coldest and warmest temperature regimes.

The results of our temperature quantile analysis in Fig. 5 show a relatively smooth evolution of forecast performance over the seasonal temperature cycle, with modestly higher errors in the upper and lower quantiles but no evidence of substantial deterioration in the tails. These findings indicate that the models remain robust even at temperature extremes.

The mid-90-day horizon, discussed previously (see Fig. 4), begins 4.5 months after the end of the training set, while the -30 day looks at temperature forecasts for the final 30 days of the test set. Even though the full year horizon is useful for identifying unbiased models, we expect few market participants would want to hedge temperature risk over one whole year, and fewer still to hedge against temperature fluctuations for the first 30 days starting from the most recent known temperature. The mid-90-day and -30 day horizons represent the more likely hedge windows in real life, but we have included all the extra horizons in the performance analysis presented in subsequent sections so that we can identify the most robust model and also analyse how performance evolves over these weather risk windows. For each forecast window, lower APE indicates a better performing model at that horizon, and this model is highlighted in bold font.

From Table 5, we observe that Histogram Gradient Boost Regression (hgbr) performs best in predicting out-of-sample cumulative average temperature across the whole test year in the most locations (11) with the Support Vector Machine being the next best (7 locations). More generally, the machine learning algorithms perform well at this horizon for this metric being the best in all but 13 locations.

The detailed performance view also exposes the risk of "model blow ups" where models forecast extremely high and unrealistic temperatures (example 1.81×10^{54} CAT APE for Sydney (Observatory Hill) for the full year horizon). The linear regression model is particularly susceptible to this, blowing up in every single location at least once over the full year horizon. An inspection of the fitted linear regression models showed consistently high and implausible parameter values for some of the features used in training the models. This is a clear indication of model misspecification. Other researchers (Alexandridis et al. 2017) have noted that temperature is not linear in many of its features and a linear architecture can be too rigid forcing the model to compensate by assigning excessively large weights which

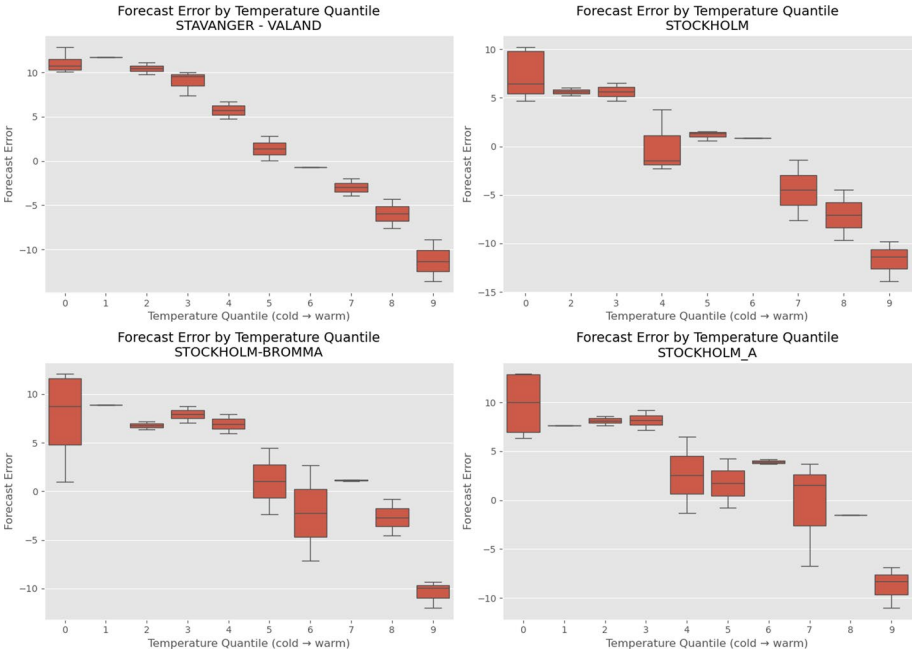


Fig. 5 Forecast error by temperature quantile for 4 locations showing forecast performance evolution across the season without significant tail deterioration

can lead to large unstable forecasts. This highlights the fact that certain model architectures might just not be suitable for the problem of temperature forecasting for pricing derivatives.

The detailed results also highlight instances where algorithms experienced severe optimisation failures where the loss function diverged during training, resulting in NaN (Not a Number) values. These NaNs represent computational failures rather than quantifiable performance metrics. For transparency in our detailed results table, we report NaN values where they genuinely occurred, clearly indicating which algorithm-dataset combinations failed to produce valid predictions. However, for statistical testing purposes, we adopt a conservative imputation strategy: NaN values are replaced with the mean performance across all algorithms for that specific dataset. This approach serves two purposes: (1) it enables statistical comparisons without arbitrarily excluding failed models, which would bias our analysis toward better performing algorithms, and (2) it provides a neutral baseline that neither penalises nor rewards failure disproportionately.

4.2 Performance evolution & bias over different forecast horizons

As mentioned in Sect. 4.1 above, analysing performance across different pricing windows exposes how model performance evolves over the forecast windows. Figures 6 and 7 present 3D plots for 6 cities' APE values for a selection of algorithms[1] for the CAT and HDD indices, respectively. We observe that performance generally deteriorates the further out the forecast horizon is. This is intuitive as predicting temperatures further away from the known

[1] These six algorithms were chosen to show contrasting algorithm performance over the forecast horizons.

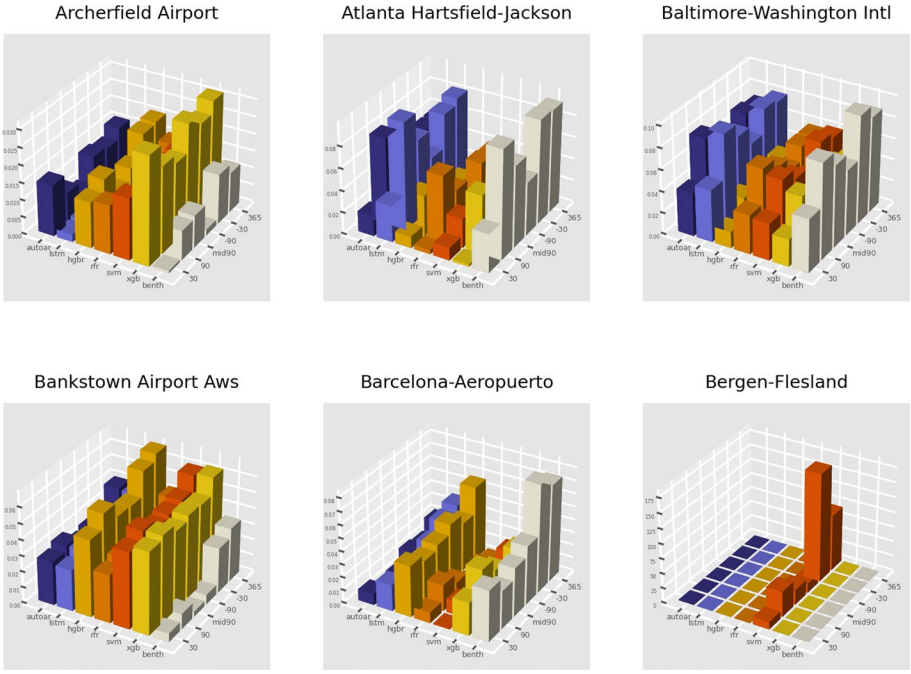


Fig. 6 The CAT APE 3D plot for the first six alphabetical locations

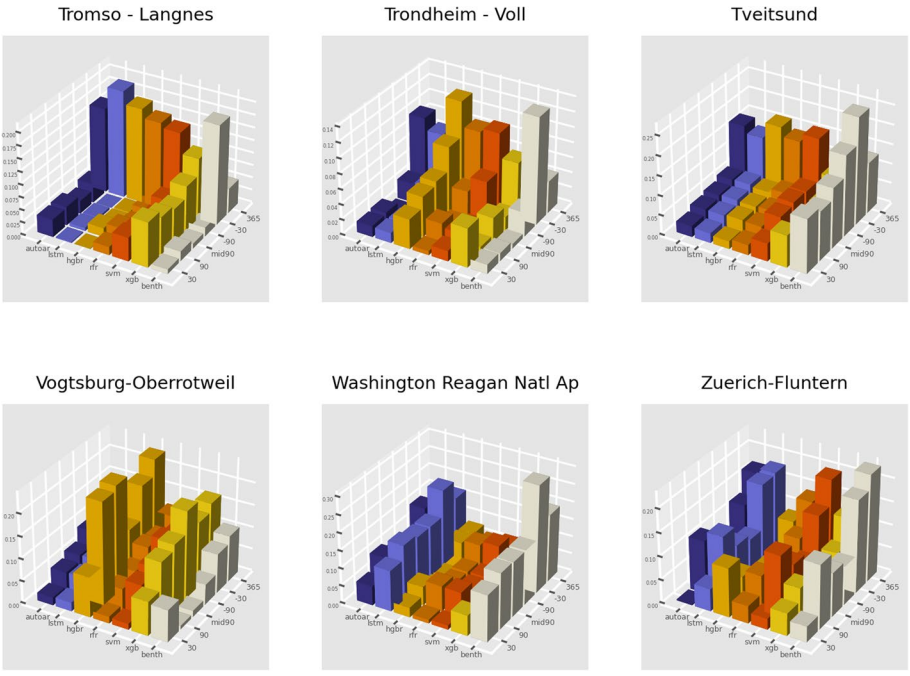


Fig. 7 Last six alphabetical locations HDD APE 3D Plot

actual is more difficult. The difficulty in predicting temperature further in the future presents greater risk to market participants who would thus prefer models that do better at those horizons. Some locations and models do not follow this general trend with Bankstown Airport AWS showing the trend clearly in the CAT APE 3D plot (see Fig. 6). The 3D plots also show the greater difficulty in forecasting windows that start further away from the current date; the performance of the 30-day horizon is generally better than the -30day horizon and the 90-day better than the -90 day horizon. This temporal evolution of forecast performance is also evident in the HDD performance. Tveitsund evidences this quite clearly in the HDD APE 3D plots in Fig. 7.

We also see that differences in model performance, across the different derivative forecasting windows, begin to emerge in the unaggregated results. It is noticeable that the model by Benth and Šaltyte Benth (2005) generally does better compared to the other approaches at the near-term forecast windows, an example of which is evidenced by the Zuerich-Fluntern 3D HDD APE plot in Fig. 7. The model by Benth and Šaltyte Benth (2005) has a lower HDD APE compared to the xgb model next to it at the 30-day horizon, but the xgb model is better than the Benth and Šaltyte Benth (2005) model at forecast windows that are further out in the future, even though both models' performance worsens as we get to later horizons. Throughout the rest of the results and analysis we highlight this shift in performance; we assert that it is of greater utility to market participants to hedge weather risks further out in the future than weather risk closer to the present.

4.3 Performance rank summaries

To get an understanding of the overall performance of the models and assess their ability to forecast temperature for pricing weather derivatives across a wide range of locations, we analyse ranked model performance across the different pricing windows. Tables 6 and 7 below show how many times each model was ranked the best based on its CAT and HDD APE performance, respectively. The model with the largest number of top-ranked locations is highlighted in bold font per horizon, allowing us to visually see the better performing models at each horizon and providing direction to the statistical tests we perform later.

From Table 6 it is evident that the model by Benth and Šaltyte Benth (2005) generally outperforms all others at the earlier or very near-term forecast horizons; however, this performance superiority diminishes the further out the forecast horizon extends and turns in favour of the xgb model for horizons with a lead start time. By the last 30 day horizon xgb is a clear favourite. A similar performance pattern is repeated for the top-ranked HDD results shown in Table 7 with the xgb model performing well at horizons with a lead time and the clear favourite at the last 30 days horizon whilst the model by Benth and Šaltyte Benth (2005) is the best at the opposite near term 30 day horizon. For the HDD predictions however, the performance is less dominated by just xgb and the model by Benth and Šaltyte Benth (2005) as other models like arima and tcn also take a share of the performance superiority.

We also see a similar theme emerge across the other performance metrics. In Table 8, the model from Benth and Šaltyte Benth (2005) still ranks best in a lot more locations than all others for near term horizons and even some of the later horizons, taking those away from the machine learning algorithms. This diminished performance by the ML models in the horizons further out is however reinstated for the other metrics; WMAPE (Table 9), RMSE

Table 6 Number of data sets model ranked best for CAT APE per horizon - most top ranked model per forecast horizon in bold font

Model	30	60	90	120	180	270	365	ISL	IISL	−270	−180	−120	−90	−60	−30
Arima	5	**12**	4	7	**7**	**10**	3	4	7	5	0	4	5	6	6
Arma	1	2	4	2	5	2	1	5	1	1	1	1	0	2	1
Autoar	2	3	1	3	2	2	2	2	0	2	2	3	4	2	0
Alaton	2	4	1	4	2	2	2	1	1	1	0	0	0	3	2
Benth	**12**	3	**15**	**10**	5	3	6	4	3	1	4	5	5	4	1
dtr	6	9	8	**10**	**7**	**10**	6	7	**10**	4	3	5	5	4	7
gbr	2	0	1	2	1	5	3	4	2	1	1	0	1	1	3
hgbr	3	3	6	2	4	6	**11**	2	8	5	8	8	6	8	6
knn	2	4	2	1	3	3	3	3	7	4	4	3	5	1	4
lr	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lstm	7	6	3	4	2	1	2	4	4	4	5	2	**7**	1	3
lsvr	7	1	1	1	3	1	3	2	4	7	6	6	6	6	3
mlp	3	2	5	5	1	4	6	0	0	5	5	3	1	3	1
rfr	1	5	2	1	5	6	4	5	2	6	5	5	3	2	2
svm	6	3	2	2	**7**	3	7	4	3	4	8	7	5	4	1
tcn	5	5	7	5	5	4	2	7	3	3	4	2	5	5	6
xgb	1	3	3	6	6	3	4	**11**	**10**	**12**	**9**	**11**	**7**	**13**	**19**

(Table 10) and MAE (Table 11) all evidencing better forecast performance of the xgb model for horizons starting further in the future.

We also analysed the relative forecast performance of the models in terms of their average ranks across all locations as opposed to counting the number of locations they were the best model for. Table 12 shows the average ranks for the 30 day horizon across the different metrics and similar to the patterns of performance observed in the rank totals, the model by Benth and Šaltyte Benth (2005) is superior to the others. The reader is reminded that the two important metrics for the pricing of weather derivatives are the CAT and HDD APEs and for these, the superiority of the model by Benth and Šaltyte Benth (2005) is shared with an ML model, the lsvr. Table 13 shows the opposite horizon measuring average rank performance of the last 30 days of the test set, and here the xgb model's superiority in the rank totals is also repeated, especially and importantly for the two APE metrics. At the full year horizon shown in Table 14, knn is the best model for the APE metrics, whilst the auto arima model is superior for the other forecast metrics.

As we mentioned earlier, temperatures further out from the current known actuals present a greater risk to businesses and represent the temperature derivative pricing windows for which market participants would typically want to hedge against. A ski business is far more likely to hedge temperature risk months in advance of the snow season rather than wait a day before the start of the ski season to purchase temperature derivatives. Models that perform better at later horizons will be preferred to those that do well for near-term horizons at the expense of later horizons.

4.4 Statistical tests

To ascertain the statistical significance of the temporal performance bias we had observed in the rank summaries and averages, we also performed statistical tests on the ranked averages.

4.4.1 The Friedman and Nemenyi post hoc tests

We first performed the Friedman (Friedman 1937) test on the APE results per horizon and found significant differences at all horizons at the 5% level (see Friedman Test P values in Tables 15, 16, 17 and 18). The Friedman test is a non-parametric statistical test used to test significant differences between ranked performance across multiple test attempts, where the null hypothesis, in our case, was that there is no significant difference between the different model forecast performance across the different horizons. To confirm which pairs of algorithms differed significantly from each other, we also performed the Nemenyi (Nemenyi 1963) post hoc for the Friedman test. The Nemenyi test does pairwise comparison for unduplicated blocked data and is conducted following significant results from the Friedman test. Tables 15 and 16 show the results of the Nemenyi post hoc tests on CAT APE performance ranks of the models whilst Tables 17 and 18 show the Nemenyi results for the models in forecasting HDD temperature. The tables are ordered to show the best (or control) model at the top of each respective horizon, along with the average ranks and adjusted p-values from the Nemenyi test.

It is noticeable from the post hoc tests that even though there is a statistically significant difference between the models when looked at together, the superiority of the top performing models is not statistically significantly different from the others, except for those very

Table 7 Number of data sets model ranked best for HDD APE - most top ranked model per forecast horizon in bold font

Model	30	60	90	120	180	270	365	ISL	IISL	−270	−180	−120	−90	−60	−30
Arima	6	**14**	**14**	14	12	5	6	**21**	7	2	4	7	7	7	9
Arma	5	5	5	3	2	6	2	17	1	2	2	2	4	7	7
Autoar	8	5	3	2	1	4	3	16	0	1	3	6	6	8	6
Alaton	6	8	7	8	3	2	2	18	1	3	0	1	4	8	7
Benth	**15**	7	11	7	5	5	6	**21**	6	4	3	6	7	10	8
dtr	11	7	7	4	2	6	5	19	**11**	5	4	8	10	9	13
gbr	5	6	3	4	1	3	3	20	1	3	1	3	5	6	9
hgbr	8	5	5	4	2	6	**8**	15	9	**9**	**9**	9	9	12	11
knn	5	4	4	2	3	1	6	15	3	5	2	4	9	6	11
lr	4	3	2	1	0	0	0	9	0	0	0	1	4	4	4
lstm	10	7	7	5	4	2	2	19	3	3	5	3	**12**	5	7
lsvr	9	7	7	5	3	4	3	16	3	1	8	6	10	12	10
mlp	9	5	10	9	3	4	5	16	2	3	2	6	5	9	7
rfr	4	7	4	3	2	2	5	17	5	4	3	4	7	7	7
svm	7	4	4	3	2	5	4	15	3	8	5	9	8	9	7
tcn	13	12	11	**16**	**17**	**7**	4	17	7	4	6	6	11	7	8
xgb	4	7	6	5	3	3	1	16	3	8	8	**13**	11	**17**	**25**

poor performing models at the bottom. We could see hints of this in the rank and average results, where although clear patterns of superiority exist (see for example, xgb at the last 30 day horizon and Benth and Šaltyte Benth (2005) model at the 30 day horizon), no model completely dominates all the others by a huge margin. This result is perhaps not surprising and an indication that several of the models might not be the outright best but deliver good average performance nonetheless. We also assert that for pricing the derivative, one will still prefer the best model compared to the rest per horizon; for example, the xgb model for forecasting the last 30 days and the model by Benth and Šaltyte Benth (2005) for the first 30 days temperature, given their superiority in the number of locations they are better at, as well as their average ranked performance.

4.4.2 Blocked bootstrap tests

Since the 65 locations of the study were selected based on prior research and the weather derivative markets, it is not unlikely that some of the temperature data of the different locations used in the study are correlated. To include tests that are robust to location correlations, we also conducted blocked bootstrapped ranked tests of our results. To achieve this, we first clustered the 65 locations into blocks that preserve location correlation, computed rank-based statistics within each block and then built empirical distributions to test the significance of the block-based results.

To generate the bootstrapping blocks, we employed minimum distance *k*-means clustering. This ensured that locations that are similar will be put in the same block whilst limiting the risk of cross-mingling locations in different sub-regional weather zones (e.g., US East Coast locations will be separated from West Coast Locations). The 65 locations were grouped into 15 blocks, the raw forecast scores for each model were sampled with replacement a thousand times, and then these bootstrapped scores were converted into ranks within each block, similar to the Friedman Test.

Tables 19 and 20 show the results of the Wilcoxon signed rank tests (Wilcoxon 1992) performed on the averaged bootstrapped ranks for each model, across the locations in the different blocks, for both CAT and HDD temperatures across the various forms of the one-season (90 day) forecast. For both CAT and HDD forecasts, the block bootstrap results show a statistically significant difference between the top-performing model and all the others, implying that when we account for correlated locations, the difference in model performance is more statistically significant. The model by Benth and Šaltyte Benth (2005), for example, is statistically significantly better than all others when forecasting CAT and HDD temperatures for the first season, whilst knn is statistically significantly better at forecasting CAT and HDD temperatures for the final season of the year when we control for similarity in locations. Random forest and lstm are best for the mid-90-day horizon forecast of the CAT and HDD temperatures, respectively. This result shows that for temperature forecasts spanning a season (90 days) machine learning algorithms are statistically significantly better than the other state-of-the-art methods when we control for location similarity and consider the latter horizons where weather risk will be greatest and the need for hedging more important for businesses. ML models generalise to diverse locations better than the current state of the art at these business-critical horizons.

Table 8 Number of data sets model ranked best for DAT MAPE - most top ranked model per forecast horizon in bold font

Model	30	60	90	120	180	270	365	ISL	IISL	−270	−180	−120	−90	−60	−30
Arima	9	13	11	9	**10**	10	**14**	**11**	5	**12**	**13**	**10**	9	6	3
Arma	3	2	2	2	2	3	7	4	1	7	2	1	0	1	3
Autoar	1	0	1	0	2	3	6	4	1	5	2	0	2	0	2
Alaton	11	5	8	9	6	2	1	2	5	0	1	2	3	3	4
Benth	**12**	**18**	**14**	**13**	7	10	10	3	**7**	4	10	**10**	**10**	7	4
dtr	3	2	3	4	5	7	4	7	**7**	2	0	2	3	1	4
gbr	1	2	3	3	3	1	2	1	5	3	4	3	2	2	1
hgbr	3	5	3	5	8	**11**	5	6	**7**	3	2	4	3	6	4
knn	2	1	2	2	2	0	1	0	4	6	7	2	4	3	2
lr	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lstm	3	5	4	4	3	1	4	6	1	4	4	4	7	**14**	9
lsvr	5	5	3	3	4	3	1	1	6	3	3	3	4	5	2
mlp	1	2	4	2	4	6	3	5	4	2	2	4	4	1	3
rfr	4	0	2	2	2	2	4	6	4	5	6	7	4	7	3
svm	4	2	2	4	4	4	1	3	5	4	3	6	3	2	2
tcn	1	0	0	0	0	1	1	0	1	2	2	2	1	1	1
xgb	2	3	3	3	3	1	1	6	2	3	4	5	6	6	**18**

4.5 Performance by temperature moments clusters

Model generalisability is one of the main motivations behind our study, so we also investigate forecast performance across different groups of locations to reinforce the need to select models that are not tailor-made for just one particular location. We performed *k*-means clustering on the first four moments of the DATs (mean, standard deviation, skewness and kurtosis) to group locations by the descriptive statistics of their temperatures. *k*-means returned the four clusters shown earlier in Fig. 1 and we analysed the performance of the different models by these clusters. Clusters 1 and 3 include locations of generally lower latitudes and warmer average temperatures compared to Clusters 0 and 2. From the box plots of the CAT APEs shown in Fig. 8, Cluster 1 has the smallest average APE and the smallest deviation around that average. The Empirical Density Function plots shown in Fig. 9 also show Cluster 1 as the best performer with Cluster 3 next in performance while Clusters 0 and 2 perform relatively less well, particularly when we focus on the bulk performance excluding outliers. We performed Kolmogorov–Smirnov tests, which are non-parametric tests for checking if two samples came from the same distribution (Dimitrova et al. 2020) and the results shown in Table 21 confirm that the forecast performance are statistically significantly different across the four location clusters. We also observe location bias when we look at the HDD box plot in Fig. 8 but here it is the higher latitude locations (cluster zero and two) that perform better. The better temperature prediction performance for lower latitude or warmer temperatures has been hinted by other researchers including Oetomo and Stevenson (2005) and Šaltyte Benth and Benth (2012) who suggest the lower volatility in warmer temperatures as a cause but to our knowledge our observed location bias specifically for HDD temperature is not widely documented. One reason why the higher latitudes have better relative performance to lower latitudes for the HDD could be the fact that the higher latitudes are more likely to have non-zero HDD temperatures making them slightly easier to forecast compared to warmer locations. In warmer locations, temperatures might not dip below 65 (°F) or 18 (°C) often, returning a lot more zeroes in the actual HDD test set compared to higher latitudes with colder temperatures. The occurrence of zeroes is more difficult to forecast whilst it also contributes to a smaller denominator in the APE equation (Eq. 9) leading to higher APEs. It is also very interesting to note that the HDD performance is generally worse for all locations compared to the CAT performance.

This analysis of the clusters confirms location bias within temperature modelling performance and highlights the need for models that have been trained on data from many varied locations and can therefore generalise better.

4.6 Pricing the temperature derivative

To demonstrate economic value and highlight the importance of accurate temperature forecasts to pricing weather derivatives, we simulated pricing a seasonal derivative written on HDDs for two models with contrasting performance across two different forecast horizons. The first seasonal prices cover the 90 days that start immediately after the training data set (90 day horizon) to represent a market participant trying to hedge weather risk for a season starting a day after the purchase of the derivative. The second experiment analyses prices for the mid 90-day horizon and mimics buying the derivative to hedge temperature risk also over a season but 4.5 months in advance of the season start. As mentioned earlier, the

Table 9 Number of data sets model ranked best for DAT WMAPE - most top ranked model per forecast horizon in bold font

Model	30	60	90	120	180	270	365	ISL	IISL	−270	−180	−120	−90	−60	−30
Arima	8	**11**	8	8	**10**	**9**	**8**	11	4	8	4	4	5	6	3
Arma	2	2	2	3	4	6	6	4	2	2	2	0	1	1	2
Autoar	2	2	2	1	6	7	**8**	5	0	2	1	1	0	0	1
Alaton	4	4	3	8	2	2	1	1	3	1	2	1	1	2	3
Benth	**13**	9	**13**	**12**	7	8	4	1	8	1	4	4	3	2	2
dtr	3	3	5	6	2	6	6	9	8	4	3	5	4	5	4
gbr	2	2	3	3	2	1	3	1	3	3	2	2	7	4	0
hgbr	2	4	4	6	8	8	7	5	**10**	4	7	7	4	6	6
knn	0	2	2	1	2	0	1	0	2	5	6	3	2	2	4
lr	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lstm	7	5	5	2	3	0	2	3	1	3	2	2	8	9	9
lsvr	9	6	4	3	2	2	2	2	5	1	5	5	4	4	2
mlp	3	6	4	2	2	5	6	3	1	5	5	6	3	2	1
rfr	4	0	4	2	6	5	3	5	6	7	3	7	6	6	4
svm	2	4	3	4	4	3	3	3	5	4	4	5	4	4	2
tcn	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
xgb	4	5	3	4	5	3	5	**12**	7	**15**	**15**	**13**	**13**	**12**	**21**

second situation better represents what happens in the real derivative market, but we have also included the first for completeness and to show the contrasting performances in terms of price impact for the different models.

To achieve this, we simulated 10, 000 future temperature paths by combining the empirical out-of-sample forecasts error from the test data not seen by any of the models with the forecasts from each respective model for the horizon of interest. The OU, AR and Neural Network models can be set up to generate probabilistic forecasts, but since most of our classical ML models are deterministic models, we implemented this Monte Carlo simulation approach to be able to more easily compare the results without overcomplicated transformations. To facilitate the pricing performance comparison, we selected the models with the best block bootstrapped forecast performance at the 90-day (benth) and mid-90-day (lstm) horizons respectively (see Table 20). Since we were attempting to price HDDs for this experiment, we selected two locations with a winter season for the 90-day and mid 90-day horizon (see Figs. 10 and 11) across both the northern and southern hemispheres to cover as diverse a location as we possibly could. Figure 10 shows the forecast horizon for the selected 90-day experiment for Innsbruck, Austria, and Fig. 11 shows the mid-90-day forecast for Archfield Airport in Australia over their winter period.

The Monte Carlo simulations allowed us to capture the empirical distribution of the HDD outcomes, including skewness and other non-Gaussian features. We were then able to generate a distribution of the payoffs from the simulated HDD forecasts for each model by a simple pricing formula (Eq. 12 where α is the tick size, H the HDDs over the forecast horizon and K is the strike) by multiplying the above strike accumulated HDD over the season by the tick size which was set at 20 for both experiments. Taking the expectation from the payoff distribution, we were able to calculate a simple Call option price using the discounted value of the expected pay off (see Eq. 13 where e is Euler's Number, r the risk free rate set at 0.01 for this simulation, TTM is time to maturity in years and $\mathbb{E}$ the expected or mean payoff). This simplified pricing setup allowed us to assess the payoff characteristics of the two models and provide insights into the effect of the different forecast performance of the two models across the different horizons.

$$\text{Call Payoffs} = \alpha \cdot \max(H_{\text{sums}} - K, 0), \tag{12}$$

$$\text{Option Price} = e^{-r \cdot TTM} \cdot \mathbb{E}[\text{Call Payoffs}]. \tag{13}$$

From the results in Figs. 12 and 13, we can see the lstm underestimates the payoff of the 90-day HDD over the season compared to the more accurate benth model, highlighting the risk in using a model that is not accurate for this horizon. For the mid90 payoffs (see Fig. 13), it is the poorer performing benth model which underestimates the temperature risk at this horizon. We again assert that most market participants will be looking to hedge risks with a forecast lead, so the mid90 pricing horizon represents the more likely scenario and for this the lstm model, which outperforms benth and all other models at this horizon, is the best choice for modelling temperature in order to price weather derivatives. Table 22 shows the risk metrics of the payoff distributions where μH is the mean and σH the variance of the cumulative HDD distribution, Mean, Stdv, Skew and Kurt describe the call payoff distribution, and VAR5 is Value at Risk at 5%. The price is the expected discounted payoff for each model horizon combination. It is clear from the metrics table that the worse performing

Table 10 Number of data sets model ranked best for HDD RMSE - most top ranked model per forecast horizon in bold font

Model	30	60	90	120	180	270	365	ISL	IISL	−270	−180	−120	−90	−60	−30
Arima	7	10	9	8	11	**11**	5	**22**	**11**	5	3	2	7	7	6
Arma	6	6	5	6	4	2	3	**22**	4	2	0	1	5	6	7
Autoar	5	5	4	4	3	7	7	18	1	1	2	1	4	5	5
Alaton	8	4	2	5	3	1	1	13	3	1	1	1	2	3	4
Benth	**20**	**14**	**19**	**18**	**13**	10	**11**	**22**	10	8	9	7	7	8	7
dtr	8	7	5	5	4	6	4	**22**	**11**	4	3	5	6	8	10
gbr	4	3	3	3	3	2	3	15	4	1	2	5	7	5	7
hgbr	8	5	3	3	3	4	4	17	9	3	4	6	8	11	8
knn	4	4	3	3	1	1	3	18	3	7	5	4	7	5	6
lr	2	2	2	2	2	3	4	1	0	4	4	4	4	4	4
lstm	7	9	7	8	6	1	2	21	2	1	2	4	8	14	13
lsvr	11	11	7	5	2	5	2	16	3	3	3	7	9	7	9
mlp	6	6	4	4	2	4	7	16	5	4	3	3	7	8	7
rfr	7	4	4	3	1	2	3	19	3	4	7	7	13	13	12
svm	10	9	8	8	5	3	4	15	5	6	5	7	10	10	8
tcn	2	2	2	2	0	0	0	3	0	0	0	2	1	3	3
xgb	7	6	5	5	2	3	2	19	4	**11**	**12**	**13**	**14**	**15**	**18**

model for each horizon leads to an underestimation of the distribution variance/risk and consequently, the payoff and the price of the option. This underestimation is much bigger in the case of the benth model for the mid 90-day HDD call price, mirroring its much poorer performance compared to the lstm model at this horizon (see Table 18).

4.7 Operational deployment of ML models

Temperature forecasting is a core component of temperature derivative pricing, and we expect forecasting models to form an integral part of any operational pricing pipeline. In practice, pricing is typically carried out by trading houses or exchanges (e.g., CME), who will generally have the computational capacity to run the more resource-intensive models (see Table 4 for sample model training core times). We also expect these models to be trained offline to price derivatives at a seasonal rather than daily cadence thus limiting the need for live performance. However, for real time or near real time deployment, computational cost becomes an important factor, particularly for organisations with limited compute resources. In such settings, choosing a model will need to be a balance between forecast accuracy, runtime and hardware requirements. Spatio-temporal models, which have been used in meteorological forecasting (Grover et al. 2015), could also offer computational advantages by enabling joint modelling across locations. However, additional work would be required to translate these raster based predictions into the station level forecasts currently needed for pricing weather derivatives.

Operational deployment also requires automated data sourcing and preprocessing. For locations where temperature data is not readily available, for example, stations not included in the NOAA repository, practical considerations should include the ease of acquiring the data and the ability to automate its ingestion and cleaning. In addition, any production pipeline must incorporate model monitoring mechanisms to ensure that forecast performance remains within acceptable thresholds and that failures or outlier behaviour are detected promptly. These aspects, while outside the scope of this paper, will be essential for robust real time deployment.

Finally, the current implementation relies heavily on Python libraries, which evolve over time. A production grade forecasting system must therefore be resilient to library updates and maintain reproducibility across software versions.

4.8 Discussion

Our results reveal some interesting insights about the challenge of forecasting temperature for the purpose of pricing weather derivatives. First, as observed by other researchers, temperature prediction generally becomes more difficult the further out from the forecasting point one has to predict. This can clearly be seen in the 3D performance plots (see Figs. 6 and 7) and translates into greater risk for hedging temperature at later forecast horizons whilst highlighting the importance of developing models that work well at those horizons. The second, which is novel to this study, is model performance bias across the different horizons, where some models perform better than others at the near-term horizons and vice versa. The model by Benth and Šaltyte Benth (2005), which is one of the most cited in the temperature derivative literature, generally performs best at short-term horizons close to the realised temperature. Both the model by Benth and Šaltyte Benth (2005) and Ala-

Table 11 Number of data sets model ranked best for HDD MAE - most top ranked model per forecast horizon in bold font

Model	30	60	90	120	180	270	365	ISL	IISL	−270	−180	−120	−90	−60	−30
Arima	9	9	8	9	**10**	**10**	4	20	4	3	2	2	9	6	5
Arma	5	5	4	5	5	5	5	20	3	0	1	1	5	6	7
Autoar	6	5	4	4	7	7	5	19	1	1	2	2	4	5	6
Alaton	9	6	4	8	2	1	1	11	5	1	1	1	2	3	3
Benth	**15**	9	**13**	**10**	8	7	**7**	21	7	5	7	4	5	7	6
dtr	9	8	6	6	5	7	4	**24**	**10**	6	3	6	8	9	9
gbr	5	5	4	4	2	2	2	15	4	2	1	3	9	8	4
hgbr	7	6	6	6	3	5	**7**	22	**10**	5	5	7	7	10	10
knn	4	3	3	3	1	2	2	20	4	4	6	4	6	8	8
lr	2	2	2	2	2	2	3	1	0	3	4	4	4	4	4
lstm	9	9	8	9	6	0	2	20	2	3	3	3	9	10	13
lsvr	12	**11**	8	6	4	4	5	16	7	3	4	7	8	8	7
mlp	7	8	6	4	2	5	**7**	17	4	9	6	6	7	7	6
rfr	6	3	3	2	1	1	1	19	7	3	4	7	10	12	9
svm	7	8	6	6	4	3	4	14	4	3	3	6	8	9	7
tcn	2	2	2	2	0	0	0	3	0	0	0	2	1	3	4
xgb	8	8	5	6	3	4	6	22	6	**14**	**13**	**14**	**17**	**17**	**26**

ton et al. (2002) are Ornstein-Uhlenbeck models that explicitly model future forecasts as innovations from the latest realised actual temperature. It is therefore to be expected that at short horizons, they perform better than other methods that do not explicitly predict future temperatures as a change to the current realised actual ones. The superiority of the model by Benth and Šaltyte Benth (2005) diminishes as forecasts further away from the realised actuals are included in the performance tests and turns in favour of ML models for forecast horizons closer to the end of the forecast year. The xgb model, for example, shows a very consistent superiority at forecasting the last 30 days of the year, whereas its performance at near-term horizons is not that superior. The near-term horizons are not where we expect a lot of market participants to be interested in, as they represent less risk to businesses since business decisions and commitments will need to be made in advance. Also, more accurate meteorological weather predictions from physical models are widely available for near-term horizons, making weather in this horizon less of a risk or an unknown.

It is also noteworthy that although there are clearly superior models at certain horizons, this does not translate to total statistical dominance. In the results, the xgb model for example, ranks best in a lot more locations than others at the -30 day horizon (see Table 6) and the rank averages (see Table 13) but is only statistically significantly better than less than half of the other models (see Table 15). This suggests greater similarity in the ranked average model performance than is immediately evident and hints that model architecture is not necessarily the only requirement for model dominance.

We do see statistically significant differences in the model performance when we account for location similarity, however. As an example the benth and rfr models are the best average ranked models for the CAT 90 day and mid90 day horizons respectively but not statistically significantly different from most of the other models (see Table 16) for the Nemenyi post hoc test but we do see statistically significant difference in the blocked bootstrap average ranks statistical tests to all the other metrics (see Tables 19). Since the blocked bootstrap tests allow us to compare performance while controlling for location similarity, our results suggest that there is also some location bias within the models. That is, models perform better not just for certain horizons but certain locations too.

It is also interesting that several ML model architectures, including classical ML models, outperform the state-of-the-art OU models at the later horizons. For example, both rfr and knn are amongst the best at predicting mid90 and -90day CAT temperatures (see Table 16), representing great options for pricing seasonal temperature derivatives using various ML architectures immediately, but also implying different ML architectures are suitable for this domain. If the ML models that outperform the current state of the art were restricted to the computationally expensive neural network flavours, this would have posed a challenge to suggest ML models as an alternative.

Despite the range of viable ML architectures considered, our results indicate that some models are poorly suited to this forecasting domain. In particular, the linear regression model frequently produced unrealistically high temperature forecasts at various horizons, reflecting the limitations of its rigid linear functional form and its tendency to generate unstable parameter estimates when applied to complex, non stationary temperature dynamics. Although previous studies have noted that linear frameworks may be inadequate for temperature modelling, our findings provide the first systematic evidence of linear models exhibiting explosive behaviour when forecasting daily average temperature across multiple locations.

Table 12 30 Day rank averages - model with best average rank in bold

Model	CAT APE	HDD APE	DAT APE	HDD RMSE	HDD MAE	DAT WMAPE
Arima	8.88	9.28	6.77	8.98	8.71	7.82
Arma	9.98	9.68	9.24	9.22	9.11	9.08
Autoar	8.21	8.42	7.78	7.62	7.82	7.54
Alaton	13.72	13.31	10.77	12.11	11.92	12.24
Benth	8.28	**8.01**	**5.65**	**6.59**	**7.56**	**6.98**
dtr	9.09	9.12	10.64	9.63	9.62	10.18
gbr	10.65	10.69	10	10.05	9.74	9.63
hgbr	9.19	9.14	9.02	8.8	8.35	8.75
knn	8.84	9.28	8.78	8.77	8.72	8.71
lr	18	17.15	18	17.55	17.53	18
lstm	8.5	8.55	7.62	8.1	8.73	8.07
lsvr	**8.05**	8.3	9.09	8.67	8.49	8.46
mlp	8.28	8.02	8.8	8.35	8.35	8.47
rfr	10.15	10.35	9.99	9.64	9.73	9.72
svm	8.85	9.12	10.35	9.32	9.87	10.12
tcn	8.87	8.35	14.68	15.3	15.12	15.37
xgb	12.15	12.45	12.48	10.55	9.88	10.55

Table 13 Last 30 day rank averages - model with best average rank in bold

Model	CAT APE	HDD APE	DAT APE	HDD RMSE	HDD MAE	DAT WMAPE
Arima	11.78	11.98	10.02	11.31	11.76	11.25
Arma	8	8.18	7.18	7.68	7.87	7.68
Autoar	7.8	8.02	6.94	7.31	7.62	7.32
Alaton	14.63	14.36	14.59	14.92	15.15	14.83
Benth	13.75	13.48	11.25	12.51	12.84	12.49
dtr	8.3	8.22	9.36	8.77	8.42	8.54
gbr	9.66	9.54	9.52	9.72	9.82	9.8
hgbr	9.06	8.98	9.47	8.94	8.74	8.76
knn	7.87	7.78	6.92	7.05	6.89	6.79
lr	18	17.35	18	16.95	16.94	18
lstm	8.17	8.75	7.76	8.2	8.22	8.08
lsvr	9.22	9.05	9.78	8.95	9.15	9.58
mlp	9.73	9.55	9.62	9.58	9.3	9.41
rfr	7.35	7.49	**6.82**	6.61	6.63	6.63
svm	9.42	9.25	9.94	9.19	9.3	9.52
tcn	10.53	10.93	14.99	15.32	14.64	14.9
xgb	**6.55**	**6.3**	7.48	**6.12**	**5.88**	**6.12**

Within-class superiorities also begin to emerge from our results; the auto arima model generally does better than the other AR models, whilst the benth model is the better of the OU models. Alaton et al. (2002) also model and forecast temperature as a direct difference from known actuals but their model does not perform as well as the model by Benth and Šaltyte Benth (2005). This is potentially because the Alaton et al. (2002) model was built specifically for Stockholm and does not generalise well to other locations.

Table 14 365 Day rank averages - model with best average rank in bold

Model	CAT APE	HDD APE	DAT APE	HDD RMSE	HDD MAE	DAT WMAPE
Arima	11.29	9.79	7.01	9.17	9.89	9.53
Arma	8.09	7.76	7.27	7.79	7.66	7.49
Autoar	7.85	**7.58**	**5.42**	**5.92**	**5.88**	**5.75**
Alaton	14.06	12.95	14.25	15.29	15.68	15.42
Benth	10.85	9.65	7.76	9.13	10.21	10.26
dtr	9.44	9.69	10.72	10.42	10.04	9.76
gbr	9.83	9.44	8.95	9.12	9.25	9.11
hgbr	9.12	9.5	9.52	8.9	8.48	8.61
knn	**7.74**	**7.58**	8.15	7.32	7	7.08
lr	18	17.93	18	16.95	17.19	18
lstm	8.04	7.78	8.14	8.78	9.05	9.22
lsvr	9.43	9.35	8.72	8.98	9.05	9.33
mlp	8.25	8.7	8.54	7.8	7.95	7.98
rfr	7.75	8.34	8.85	8.12	7.48	7.38
svm	9.25	10.14	10.32	9.68	9.67	9.79
tcn	10.14	10.8	15.73	16.22	16.23	16.12
xgb	10.58	12.65	12.42	10.15	9	8.89

Table 15 CAT Nemenyi post hoc test for 30 day, 365 day and -30 day horizon

CAT 30			CAT 365			CAT -30		
Friedman test (P_{frd} 3.15e−47)			Friedman test (P_{frd} 1.88e−50)			Friedman test (P_{frd} 6.98e−68)		
Model	AvgRank	P_{nem}	Model	AvgRank	P_{nem}	Model	AvgRank	P_{nem}
lsvr	8.05	1.00	knn	7.74	1.00	xgb	6.55	1.00
autoar	8.21	0.90	rfr	7.75	0.90	rfr	7.35	0.90
mlp	8.28	0.90	autoar	7.85	0.90	autoar	7.80	0.90
benth	8.28	0.90	lstm	8.04	0.90	knn	7.87	0.90
lstm	8.50	0.90	arma	8.09	0.90	arma	8.00	0.90
knn	8.84	0.90	mlp	8.25	0.90	lstm	8.17	0.90
svm	8.85	0.90	hgbr	9.12	0.90	dtr	8.30	0.85
tcn	8.87	0.90	svm	9.25	0.90	hgbr	9.06	0.28
arima	8.88	0.90	lsvr	9.43	0.89	lsvr	9.22	0.19
dtr	9.09	0.90	dtr	9.44	0.87	svm	9.42	0.11
hgbr	9.19	0.90	gbr	9.83	0.60	gbr	9.66	**0.04**
arma	9.98	0.73	tcn	10.10	0.36	mlp	9.73	**0.04**
rfr	10.20	0.60	xgb	10.60	0.11	tcn	10.50	**0.001**
gbr	10.70	0.20	benth	10.80	**0.04**	arima	11.80	**0.001**
xgb	12.10	**0.001**	arima	11.30	**0.006**	benth	13.70	**0.001**
alaton	13.70	**0.001**	alaton	14.10	**0.001**	alaton	14.60	**0.001**
lr	18.0	**0.001**	lr	18.00	**0.001**	lr	18.00	**0.001**

Finally, our pricing experiment showed the economic importance of building accurate models to forecast temperature for pricing weather derivatives. The payoff risks table (Table 22) evidenced the implications of underestimating the variability in the distribution of future temperature used in pricing weather derivatives and how that can lead to underestimating the risks, payoff and temperature derivative prices.

Table 16 CAT Nemenyi post hoc test for 90 day, mid90 day and -90 day horizon

CAT 90			CAT Mid90			CAT -90		
Friedman test (P_{frd} 3.70e−58)			Friedman test (P_{frd} 3.62e−61)			Friedman test (P_{frd} 2.02e−59)		
Model	AvgRank	P_{nem}	Model	AvgRank	P_{nem}	Model	AvgRank	P_{nem}
benth	7.19	1.00	rfr	7.15	1.00	autoar	7.03	1.00e+00
arma	7.91	0.90	autoar	7.50	0.90	arma	7.40	0.90
knn	7.93	0.90	arma	7.52	0.90	knn	7.68	0.90
lstm	8.01	0.90	mlp	7.99	0.90	lstm	8.17	0.90
autoar	8.05	0.90	knn	8.24	0.90	rfr	8.32	0.90
mlp	8.18	0.90	hgbr	8.54	0.90	lsvr	8.35	0.90
arima	8.56	0.90	lstm	8.64	0.90	dtr	8.79	0.82
tcn	9.04	0.80	svm	8.89	0.86	mlp	9.18	0.55
hgbr	9.24	0.65	xgb	8.98	0.78	svm	9.38	0.38
lsvr	9.42	0.53	dtr	9.10	0.70	xgb	9.48	0.32
rfr	9.55	0.42	lsvr	9.57	0.35	gbr	9.77	0.14
dtr	9.91	0.17	gbr	9.68	0.27	hgbr	9.80	0.13
svm	10.20	0.08	tcn	10.60	**0.01**	tcn	10.40	**0.02**
gbr	11.10	**0.002**	arima	11.70	**0.001**	arima	11.40	**0.001**
xgb	13.00	**0.001**	benth	13.00	**0.001**	benth	11.50	**0.001**
alaton	14.50	**0.001**	alaton	15.40	**0.001**	alaton	15.10	**0.001**
lr	18.00	**0.001**	lr	17.40	**0.001**	lr	18.00	**0.001**

Table 17 HDD Nemenyi post hoc test for 30 day, 365 day and -30 day horizon

HDD 30			HDD 365			HDD -30		
Friedman test (P_{frd} 6.62e−17)			Friedman test (P_{frd} 4.23e−46)			Friedman test (P_{frd} 8.58e−48)		
Model	AvgRank	P_{nem}	Model	AvgRank	P_{nem}	Model	AvgRank	P_{nem}
benth	8.01	1.00	knn	7.58	1.00	xgb	6.30	1.00
mlp	8.02	0.90	autoar	7.58	0.90	rfr	7.49	0.90
lsvr	8.30	0.90	arma	7.76	0.90	knn	7.78	0.90
tcn	8.35	0.90	lstm	7.78	0.90	autoar	8.02	0.89
autoar	8.42	0.90	rfr	8.34	0.90	arma	8.18	0.77
lstm	8.55	0.90	mlp	8.70	0.90	dtr	8.22	0.74
svm	9.12	0.90	lsvr	9.35	0.84	lstm	8.75	0.34
dtr	9.12	0.90	gbr	9.44	0.77	hgbr	8.98	0.19
hgbr	9.14	0.90	hgbr	9.50	0.72	lsvr	9.05	0.15
arima	9.28	0.90	benth	9.65	0.60	svm	9.25	8.22e−02
knn	9.28	0.90	dtr	9.69	0.59	gbr	9.54	**0.029**
arma	9.68	0.90	arima	9.79	0.49	mlp	9.55	**0.028**
rfr	0.10	0.44	svm	10.10	0.25	tcn	10.90	**0.001**
gbr	0.11	0.19	tcn	10.80	**0.02**	arima	12.00	**0.001**
xgb	0.12	**0.001**	xgb	12.60	**0.001**	benth	13.50	**0.001**
alaton	0.13	**0.001**	alaton	13.00	**0.001**	alaton	14.40	**0.001**
lr	0.17	**0.001**	lr	17.90	**0.001**	lr	17.40	**0.001**

Table 18 HDD Nemenyi post hoc test for 90 day, mid 90day and -90 day horizon

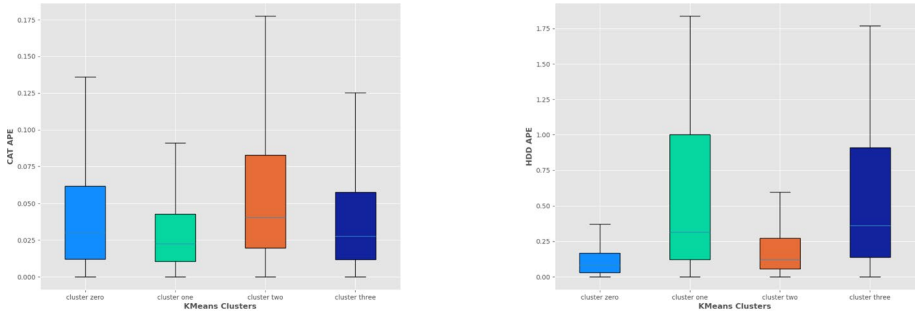
HDD 90			HDD Mid90			HDD -90		
Friedman test (P_{frd} 1.84e−15)			Friedman test (P_{frd} 7.48e−19)			Friedman test (P_{frd} 1.38e−48)		
Model	AvgRank	P_{nem}	Model	AvgRank	P_{nem}	Model	AvgRank	P_{nem}
benth	6.78	1.00e+00	lstm	7.83	1.00	autoar	7.42	1.00
arima	7.33	0.90	autoar	8.17	0.90	arma	7.55	0.90
tcn	7.65	0.90	arima	8.54	0.90	knn	7.73	0.90
lstm	7.83	0.90	arma	8.56	0.90	lsvr	8.24	0.90
mlp	8.40	0.90	knn	8.72	0.90	rfr	8.26	0.90
knn	8.49	0.89	mlp	8.79	0.90	lstm	8.57	0.90
autoar	8.53	0.87	rfr	9.20	0.90	dtr	8.68	0.90
arma	8.67	0.76	hgbr	9.31	0.90	mlp	9.22	0.80
lsvr	8.78	0.67	benth	9.58	0.84	xgb	9.38	0.68
hgbr	9.97	**0.03**	gbr	9.70	0.76	svm	9.48	0.60
rfr	10.00	**0.03**	dtr	9.84	0.64	gbr	9.62	0.50
gbr	10.70	**0.001**	lsvr	10.20	0.39	hgbr	9.77	0.37
svm	10.80	**0.001**	svm	10.30	0.32	tcn	10.20	0.12
dtr	11.00	**0.001**	tcn	11.50	**0.003**	arima	11.60	**0.001**
alaton	13.10	**0.001**	xgb	12.00	**0.001**	benth	11.80	**0.001**
xgb	14.00	**0.001**	lr	13.00	**0.001**	alaton	14.50	**0.001**
lr	17.20	**0.001**	alaton	14.40	**0.001**	lr	17.30	**0.001**

Table 19 CAT block bootstrap Wilcoxon test for 90 day, mid 90day and -90 day horizon

CAT 90			CAT Mid90			CAT -90		
Model	AvgRank	P_{wcx}	Model	AvgRank	P_{wcx}	Model	AvgRank	P_{wcx}
benth	1.88	nan	rfr	1.81	nan	knn	1.96	nan
knn	3.31	**1.48e−57**	knn	3.85	**5.51e−142**	lsvr	3.38	**1.31e−65**
lstm	3.69	**1.08e−91**	arma	4.04	**1.19e−102**	rfr	4.01	**1.64e−135**
mlp	4.08	**8.73e−84**	hgbr	4.36	**3.61e−110**	lstm	4.80	**4.43e−143**
rfr	6.55	**3.53e−165**	autoar	5.18	**2.19e−142**	mlp	5.51	**8.10e−159**
hgbr	7.44	**8.07e−162**	mlp	6.13	**1.09e−154**	xgb	7.00	**2.30e−165**
tcn	7.48	**3.27e−166**	dtr	6.66	**5.26e−155**	hgbr	7.77	**1.36e−167**
lsvr	8.04	**7.81e−165**	xgb	7.18	**1.38e−148**	tcn	8.11	**4.11e−166**
arma	9.11	**2.99e−145**	lstm	7.70	**1.96e−164**	autoar	8.85	**3.35e−118**
autoar	9.48	**2.39e−155**	lsvr	9.71	**2.40e−166**	arma	9.24	**1.98e−141**
arima	9.93	**2.37e−166**	tcn	11.10	**2.14e−167**	benth	9.98	**1.20e−167**
xgb	10.40	**4.16e−167**	gbr	12.20	**1.92e−166**	alaton	11.40	**2.05e−169**
alaton	11.40	**3.82e−167**	arima	12.90	**3.13e−168**	dtr	11.90	**2.01e−164**
dtr	13.20	**2.21e−166**	svm	13.60	**3.85e−167**	arima	12.00	**4.82e−169**
svm	14.40	**2.49e−169**	benth	14.30	**5.20e−169**	svm	14.50	**8.19e−171**
gbr	15.70	**5.90e−169**	alaton	15.60	**9.23e−169**	gbr	15.70	**2.20e−168**
lr	17.00	**7.27e−176**	lr	16.80	**2.19e−169**	lr	17.00	**7.18e−170**

Table 20 HDD block bootstrap Wilcoxon test for 90 day, mid 90day and -90 day horizon

HDD 90			HDD Mid90			HDD -90		
Model	AvgRank	P_{wcx}	Model	AvgRank	P_{wcx}	Model	AvgRank	P_{wcx}
benth	6.31	nan	lstm	6.96	nan	knn	8.31	nan
lstm	6.78	**6.74e−52**	autoar	7.26	**2.39e−13**	lsvr	8.36	**4.54e−02**
mlp	7.29	**3.21e−56**	arima	7.71	**4.51e−15**	rfr	8.58	**1.35e−17**
knn	7.75	**4.40e−60**	arma	7.81	**1.71e−39**	lstm	8.60	**1.32e−17**
lsvr	8.01	**6.02e−60**	knn	7.82	**2.93e−38**	mlp	8.69	**2.99e−18**
rfr	8.50	**1.63e−60**	mlp	7.90	**4.48e−32**	tcn	8.71	**8.36e−14**
hgbr	8.60	**2.29e−60**	hgbr	8.24	**5.31e−40**	xgb	8.81	**2.58e−18**
tcn	8.72	**1.99e−57**	benth	8.40	**2.79e−37**	hgbr	8.92	**2.27e−18**
arima	8.75	**1.81e−44**	rfr	8.71	**7.29e−50**	arma	9.04	**1.18e−14**
autoar	9.18	**2.79e−59**	dtr	8.90	**1.74e−51**	autoar	9.05	**8.57e−15**
arma	9.28	**5.57e−60**	gbr	9.44	**1.82e−54**	benth	9.06	**2.48e−18**
xgb	9.34	**2.07e−60**	lsvr	9.82	**1.99e−54**	alaton	9.21	**1.50e−18**
alaton	9.86	**7.34e−61**	xgb	10.30	**9.79e−55**	dtr	9.27	**4.30e−18**
dtr	10.30	**4.60e−61**	svm	10.60	**1.79e−54**	arima	9.29	**1.86e−18**
svm	11.20	**5.25e−63**	tcn	10.60	**1.58e−54**	svm	9.62	**2.03e−18**
gbr	11.30	**5.56e−62**	alaton	11.10	**1.01e−54**	gbr	9.66	**2.24e−18**
lr	11.80	**1.86e−65**	lr	11.30	**1.03e−54**	lr	9.80	**1.64e−18**



(a) CAT APE (b) HDD APE

Fig. 8 Forecast performance by the location clusters shown in Fig. 1

5 Conclusion

In this study, we explored the effectiveness of different machine learning architectures for modelling daily average temperature for the purpose of pricing temperature derivatives. We compared these machine learning models to current state-of-the-art approaches deployed for pricing temperature derivatives across a wide range of locations around the world. We analysed model performance for 65 locations across various forecast horizons that closely represent the pricing windows a typical business would want to hedge its temperature risks against.

Besides focusing on actual business risk windows more precisely, our study is novel in its application of different models to a wide range of locations, allowing for the selection of

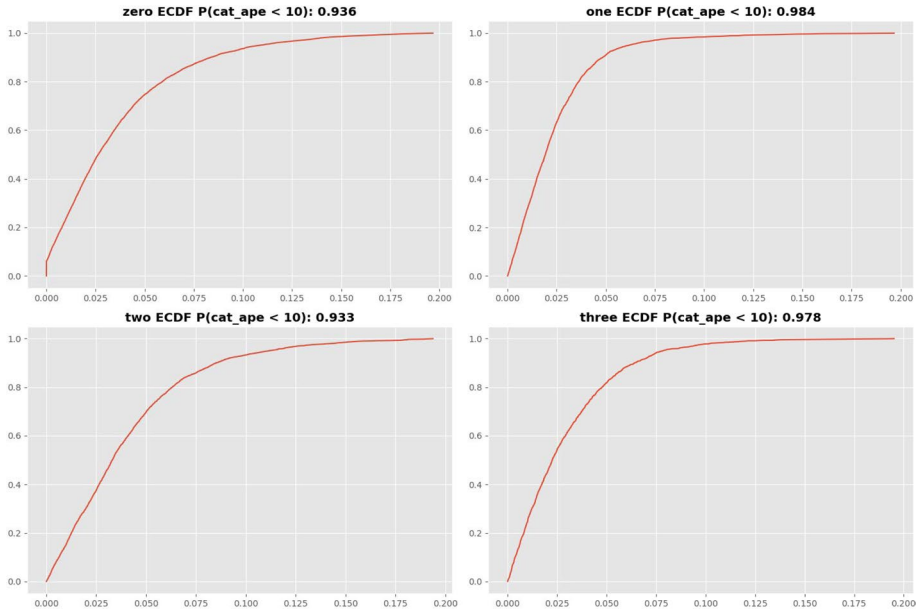


Fig. 9 Empirical CDF of CAT APE across KNN-derived location clusters

Table 21 Kolmogorov–Smirnov *p*-values

	Cluster 0	Cluster 1	Cluster 2
Cluster 1	5.32×10^{-43}		
Cluster 2	3.22×10^{-44}	7.60×10^{-116}	
Cluster 3	5.96×10^{-7}	4.45×10^{-17}	6.64×10^{-34}

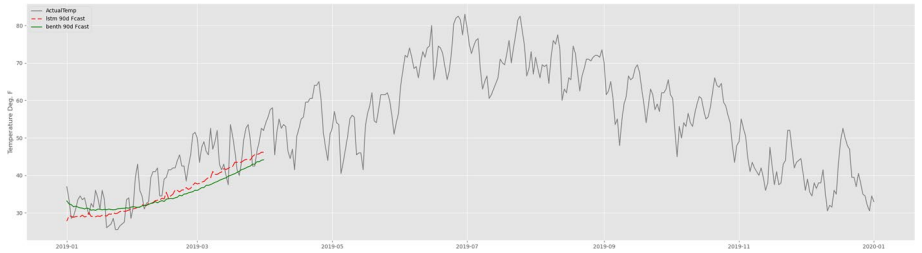


Fig. 10 LSTM and benth model 90 day forecast for Innsbruck-Austria

a model that can generalise well in forecasting DAT needed for pricing temperature derivatives. We differentiated between near-term and future-term forecast horizons, which are better suited for pricing weather derivatives, and identified the best models for this risk window across a wide range of locations. We also differentiated model performance by location, identifying models whose performance is robust to location correlation and are therefore good candidates for pricing temperature derivatives across a lot of different locations with uncorrelated weather.

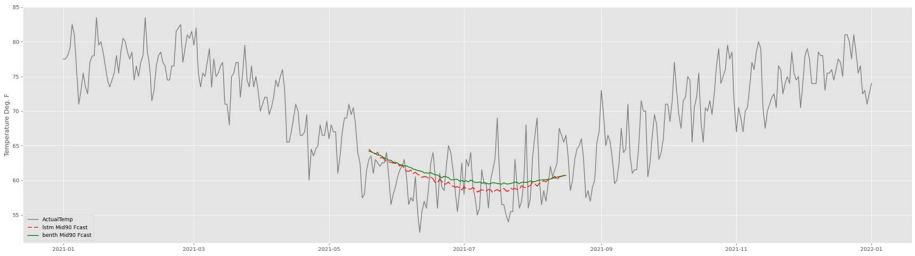


Fig. 11 LSTM and benth model mid 90 day forecast for Archfield Airport-Australia

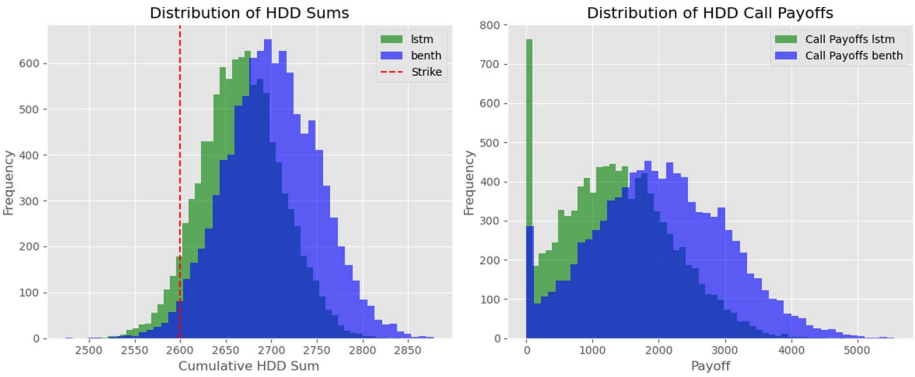


Fig. 12 LSTM and benth models 90 day HDD and call option payoff distributions for Innsbruck

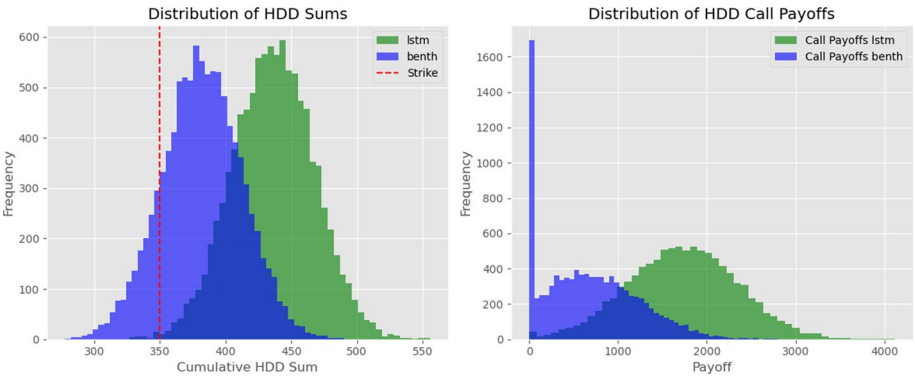


Fig. 13 lstm and benth models Mid90 day HDD and call option payoff distributions for Archfield Airport

We priced temperature derivatives for two very different locations, demonstrating the economic implications of model forecasting performance on temperature derivative payoffs and pricing.

Our results show that machine learning algorithms outperform the other approaches in forecasting temperature for later-start horizons, with several ML architectures proving to be equal to, if not better than, the current state-of-the-art models. The model by Benth and

Table 22 Risk metrics for HDD payoff distributions

HDD 90			HDD Mid90	
Metric	lstm	Benth	lstm	Benth
μH	2667.08	2700.99	436.38	381.97
σH	44.23	50.82	30.98	30.43
Price	1362.41	2023.52	1723.88	684.82
Mean	1365.78	2028.51	1728.14	686.51
Stdv	836.65	995.07	617.94	531.94
Skew	0.31	0.14	0.01	0.53
Kurt	−0.36	−0.28	−0.07	0.31
VAR5	0.00	349.91	716.42	0.00

Šaltyte Benth (2005) outperformed all others at short-term horizons, and we found a statistically significant difference for their model at the 90-day horizon over all other models. The XGBoost model was better at predicting temperature for later forecasting windows and was statistically significantly better than all others at the -30 day horizon, whilst Random Forest and KNN were better at the mid 90day and last 90day respectively. For the purpose of hedging weather risk, models that do better at later horizons will be preferred since there is greater weather risk at those horizons, and this makes ML models a better alternative to the current state of the art for modelling and forecasting weather derivative temperature.

Future research could focus on deepening our understanding of how machine learning model architecture influences temperature forecast performance, with the aim of identifying design principles that improve existing models. Hybrid approaches that combine the strengths of different architectures may also offer benefits for the various components of the temperature forecasting problem. Another promising direction is the development of richer feature sets and temperature covariates, moving beyond those used here, which were chosen to mirror aspects of the Ornstein–Uhlenbeck framework.

In addition, future work could explore operationalising the top performing machine learning models identified in this study within a production grade temperature derivative pricing pipeline. A more comprehensive analysis of tail behaviour in the seasonal context relevant for derivative pricing would also be valuable. Emerging forecasting architectures including recent advances in Generative AI also present opportunities for further performance gains and merit investigation in this domain.

Finally, a further direction for future research will be to incorporate the spatial dimension of temperature modelling. In this study, models were developed on a station by station basis, reflecting the structure of temperature derivative markets, where contracts are written against a single reference station rather than a spatial aggregate. Recent advances in spatio-temporal machine learning architectures (Yu et al. (2024), Feng et al. (2024)) that learn jointly from multiple locations, offer a promising avenue for future work. Such approaches could reduce computational overhead by enabling shared learning across locations, while also providing a richer representation of spatial temperature correlations.

Appendix: Mathematical formulations of the machine learning models

This appendix summarises the core mathematical formulations of the forecasting models used in this study. Let $\{(x_t, y_t)\}_{t=1}^{T}$ denote the input output pairs, where $x_t \in \mathbb{R}^d$ and $y_t \in \mathbb{R}$.

Linear regression

Linear regression assumes a linear relationship

$$\hat{y}_t = x_t^\top \beta,$$

where β is estimated by minimising the squared error

$$\min_{\beta} \sum_{t=1}^{T} (y_t - x_t^\top \beta)^2.$$

k-nearest neighbours regression (kNN)

The prediction is the average of the k nearest neighbours of x_t:

$$\hat{y}_t = \frac{1}{k} \sum_{i \in \mathcal{N}_k(x_t)} y_i,$$

where $\mathcal{N}_k(x_t)$ denotes the index set of the k closest points under a chosen distance metric.

Decision tree regression

A decision tree partitions the input space into M disjoint regions $\{R_m\}_{m=1}^{M}$. The prediction is

$$\hat{y}_t = \sum_{m=1}^{M} c_m \, 1(x_t \in R_m),$$

where each c_m is the mean of y_t in region R_m. Splits are chosen to minimise the impurity

$$\sum_{m=1}^{M} \sum_{t: x_t \in R_m} (y_t - c_m)^2.$$

Random forest regression

A random forest is an ensemble of B decision trees $\{f_b(x)\}_{b=1}^{B}$ trained on bootstrap samples with random feature subsets. The prediction is

$$\hat{y}_t = \frac{1}{B} \sum_{b=1}^{B} f_b(x_t).$$

Support vector regression (SVR)

SVR solves

$$\min_{w,b,\xi_t,\xi_t^*} \frac{1}{2}\|w\|^2 + C\sum_{t=1}^{T}(\xi_t + \xi_t^*)$$

subject to

$$\begin{aligned} y_t - (w^\top \phi(x_t) + b) &\le \varepsilon + \xi_t, \\ (w^\top \phi(x_t) + b) - y_t &\le \varepsilon + \xi_t^*, \\ \xi_t, \xi_t^* &\ge 0, \end{aligned}$$

where $\phi(\cdot)$ is a feature map and ε is the insensitive margin.

Linear support vector regression

This is the special case of SVR where $\phi(x) = x$, giving a linear predictor

$$\hat{y}_t = w^\top x_t + b.$$

Gradient boosting regression

Gradient boosting constructs an additive model

$$f_M(x) = \sum_{m=1}^{M} \nu\, h_m(x),$$

where each h_m is a regression tree fitted to the negative gradient of the loss at iteration m, and ν is the learning rate.

Histogram gradient boosting regression

Histogram gradient boosting approximates the gradient boosting procedure by binning each feature into discrete histogram buckets. Trees are fitted on these binned features, reducing the optimisation to a piecewise-constant approximation of the gradient.

XGBoost

XGBoost minimises a regularised objective

$$\mathcal{L} = \sum_{t=1}^{T} \ell(y_t, f(x_t)) + \sum_{m=1}^{M} \Omega(h_m),$$

where $f(x) = \sum_{m=1}^{M} h_m(x)$ and each tree h_m has regularisation

$$\Omega(h_m) = \gamma J + \frac{\lambda}{2} \sum_{j=1}^{J} w_j^2,$$

with J leaves and leaf weights w_j. A second-order Taylor expansion of the loss guides tree construction.

Multilayer perceptron (MLP)

An MLP with L layers computes

$$h^{(0)} = x_t, \qquad h^{(\ell)} = \sigma(W^{(\ell)} h^{(\ell-1)} + b^{(\ell)}), \quad \ell = 1, \ldots, L,$$

with prediction $\hat{y}_t = h^{(L)}$. Parameters are learned by minimising a loss via gradient descent.

Long short-term memory (LSTM)

For input x_t and hidden state (h_{t-1}, c_{t-1}), an LSTM updates

$$\begin{aligned}
i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i), \\
f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f), \\
o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o), \\
\tilde{c}_t &= \tanh(W_c x_t + U_c h_{t-1} + b_c), \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \\
h_t &= o_t \odot \tanh(c_t),
\end{aligned}$$

with $\hat{y}_t$ obtained from a final linear layer.

Temporal convolutional network (TCN)

A TCN applies dilated causal convolutions:

$$h_t^{(l)} = \sigma\left(\sum_{k=0}^{K-1} W_k^{(l)} \, h_{t-d_l k}^{(l-1)} + b^{(l)} \right),$$

where K is the kernel size and d_l the dilation factor at layer l. The final layer maps $h_t^{(L)}$ to $\hat{y}_t$.

Author Contributions NNB, MK and PK made substantial contributions to the conception and design of the work. NNB wrote the main manuscript with ML and PK providing a critical review.

Funding The authors did not receive support from any organization for the submitted work.

Data availibility The data that was used in this study is publicly available and has been fully referenced in the manuscript but the downloaded and processed copy used can be made available on request.

Materials availability Not applicable to this work.

Code availability The code that was used to produce this work has been made publicly available on Github here: https://github.com/ganyohbi/TemperatureForecastIngPythonFrameWork.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Ethical approval and consent to participate Not applicable to this work.

Consent for publication The authors consent to the submission of this manuscript to be considered for publication by Artificial Intelligence Review.

References

Alaton P, Djehiche B, Stillberger D (2002) On modelling and pricing weather derivatives. Appl Math Finance 9(1):1–20

Alexandridis A, Zapranis A (2013a) Wind derivatives: modeling and pricing. Comput Econ 41:299–326. https://doi.org/10.1007/s10614-012-9350-y

Alexandridis AK, Zapranis AD (2013b) Wavelet neural networks: a practical guide. Neural Netw 42:1–27. https://doi.org/10.1016/j.neunet.2013.01.008

Alexandridis AK, Zapranis AD (2013c) Weather derivatives: modeling and pricing weather-related risk. Springer

Alexandridis A, Kampouridis M, Cramer S (2017) A comparison of wavelet networks and genetic programming in the context of temperature derivatives. Int J Forecast 33:21–47. https://doi.org/10.1016/j.ijforecast.2016.07.002

Alexandridis AK, Gzyl H, ter Horst E, Molina G (2021) Extracting pricing densities for weather derivatives using the maximum entropy method. J Oper Res Soc 72:2412–2428. https://doi.org/10.1080/01605682.2020.1796532

Alfonsi A, Vadillo N (2024) A stochastic volatility model for the valuation of temperature derivatives. IMA J Manag Math 35(4):737–785

Ben Bouallègue Z, Clare MC, Magnusson L, Gascon E, Maier-Gerber M, Janoušek M, Rodwell M, Pinault F, Dramsch JS, Lang ST, Raoult B, Rabier F, Chevallier M, Sandu I, Dueben P, Chantry M, Pappenberger F (2024) The rise of data-driven weather forecasting: a first statistical assessment of machine learning-based weather forecasts in an operational-like context. Bull Am Meteor Soc 105(6):E864–E883

Benth F (2003) On arbitrage-free pricing of weather derivatives based on fractional Brownian motion. Appl Math Finance 10:303–324. https://doi.org/10.1080/1350486032000174628

Benth F, Šaltyte Benth J (2005) Stochastic modelling of temperature variations with a view towards weather derivatives. Appl Math Finance 12:53–85. https://doi.org/10.1080/1350486042000271638

Berhane T, Shibabaw N, Awgichew G, Kebede T (2020) Option pricing of weather derivatives based on a stochastic daily rainfall model with analogue year component. Heliyon 6:e03212. https://doi.org/10.1016/J.HELIYON.2020.E03212

Berhane T, Shibabaw A, Awgichew G, Walelgn A (2021) Pricing of weather derivatives based on temperature by obtaining market risk factor from historical data. Model Earth Syst Environ 7:871–884. https://doi.org/10.1007/S40808-020-00925-4
Box GEP, Jenkins GM (1976) Time series analysis: forecasting and control, 2nd edn. Holden-Day, San Francisco
Brody DC, Syroka J, Zervos M (2002) Dynamical pricing of weather derivatives. Quant Finance 2(3):189. https://doi.org/10.1088/1469-7688/2/3/302
Cabrales S, Bautista R, Madiedo I, Galindo M (2022) A methodology for temperature option pricing in the equatorial regions. Eng Econom 67:96–111. https://doi.org/10.1080/0013791X.2021.2000086
Campbell SD, Diebold FX (2005) Weather forecasting for weather derivatives. J Am Stat Assoc 100:6–16. https://doi.org/10.1198/016214504000001051
Carmona R (1999) Calibrating degree day options. In: 3rd Seminar on stochastic analysis, random field and applications. Ecole Polytechnique de Lausanne, Switzerland
CFTC (2020) Managing climate risk in the U.S. financial system. Report of the Climate Related Market Risk Subcommittee
Cheng T, Poreddy SR, Chen K (2025) Tail risk in weather derivatives. Commodities 4(2):11
De Saa E, Ranathunga L (2020) Comparison between Arima and deep learning models for temperature forecasting. arXiv preprint arXiv:2011.04452
Dehvari M, Farzaneh S, Forootan E (2025) Forecasting rainfall events based on zenith wet delay time series utilizing extreme gradient boosting (xgboost). Adv Space Res 75:2584–2598
Dimitrova DS, Kaishev VK, Tan S (2020) Computing the Kolmogorov-Smirnov distribution when the underlying cdf is purely discrete, mixed, or continuous. J Stat Softw 95:1–42
Dimri T, Ahmad S, Sharif M (2020) Time series analysis of climate variables using seasonal Arima approach. J Earth Syst Sci 129(1):149
Dischel B (1999) Shaping history for weather risk management. Energy Power Risk 12(8):13–15
Dueben PD, Bauer P (2018) Challenges and design choices for global weather and climate models based on machine learning. Geosci Model Dev 11(10):3999–4009. https://doi.org/10.5194/gmd-11-3999-2018
Eggen MD, Dahl KR, Näsholm SP, Mæland S (2022) Stochastic modeling of stratospheric temperature. Math Geosci 54(4):651–678
Elias RS, Wahab MI, Fang LL (2014) A comparison of regime-switching temperature modeling approaches for applications in weather derivatives. Eur J Oper Res 232:549–560. https://doi.org/10.1016/J.EJOR.2013.07.015
Elshewey AM, Shams MY, Elhady AM, Shohieb SM, Abdelhamid AA, Ibrahim A, Tarek Z (2022) A novel wd-Sarimax model for temperature forecasting using daily Delhi climate dataset. Sustainability 15(1):757
Feng R, Chen M, Song Y (2024) Learning traffic as videos: short-term traffic flow prediction using mixed-pointwise convolution and channel attention mechanism. Expert Syst Appl 240:122468
Franses PH, Neele J, van Dijk D (2001) Modeling asymmetric volatility in weekly Dutch temperature data. Environ Modell Softw 16(2):131–137
Friedman M (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J Am Stat Assoc 32(200):675–701
Frnda J, Durica M, Nedoma J, Zabka S, Martinek R, Kostelansky M (2019) A weather forecast model accuracy analysis and ecmwf enhancement proposal by neural network. Sensors 19(23):5144
Geyser JM (2004) Weather derivatives: concept and application for their use in South Africa. Agrekon 43:444–464. https://doi.org/10.1080/03031853.2004.9523660
Grover A, Kapoor A, Horvitz E (2015) A deep hybrid model for weather forecasting. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pp 379–386
Gyamfi B (2024) A stochastic weather model: a case of bono region of ghana. arXiv preprint arXiv:2409.06731
Gyamfi B, Boiquaye PA, Gyamerah SA (2025) Stochastic modelling of temperature for pricing weather derivatives. Environ Res Commun 7(5):055016
Ham YG, Kim JH, Luo JJ (2019) Deep learning for multi-year enso forecasts. Nature 573(7775):568–572
Haque E, Tabassum S, Hossain E (2021) A comparative analysis of deep neural networks for hourly temperature forecasting. IEEE Access 9:160646–160660
He R, Zhang L, Chew AWZ (2022) Modeling and predicting rainfall time series using seasonal-trend decomposition and machine learning. Knowl-Based Syst 251:109125
Hossain M, Rekabdar B, Louis SJ, Dascalu S (2015) Forecasting the weather of Nevada: a deep learning approach. In: 2015 International joint conference on neural networks (IJCNN). IEEE, pp 1–6
Hyndman RJ, Athanasopoulos G (2018) Forecasting: principles and practice (2 ed.). OTexts
Jewson S (2004) Introduction to weather derivative pricing. J Altern Invest 7:57–64
Jewson S, Brix A (2005) Weather derivative valuation: the meteorological, statistical, financial and mathematical foundations. Cambridge University Press

Johnston J, DiNardo JE (2007) Econometric methods, 4th edn. McGraw-Hill, New York
Karevan Z, Suykens, JA (2018) Spatio-temporal stacked lstm for temperature prediction in weather forecasting. arXiv preprint arXiv:1811.06341
Kreuzer D, Munz M, Schlüter S (2020) Short-term temperature forecasts using a convolutional neural network–an application to different weather stations in germany. Mach Learn Appl 2:100007
Li P, Wang F, Wang L, Liu K, Shen J, Zhong W (2025) Pricing weather derivatives and managing weather risks under regime switching. IMA J Manag Math: dpaf003
Liu JN, Hu Y, You JJ, Chan PW (2014) Deep neural network based feature representation for weather forecasting. In: Proceedings on the international conference on artificial intelligence (ICAI), pp 1
Löning M, Bagnall A, Ganesh S, Kazakov V, Lines J, Király FJ (2019) Sktime: A Unified interface for machine learning with time series. In: Workshop on systems for ML at NeurIPS 2019
Moreno M (2000) Riding the temp. Futures and Options World 11
Müller A, Grandi M (2000) Weather derivatives: a risk management tool for weather-sensitive industries. Geneva Papers Risk Insur Issues Pract 25(2):273–287
Murat M, Malinowska I, Gos M, Krzyszczak J (2018) Forecasting daily meteorological time series using Arima and regression models. Int Agrophys 32(2)
Nemenyi PB (1963) Distribution-free multiple comparisons. Princeton University
NOAA (2024) National oceanography and atmospheric administration. https://www.ncei.noaa.gov/access/past-weather Accessed: 2024-04-04
Oetomo T, Stevenson M (2005) Hot or cold? A comparison of different approaches to the pricing of weather derivatives. J Emerg Market Finance 4:101–133. https://doi.org/10.1177/097265270500400201
Oettli P, Nonaka M, Richter I, Koshiba H, Tokiya Y, Hoshino I, Behera SK (2022) Combining dynamical and statistical modeling to improve the prediction of surface air temperatures 2 months in advance: A hybrid approach. Front Climate 4:862707
Ornstein LS (1930) On the theory of the Brownian motion. Phys Rev 36:823–841
Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
Price I, Sanchez-Gonzalez A, Alet F, Andersson TR, El-Kadi A, Masters D, Ewalds T, Stott J, Mohamed S, Battaglia P et al (2025) Probabilistic weather forecasting with machine learning. Nature 637(8044):84–90
Ren X, Li X, Ren K, Song J, Xu Z, Deng K, Wang X (2021) Deep learning-based weather prediction: a survey. Big Data Res 23:100178
Šaltyte Benth J, Benth F (2012) A critical view on temperature modelling for application in weather derivatives markets. Energy Econom 34:592–602. https://doi.org/10.1016/j.eneco.2011.09.012
Schiller F, Seidler G, Wimmer M (2012) Temperature models for pricing weather derivatives. Quant Finance 12:489–500. https://doi.org/10.1080/14697681003777097
Sharma S, Singh UP (2024) A comparative study of adaptive time series models for weather forecasting in Indian scenarios. In: 2024 IEEE 21st India council international conference (INDICON). IEEE, pp 1–6
Singh S, Kaushik M, Gupta A, Malviya AK (2019) Weather forecasting using machine learning techniques. In: Proceedings of 2nd international conference on advanced computing and software engineering (ICACSE)
Supto STJ (2025) Next-generation climate modeling: Ai-enhanced, machine-learning, and hybrid approaches beyond conventional gcms. Environ Earth Sci Proc 34(1):15
Tallarico MH, Olivares P (2024) Neural and time-series approaches for pricing weather derivatives Performance and regime adaptation using satellite data arXiv preprint arXiv:2411.12013
Tol RS (1996) Autoregressive conditional heteroscedasticity in daily temperature measurements. Environmetrics 7(1):67–75
Tong KZ, Liu A, Liu A (2020) Modeling temperature and pricing weather derivatives based on subordinate Ornstein-Uhlenbeck processes. Green Finance 2(1):1–19
Türkvatan A, Hayfavi A, Omay T (2020) A regime switching model for temperature modeling and applications to weather derivatives pricing. Math Finance Econ 14:1–42. https://doi.org/10.1007/s11579-019-00242-0
Van Rossum G, Drake FL (2009) Python 3 reference manual. CreateSpace, Scotts Valley
Wenjie J et al (2024) Meteorological composite index prediction based on WD-SSA-LSTM modeling and pricing study of weather derivatives. Acad J Busin Manag 6(7):175–180
Wilcoxon F (1992) Individual comparisons by ranking methods, breakthroughs in statistics: methodology and distribution, 196–202. Springer
Yan J, Mu L, Wang L, Ranjan R, Zomaya AY (2020) Temporal convolutional networks for the advance prediction of enso. Sci Rep 10(1):8055
Yang CC, Brockett PL, Wen MM (2009) Basis risk and hedging efficiency of weather derivatives. J Risk Finance 10:517–536. https://doi.org/10.1108/15265940911001411

Yu M, Huang Q, Li Z (2024) Deep learning for spatiotemporal forecasting in earth system science: a review. Int J Digit Earth 17(1):2391952
Yu Y, Xie Y, Tao Z, Ju H, Wang M (2023) Global temperature prediction models based on Arima and lstm. In: Chinese conference on image and graphics technologies. Springer, pp 301–314
Zapranis A, Alexandridis A (2008) Modelling the temperature time-dependent speed of mean reversion in the context of weather derivatives pricing. Appl Math Finance 15:355–386. https://doi.org/10.1080/13504860802006065
Zhang H, Liu Y, Zhang C, Li N (2025) Machine learning methods for weather forecasting: a survey. Atmosphere 16(1):82
Zhu L, Li Q (2023) Global warming temperature prediction based on Arima. In: Proceedings of the 7th international conference on innovation in artificial intelligence, pp 121–128
Zong L, Ender M (2016) Spatially-aggregated temperature derivatives: agricultural risk management in China. Int J Financial Stud. https://doi.org/10.3390/ijfs4030017

Authors and Affiliations

Nee Barnor[1] · Michael Kampouridis[1] · Panagiotis Kanellopoulos[1]

✉ Nee Barnor
nb22949@essex.ac.uk

Michael Kampouridis
mkampo@essex.ac.uk

Panagiotis Kanellopoulos
panagiotis.kanellopoulos@essex.ac.uk

[1] School of Computer Science and Electronic Engineering, University of Essex, Wivenhoe Park, Colchester, Essex CO4 3SQ, UK