# Approximating the universal thermal climate index using sparse regression with orthogonal polynomials

Sabin Roman[1], Gregor Skok[2], Ljupčo Todorovski[2,1], and Sašo Džeroski[1]

[1] Department of Knowledge Technologies, Jožef Stefan Institute
sabin.roman@ijs.si
[2] Faculty of Mathematics and Physics, University of Ljubljana

**Abstract.** The Universal Thermal Climate Index (UTCI) is a measure of thermal comfort that quantifies how humans experience environmental conditions. Due to its robustness and versatility as a bioclimatic indicator, it has been extensively employed across a wide range of studies in bioclimatology and is increasingly used as an operational measure of outdoor thermal comfort. At the same time, calculating the UTCI value from the relevant environmental parameters is nominally not straightforward, which is why using a 6th-degree polynomial approximation has become the standard way to calculate UTCI values. At the same time, although it is computationally efficient, the error of this polynomial approximation can be substantial. The goal of this study was to develop an improved version of the polynomial approximation – one that retains comparable computational efficiency but is more robust in terms of numerical stability and substantially more accurate, particularly in reducing the frequency of larger errors. This goal was successfully achieved using sparse orthogonal regression, namely sparse regression with an orthogonal polynomial basis, which not only substantially reduces the average errors (i.e., the mean error, the mean absolute error, and the root mean square error) but also drastically reduces the frequency of large errors. By leveraging Legendre polynomial bases, approximation models could be constructed that efficiently populate a Pareto front of accuracy versus complexity and exhibit stable, hierarchical coefficient structures across varying model capacities. Training the new approximation models over only 20% of the data, with the testing performed over the remaining 80%, highlights successful generalization, with the results also being robust under bootstrapping. The decomposition effectively approximates the UTCI as a Fourier-like expansion in an orthogonal basis, yielding results near the theoretical optimum in the $L_2$ (least squares) sense.

**Keywords:** Universal Thermal Climate Index · Sparse regression · Orthogonal polynomials

## 1 Introduction

The Universal Thermal Climate Index (UTCI) is a measure of thermal comfort that quantifies how humans experience environmental conditions. It is derived

from an advanced thermo-physiological model (Pappenberger et al., 2015) and expressed in units of temperature. The index accounts for multiple factors, including air temperature, humidity, wind speed, radiation, and clothing insulation (Bröde et al., 2012). A notable advantage of the UTCI compared to many other bioclimatic indices is its ability to represent thermal conditions in terms that are applicable to human strain under a wide range of climatic conditions (e.g., for both hot and cold conditions, Blazejczyk et al. (2012)). Based on the UTCI value, the environmental conditions can be classified into one of the ten thermal stress categories (Bröde et al., 2012), ranging from Extreme heat stress ($UTCI > 43°\mathrm{C}$) to Extreme cold stress ($UTCI < -40°\mathrm{C}$).

Owing to its robustness and versatility as a bioclimatic indicator, the UTCI has been extensively employed across a wide range of studies in bioclimatology and related scientific disciplines. Its applications encompass diverse research areas, including the assessment of regional and local bioclimate characteristics, the study of urban bioclimate, recreation, tourism, and sports, epidemiological and health-related research, as well as the assessment and forecasting of bioclimatic changes (Błażejczyk and Kuchcik, 2021). The UTCI has also seen growing adoption across numerous countries as a standardized measure of outdoor thermal comfort and is increasingly integrated into routine operational meteorological forecasts. For example, within Europe, UTCI is used operationally in the Czech Republic, Italy, Poland, Portugal, and Slovenia (Di Napoli et al., 2021a; Kuzmanović et al., 2024).

At the same time, calculating the UTCI value from the relevant environmental parameters is nominally not straightforward. Namely, the UTCI is based on the Fiala multi-node model of human thermoregulation (Fiala et al., 2012). However, running the complete Fiala model is computationally expensive and requires expert knowledge to operate the complex simulation software (Bröde et al., 2012). This is the reason the authors of Bröde et al. (2012) provided two simplified approximate procedures for calculating the UTCI values that could be used in operational settings. The first approximation is based on a 4-dimensional look-up table of 104 643 accurate pre-calculated UTCI values that cover a wide range of relevant combinations of the meteorological parameters. Using this look-up table, interpolation from nearby data points can be used to determine approximate UTCI values for intermediate values of meteorological parameters. The second approximation is based on a 6th-degree regression polynomial with 210 coefficients.

Each approximation has its benefits and weaknesses. The look-up table approach is more accurate, but storing the tabulated values and searching for neighboring datapoints poses challenges to the implementation of this algorithm, while also resulting in a longer execution time compared to the other approach (Bröde, 2021a). In contrast, the polynomial approximation is less accurate, but computationally faster and substantially easier to implement in various programming languages and computational environments, as it relies on only the most common, primitive mathematical operators and does not require storing the tabulated values. At the same time, the motivation for improving the poly-

nomial approximation is not simply a matter of storage, since the size of the look-up table is modest in modern computational settings. Rather, an improved polynomial approximation remains attractive for several practical reasons:

(i) It is fully self-contained and does not depend on external tabulated data, which facilitates reproducibility and makes redistribution and integration into open-source software and operational tools more straightforward;
(ii) It is computationally more efficient than look-up-table-based interpolation, which has been reported to be slower by roughly three orders of magnitude (Bröde et al., 2012), an important consideration in large-scale applications such as numerical weather prediction and climate reanalysis;
(iii) It is simpler to implement and port across programming languages and computational environments, including constrained, embedded, or legacy systems, because it requires only basic arithmetic operations and avoids the additional logic needed for multidimensional interpolation, data handling, and neighborhood search;
(iv) It provides a direct, continuous, and analytically defined mapping over the domain of validity, whereas the look-up table still requires interpolation, and in some cases extrapolation, for environmental states not explicitly represented in the tabulated values;
(v) Its predictive behavior on unseen data can be assessed directly through a train–test evaluation framework; in the present case, training on 20% of the dataset and testing on the remaining 80% still yields very good predictive performance, indicating strong generalization.

For these reasons, the polynomial approximation is best viewed not as a universal replacement for the look-up-table approach, but as a complementary alternative that is particularly useful in applications where speed, portability, reproducibility, and ease of deployment are important.

Due to its simplicity and computational efficiency, the polynomial approximation has become the standard way of calculating the UTCI values. It has been incorporated into various biclimatic software packages and libraries (e.g., the Bioklima software, Błażejczyk (2025), the Thermofeel Python library, Brimicombe et al. (2022), and the pyThermalComfort Python library Tartarini and Schiavon (2020)), as well as numerical weather prediction and reanalysis systems (e.g., the ALADIN model, Termonia et al. (2018), and the ERA5 reanalysis, Di Napoli et al. (2021b)). At the same time, the error of the polynomial approximation can be substantial. For example, when evaluated on the aforementioned look-up table of accurate UTCI values, the root-mean-square-error is about 1.1°C while the frequency of absolute errors larger than 2°C is about 8%, and the frequency of errors larger than 3°C is about 2%. This is problematic since an error of a few degrees Celsius can increase the likelihood of misclassification of the thermal stress category, some of which span only a 6°C interval.

The goal of this study is to develop an improved version of the polynomial approximation – one that has comparable computational complexity to the existing approximation but is more robust in terms of numerical stability and substantially more accurate, particularly in reducing the frequency of larger errors. To

| Variable name | Description | Valid Range | Normalized range |
|---|---|---|---|
| $Ta$ | Air temperature | $-50$ to $+50$ °C | $[-1, 1]$ |
| $va$ | Wind speed at 10 m | 0.5 to 30.3 m/s | $[-1, 1]$ |
| $Tr - Ta$ | Mean Radiant–air temperature difference | $-30$ to $+70$ °C | $[-1, 1]$ |
| $rH$ | Relative humidity | 5 to 100 % | $[-1, 1]$ |
| $pa$ | Water vapour pressure | 0 to 5 kPa | Not used |

Table 1: Description of variables used in this study, following Bröde et al. (2012). The normalized ranges map each variable to $[-1, 1]$, with respect to the interval of validity, suitable for use with Legendre polynomial bases. Although water vapor pressure ($pa$) is not used directly as an input for the new approximation, it can be computed from air temperature ($Ta$) and relative humidity ($rH$), and its effect is therefore accounted for through the inclusion of $rH$.

achieve this goal, symbolic and sparse regression techniques are used as tools for interpretable and efficient function approximation. We fit the UTCI offset using sparse regression on an orthogonal Legendre polynomial basis. To emphasize this key feature and distinguish it from standard sparse regression on monomials, we refer to this approach as sparse orthogonal regression.

We also note that the aim was not to derive an approximation that was as accurate as possible. For example, a sufficiently complex neural-network-based model would likely provide more accurate estimates of the UTCI values. However, such a model would also require the use of machine-learning libraries, as well as suitable Graphics Processing Units, to function efficiently. This means that its implementation in various programming languages and computational environments would be substantially more difficult. On the other hand, replacing an existing polynomial approximation with a new one is fairly straightforward, meaning that implementing the new approximation into existing biclimatic software packages/libraries and numerical weather prediction systems would be relatively easy.

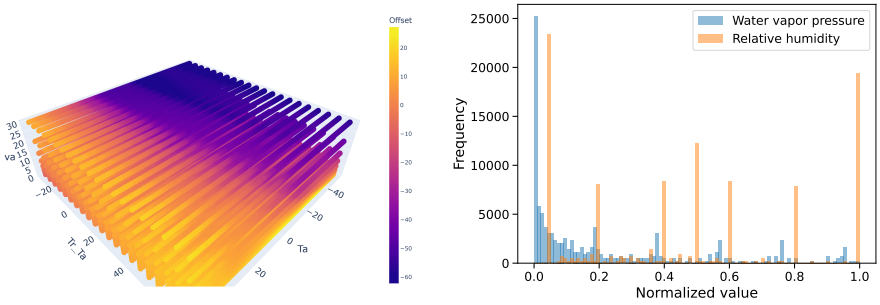
## 2 Methods

Formally, the UTCI is defined as (Bröde et al., 2012)

$$UTCI = Ta + \text{Offset}(Ta, va, Tr, rH \text{ or } pa), \tag{1}$$

where $Ta$ is the air temperature and the Offset is the physiologically equivalent temperature difference, representing how other environmental factors modify the effect of the thermal stress on the human body. The Offset function represents the deviation of the UTCI from the actual air temperature and depends on $Ta$, wind speed at 10 m ($va$), mean radiant temperature ($Tr$), which accounts for the effect of all incoming radiation, and humidity, which can be represented by either relative humidity ($rH$) or water vapour pressure ($pa$).

The dataset provided by Bröde et al. (2012) contains accurate values of the Offset function covering a wide range of environmental states. The variables




(a) (b)

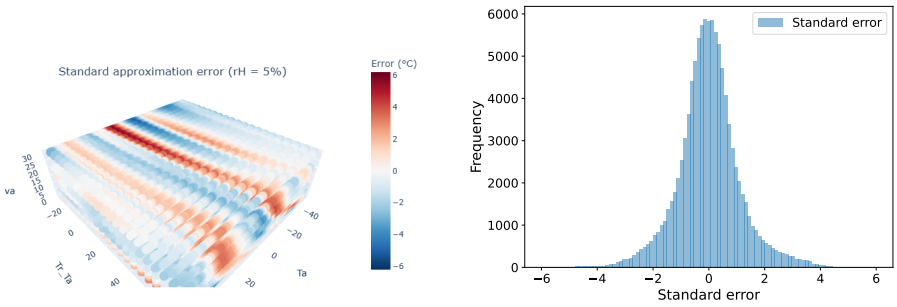
Fig. 1: (a) 3D plot of UTCI Offset Bröde et al. (2012) at 5% relative humidity, showing how wind speed ($va$), air temperature ($Ta$), and mean radiant temperature difference ($Tr - Ta$) combine to influence thermal stress. Color indicates the UTCI Offset magnitude across these environmental dimensions. (b) The different distributions of the water vapor pressure and relative humidity in the computed Offset dataset Bröde et al. (2012). The water vapor pressure is strongly peaked at zero, while the relative humidity is uniform across its range.

and their ranges are included in Table 1. The intervals of the environmental variables also represent the domain where the sixth-degree polynomial regression approximation is considered valid (Bröde et al., 2012). Using the approximation for conditions outside of these intervals can lead to large errors and unrealistic values of the Offset function and should be avoided (Bröde, 2021a).

In Fig. 1(a) we see how the UTCI Offset varies along the different environmental variables. Instead of the humidity ($rH$), the water vapor pressure ($pa$) can be used which is a nonlinear function of $rH$ and the air temperature ($Ta$). However, the variables have different distribution, see Fig. 1(b), which impacts the extent that approximations of UTCI can generalize, discussed below.

Equation discovery aims to learn interpretable mathematical expressions, either differential or algebraic equations, from measurements of the variables of a given observed system (Todorovski and Džeroski, 1997). Positioned at the intersection of symbolic machine learning and system identification, it is becoming increasingly relevant in environmental and climate science, where data-driven yet transparent models are essential (Steinmann et al., 2025; Roman, 2025b). Traditional modeling approaches rely on expert-derived formulations (Roman, 2021, 2023; Roman and Bertolotti, 2022, 2023; Roman and Palmer, 2019), but the growing complexity and volume of climate data call for automated alternatives. Symbolic regression, which iteratively combines mathematical operators and variables to fit data, forms the core of equation discovery (Bridewell et al., 2005; Todorovski and Džeroski, 2006; Džeroski et al., 2007). Most methods employ evolutionary or other (e.g., enumerative) search strategies to explore the space of candidate equations (Tanevski et al., 2016a, 2020; Mežnar et al., 2023).

Recent advances integrate probabilistic grammars to incorporate prior knowledge and constrain the search to physically meaningful expressions (Brence et al.,




(a) (b)

Fig. 2: The error of the standard polynomial UTCI approximation (Bröde et al., 2012) for relative humidity of 5%. (a) The difference between the standard UTCI approximation and the accurate values of the Offset function. (b) Histogram of the differences showing a normal distribution centered at zero.

2020, 2023; Omejc et al., 2024). This structured approach improves both model interpretability and search efficiency, especially in domains governed by established scientific principles. Equation discovery has been applied to various environmental systems (Atanasova et al., 2006a, 2011, 2008, 2006b), including ecosystem dynamics (Jeraj et al., 2006; Čerepnalkoski et al., 2012; Simidjievski et al., 2015, 2016; Tanevski et al., 2016b). In these settings, it can match or even surpass expert-built models while simultaneously revealing new relationships (Todorovski and Džeroski, 2001; Todorovski et al., 1998). Its ability to generate compact, interpretable, and physically plausable models makes it especially suitable for climate applications, where model transparency and adherence to physical principles are vital.

As already mentioned, the errors of the sixth-degree regression polynomial from Bröde et al. (2012) can be substantial. Fig. 2(a) shows the approximation error at 5% relative humidity, while 2(b) displays a histogram of the errors, revealing a normal distribution centered at zero, indicating minimal bias. We aim to improve upon this standard approximation using equation discovery and sparse regression methods by utilizing the accurate Offset dataset provided by Bröde et al. (2012).

Sparse machine learning models aim to construct parsimonious predictive functions by enforcing zero-valued coefficients in high-dimensional parameter spaces, thereby performing implicit feature selection (Brunton et al., 2016). This sparsity promotes interpretability, reduces overfitting, and improves computational tractability, especially when the number of candidate predictors is large or when strong correlations exist among inputs. Sparse regression (Brunton et al., 2016), a key instantiation of this paradigm, extends linear regression with an $L_1$-norm regularization term—most notably in the Lasso (Least Absolute Shrinkage and Selection Operator, Reid et al., 2016) —to penalize unnecessary parameters and induce a compact representation.

In this work, we employ sparse regression to identify compact, interpretable models of the UTCI, emphasizing its suitability for high-dimensional input spaces with redundant or weakly relevant features. While sparse modeling is well-established in statistical learning, its application to orthogonal polynomial bases-particularly in the context of bioclimatic indices—remains unexplored. By leveraging the structure of orthogonal polynomials, we obtain improved numerical stability and additive expansions that facilitate coefficient interpretability. To our knowledge, this is the first application of sparse regression using orthogonal bases to approximate the UTCI, addressing both predictive accuracy and model parsimony. Our results show that this approach surpasses the standard sixth-degree polynomial approximation in both accuracy and efficiency.

## 3 Results and discussion

Table 2 presents a detailed comparison of model performance across a range of polynomial degrees for both standard (non-sparse) linear regression and sparse regression techniques, evaluated in the context of approximating the UTCI. The standard approximation (Bröde et al., 2012) is a sixth-degree regression polynomial model with four variables, consisting of 210 terms and achieving a root mean squared loss of 1.12°C. This serves as the benchmark to be matched or improved upon. It is important to note that the standard approximation does not directly employ the relative humidity ($rH$), but the water vapor pressure ($pa$), which can be derived from the relative humidity ($rH$) and air temperature ($Ta$). As we noted above, in the dataset, the relative humidity is well represented across its entire range, see Fig. 1(b), while the water vapor pressure is strongly peaked close to zero. Optimization employing the water vapor pressure ($pa$) as an independent variable (instead of $rH$) is thus poorly conditioned and leads to instability in the regression coefficients, both in simple and sparse regression. While using the pa (instead of $rH$) can achieve better accuracy (lower loss), it comes at the price of losing parameter consistency across optimizations with different polynomial degrees. For this reason, we report our results employing the relative humidity ($rH$) instead of the water vapor pressure ($pa$), see Table 1.

The regression methods are applied to polynomial basis expansions of increasing degree, evaluated on the basis of root mean squared test loss and number of active parameters. Unlike many studies in the literature where models are trained on the majority of the data and evaluated on a relatively small test set, our approach inverts this paradigm: training is conducted on only 20% of the available data, while performance is assessed on the remaining 80%. Despite this stringent evaluation setting, the models achieve comparable performance on both training and test sets, underscoring their strong generalization capabilities. This performance stability is further validated through bootstrapping, which reveals minimal variance in both loss metrics and selected features across resampled datasets. The reported performance metrics—such as train/test loss and number of parameters—remain stable when the model training and evaluation process is repeated on multiple random re-samplings (bootstrapped subsets) of

| Method | | | | | | | |
|---|---|---|---|---|---|---|---|
| Standard | 1.12 (210) | | | | | | |
| | Polynomial degree | | | | | | |
| | 4th | 6th | 8th | 10th | 12th | 14th | 16th |
| Linear regression | 2.1 (70) | 1.3 (210) | 0.92 (495) | Train: 0.67 Test: 0.71 (1001) | 0.54 0.62 (1820) | 0.44 0.66 (3060) | 0.36 1.74 (4845) |
| Sparse orthogonal regression | 2.1 (65) | 1.38 (124) | 1.03 (176) | **0.88** (209) | 0.69 (355) | 0.63 (400) | 0.6 (424) |

Table 2: Root mean squared train loss [°C], test loss [°C] and the number of parameters (shown in parenthesis) in approximating the UTCI Offset. The baseline reference, labeled as "Standard," corresponds to the sixth-degree regression polynomial model with four variables (Bröde et al., 2012). Unless otherwise stated the test loss equals the train loss. Where two loss values are reported (train loss on the top and test loss below), they indicate a notable train-test discrepancy, typically suggesting overfitting. Training is done with 20% of the data and testing is performed with 80%. Results are robust under bootstrapping.

the data. This suggests that the results are not sensitive to specific data splits and that the models generalize well across different subsets of the dataset, indicating reliability and consistency in the reported findings. These findings demonstrate the robustness and reliability of the proposed framework.

To make the fitted model class explicit, let $\tilde{T}_a$, $\tilde{v}_a$, $\widetilde{\Delta T_r}$, and $\widetilde{rH}$ denote the normalized versions of $T_a$, $v_a$, $T_r - T_a$, and $rH$, respectively, each mapped to the interval $[-1, 1]$ according to the ranges in Table 1. In this formulation, relative humidity is retained as an input variable in order to account for the effect of water vapor. The approximation of the UTCI offset can then be written in the general form

$$\widehat{\text{Offset}}(T_a, v_a, T_r - T_a, rH) = \sum_{\boldsymbol{\alpha} \in \mathcal{A}_p} c_{\boldsymbol{\alpha}} \prod_{j=1}^{4} P_{\alpha_j}(x_j), \tag{2}$$

where $P_n(\cdot)$ denotes the Legendre polynomial of degree $n$, $(x_1, x_2, x_3, x_4) = (\tilde{T}_a, \tilde{v}_a, \widetilde{\Delta T_r}, \widetilde{rH})$, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ is a multi-index, and

$$\mathcal{A}_p = \left\{ \boldsymbol{\alpha} \in \mathbb{N}_0^4 : \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 \leq p \right\}$$

is the set of all basis terms up to total polynomial degree $p$. Thus, the model is a linear combination of products of Legendre polynomials in the four normalized environmental variables. For a given maximum degree $p$, the full candidate basis contains $\binom{p+4}{4}$ terms, which yields the sequence $70, 210, 495, \ldots$ reported in Table 2 for degrees $4, 6, 8, \ldots$. Sparse orthogonal regression restricts this expansion

by retaining only a subset of the candidate terms,

$$\widehat{\text{Offset}}(T_a, v_a, T_r - T_a, rH) = \sum_{\boldsymbol{\alpha} \in S_p} c_{\boldsymbol{\alpha}} \prod_{j=1}^{4} P_{\alpha_j}(x_j), \tag{3}$$

where $S_p \subseteq \mathcal{A}_p$ is selected by the Lasso regularization. The number of active parameters therefore depends on two factors: the maximum polynomial degree, which determines the size of the candidate pool, and the regularization strength, which determines how many of those candidate terms are retained in the final model. This is the reason why the number of parameters changes across polynomial degrees and also along the Pareto fronts shown in Fig. 3. In this sense, the approximation can be viewed as a Fourier-like decomposition in an orthogonal polynomial basis, where lower-order terms capture the dominant structure of the UTCI offset and higher-order terms provide progressively finer corrections. A key advantage of the orthogonal basis is that it yields order-by-order consistency, see Fig. 4: when higher-degree terms are introduced, the coefficients associated with lower-order structure remain much more stable than in regressions based on ordinary monomials.

Linear regression without any sparsity constraints shows improved performance at higher degrees, with test loss reducing as model capacity increases. However, this comes with a dramatic increase in the number of parameters; it reaches over 1800 coefficients by degree 12. Furthermore, the discrepancy between train and test losses at higher degrees (e.g., 0.62°C vs. 0.54°C at degree 12) indicates overfitting, despite the improved predictive accuracy. The resulting models are also substantially more complex, raising concerns regarding interpretation and generalization. Sparse regression with standard polynomial bases show similar performance at low degrees but fails to converge beyond the 6th degree. This indicates that enforcing sparsity in a poorly conditioned basis becomes increasingly difficult as model complexity grows.

In contrast, sparse regression using an orthogonal Legendre basis (or sparse orthogonal regression) exhibits superior stability and accuracy across all degrees. It outperforms the baseline 6th-degree polynomial fit from degree 8th onward, achieving a test loss of 0.88°C at degree 10 with only 209 parameters—almost the same count as the original benchmark model, but with improved generalization. As the degree increases to 16, the loss reduces further to 0.60°C using 424 parameters—a fraction of those used by the corresponding standard regression model. The orthogonality of the Legendre basis likely contributes to better numerical conditioning, facilitating sparse model discovery even at high degrees. These results emphasize the importance of basis selection and regularization strategy in symbolic regression tasks. Sparse methods, when combined with well-structured bases like Legendre polynomials, offer a promising path toward accurate, compact, and interpretable models in high-dimensional settings.

Furthermore, optimization of nonlinear objective functions using gradient-based algorithms can be computationally intensive, especially in high-dimensional spaces where convergence is slow and local minima may hinder performance. In contrast, the regression-based approach proposed in this article—particularly

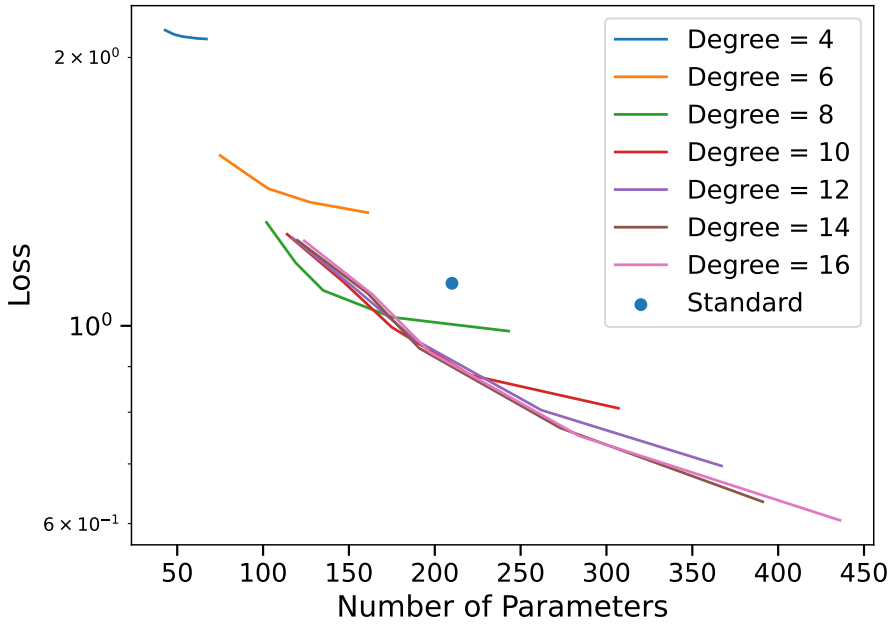



Fig. 3: Loss versus number of parameters for different polynomial degrees. The regularization parameter was varied in the lasso regression to yield a Pareto front in model accuracy and complexity for each degree.

through sparse regression with orthogonal polynomials—offers significantly faster computation. By framing the problem as a structured regression task rather than a nonlinear optimization, the method avoids costly iterative procedures and scales efficiently with dimensionality, making it highly suitable for rapid modeling of complex environmental indices like the UTCI.

Fig. 3 illustrates the relationship between model complexity (measured by the number of parameters) and prediction accuracy (log-scaled loss) for sparse regression models using Legendre polynomial bases of varying degrees. Each curve corresponds to a fixed polynomial degree, ranging from 4 to 16, with points reflecting models of increasing complexity obtained through regularization. A clear trend is observed: for a given polynomial degree, increasing the number of parameters generally results in improved model accuracy (i.e., lower loss). However, diminishing returns set in, and the rate of improvement flattens. More notably, the envelope formed by the lowest loss at each level of complexity across all degrees traces an emergent Pareto front (Smits and Kotanchek, 2005). This front captures the trade-off between model simplicity and predictive performance.

Higher-degree models (e.g., degrees 12–16) dominate this frontier at higher parameter counts, offering better loss with only marginal increases in complexity. In contrast, lower-degree models saturate quickly, highlighting their limited expressivity. The Pareto front thus reflects the optimal set of models that balance

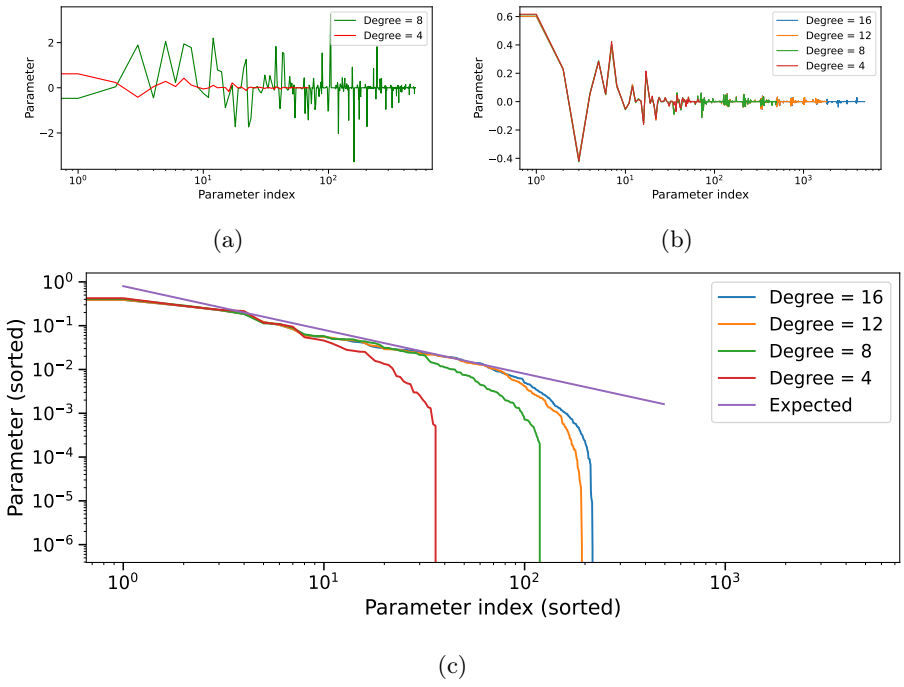



Fig. 4: Parameters (or polynomial coefficients) and how they change for different polynomial degrees for (a) simple regression and (b) sparse regression (using Legendre basis). (c) Sorted sparse-regression coefficients (Legendre basis) versus parameter index on a logarithmic x–axis show a clear, Fourier–like decay with order—approximately $1/n$—that is stable across model capacities (degrees 4, 8, 12, 16), indicating a hierarchical structure where lower–order terms dominate and higher–order terms provide incremental refinement.

accuracy and sparsity, guiding model selection under complexity constraints. The use of Legendre polynomials ensures numerical stability and encourages efficient basis representations, which supports the recovery of compact yet accurate models in this sparse regression setting.

The Fig. 4(a) and (b) we visualize the behavior of regression coefficients obtained from simple regression and sparse regression with orthogonal Legendre polynomials. Both plots use a logarithmic x-axis to indicate the parameter index and reveal how coefficients evolve as higher-degree polynomial terms are introduced. In Fig. 4(a), each line corresponds to simple regression solutions using polynomial bases of increasing degree. The x-axis denotes the index of polynomial terms (sorted or sequential), while the y-axis shows the corresponding coefficient values. A key observation is that the coefficients of lower-degree terms (left side of the plot) are not stable across model orders. As higher-degree terms are added, previously estimated lower-order coefficients shift significantly, often changing sign and magnitude.

Figure 4(b) presents coefficient values for sparse regression using Legendre polynomials, with colors indicating contributions from different polynomial degrees. Here, a contrasting pattern emerges: coefficients associated with lower-degree terms remain stable as higher-degree terms are added. New coefficients primarily emerge in the higher-order region of the x-axis, without disturbing the existing ones. This stability results from the orthogonality of the Legendre basis, which decorrelates the polynomial terms and enables additive refinement without re-tuning existing coefficients.

The contrast between the Figs. 4(a) and (b) underscores the advantage of orthogonal polynomial bases in sparse regression. Simple regression results in unstable, entangled coefficient estimates that shift with basis expansion, complicating interpretability and reuse. Sparse regression with ordinary polynomial bases fails to converge for higher degrees. In contrast, sparsity and orthogonal polynomials yield stable, hierarchical models where lower-order structure is preserved and higher-order terms incrementally enrich the representation. This behavior is particularly valuable for symbolic regression and interpretable modeling, where each term ideally reflects a distinct, meaningful contribution to the model output.

In Fourier analysis, the magnitude of coefficients typically decays as $1/n$ (where $n$ is the order of the term) for functions of bounded variation (Stein and Shakarchi, 2011) – a class that includes many naturally occurring signals and is a reasonable assumption for observational data. This decay reflects the fact that higher-order (or higher-frequency) components contribute less to the overall structure of such functions. A similar trend is observed in sparse regression using orthogonal polynomial bases, see Fig. 4(c). When coefficients are sorted by magnitude, they exhibit a clear decreasing pattern, analogous to the Fourier case, with lower-order terms capturing the dominant structure and higher-order terms refining the approximation in a controlled manner.

This suggests that through the use of sparse regression with an orthogonal polynomial basis, we have achieved a Fourier-like decomposition of the UTCI Offset in the Legendre basis (instead of the trigonometric one). This has a number of theoretical advantages: due to the orthogonality of the basis functions, the decomposition minimizes the $L_2$ distance (least squares) between approximation and function, guaranteeing the best possible polynomial fit for a given model complexity (Stein and Shakarchi, 2011). Additionally, the coefficients are uncorrelated and hierarchically structured, ensuring that lower-order components remain stable as higher-order terms are added—enhancing both interpretability and numerical robustness.

Based on the analysis results and one of the initial goals (that the new approximation should have comparable computational complexity to the existing one), we selected the sparse regression model based on tenth-degree Legendre polynomials as the most suitable approximation. The final version of the new polynomial, which has 209 coefficients, was calculated using the whole dataset of tabulated values.

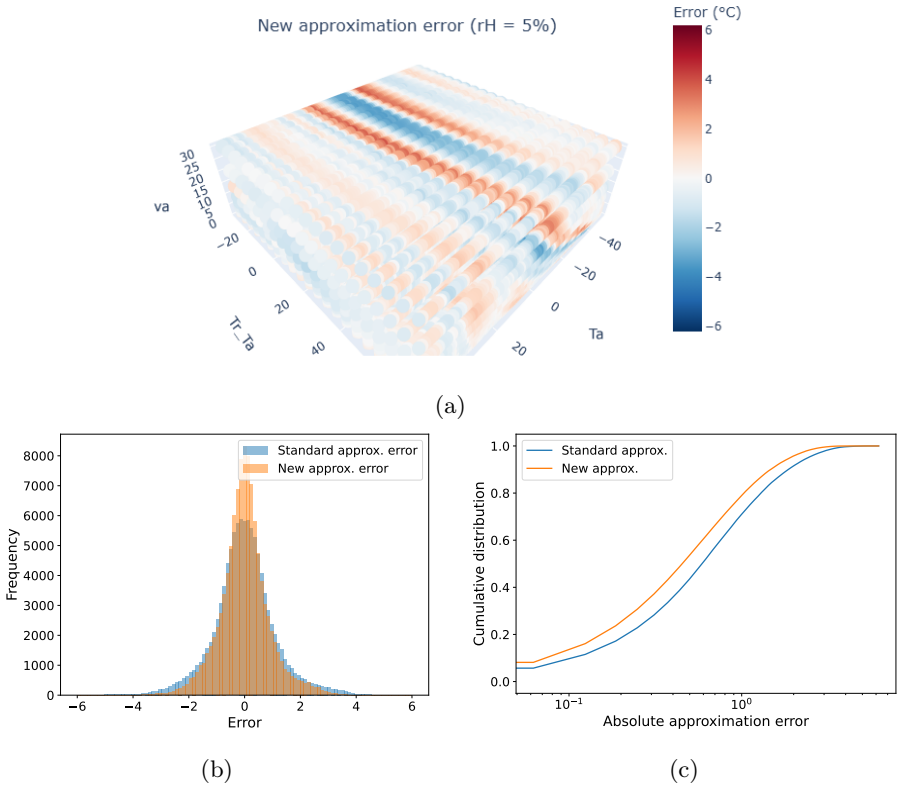



Fig. 5: (a) Spatial distribution of the UTCI Offset error (approximation minus reference) for the new sparse-model-based approximation at a fixed relative humidity of 5%, showing small, smoothly varying discrepancies. (b) Comparison of error histograms for the standard UTCI approximation and the new approximation based on the tenth-degree Legendre polynomials. (c) Cumulative distributions of the absolute errors of the two approximations.

Fig. 5(a) shows the spatial distribution of the Offset errors for the new approximation at a fixed relative humidity of 5%. The errors are small and smoothly varying, indicating good agreement across the input space. Fig. 5(b) presents a comparison of error histograms for both the standard and new approximations. The sparse-model-based approximation produces a narrower, more sharply peaked distribution centered at zero, highlighting a reduction in error variance and suggesting better generalization. Fig. 5(c) shows the cumulative distribution of absolute errors for the two approximations. The curve for the new approximation rises more steeply and reaches higher cumulative values at lower error thresholds, indicating that a larger proportion of predictions fall within smaller error margins.

Table 3 summarizes the most relevant properties of the two approximations. The results show a clear improvement in accuracy: the new approximation not

| | Standard approximation | New approximation |
|---|---|---|
| Polynomial degree | 6th | 10th |
| Basis functions | monomials | Legendre |
| Number of coefficients | 210 | 209 |
| Mean Error | $1.7 \cdot 10^{-3}$ °C (0.35 °C) | $-2.7 \cdot 10^{-15}$ °C (0.22 °C) |
| Mean Absolute Error | 0.81 °C (1.33 °C) | 0.64 °C (0.71 °C) |
| Root Mean Square Error | 1.17 °C (2.77 °C) | 0.88 °C (0.96 °C) |
| Freq. of abs. errors larger than 2°C | 8.4 % (15.5 %) | 4.2 % (5.0 %) |
| Freq. of abs. errors larger than 3°C | 2.2 % (6.3 %) | 0.50 % (0.60 %) |
| Freq. of abs. errors larger than 4°C | 0.34 % (3.8 %) | 0.011 % (0.10 %) |
| Freq. of abs. errors larger than 5°C | 0.038 % (3.3 %) | 0.00096 % (0 %) |

Table 3: Comparison of properties of the standard (Bröde et al., 2012) and new polynomial approximations of UTCI Offset function. The values outside of the parentheses reflect the evaluation of the approximations on the full dataset of 104 643 accurate Offset values provided by Bröde et al. (2012). The values shown in the parentheses reflect the evaluation using the independent dataset of 1000 accurate UTCI values (Bröde, 2021b), which were not used during the development of the new approximation. Both approximations are only valid for the intervals of environmental variables available in the full dataset (Table 1).

only substantially reduces the average errors (i.e, the mean error, the mean absolute error, and the root mean square error) but also drastically reduces the frequency of large deviations compared to the standard approximation. For example, the frequency of absolute errors larger than 2°C is halved from 8% to 4%, the frequency of errors larger than 3°C reduces from 2% to 0.5%, while the frequency of errors larger than 4°C reduces from 0.3% to 0.01%. These results clearly show the added benefits of the new approximation and confirm that the sparse regression approach can achieve comparable or improved predictive accuracy while maintaining interpretability and model parsimony.

We also evaluated the new approximation on the independent dataset of 1000 accurate UTCI values, which were not used during the development of the approximation. This dataset was prepared by the authors of the Bröde et al. (2012) paper, and is freely available on a Zenodo repository (Bröde, 2021b). Similarly to the evaluation of the new approximation on the full dataset, evaluation on the independent dataset shows a substantial reduction of the mean errors and a drastic reduction in the frequency of large errors compared to the standard approximation (Table 3).

Since the new approximation was determined using the full dataset of accurate Offset values (Bröde et al., 2012), it is, same as the standard approximation, only valid for the intervals of environmental variables available in this dataset (Table 1). Using the approximation for conditions outside of these intervals can potentially lead to large errors or unrealistic results and should be avoided.

## 4 Conclusions

The goal of this study was to develop an improved version of the polynomial approximation – one that would have comparable computational complexity to the existing approximation but would be more robust in terms of numerical stability and substantially more accurate, particularly in reducing the frequency of larger errors. This goal was successfully achieved using sparse regression with an orthogonal polynomial basis.

Sparse regression methods, such as LASSO, helped reduce overfitting and improve interpretability. As we have shown, the choice of basis functions is crucial: orthogonal polynomials like Legendre polynomials offer better numerical stability and conditioning than monomials. They enable hierarchical models where higher-order terms don't affect lower-order estimates, making them especially useful in sparse, interpretable models. Empirical results support these theoretical advantages.

Using sparse regression with an orthogonal polynomial basis (or sparse orthogonal regression), we have:

(a) Achieved substantially better accuracy – compared to the standard approximation, the new approximation not only substantially reduces the average errors (i.e, the mean error, the mean absolute error, and the root mean square error) but also drastically reduces the frequency of large errors.
(b) Retained a comparable computational complexity – the number of coefficients is almost the same for both approximations, meaning the computational complexity is comparable.
(c) Found a Pareto front for different model complexities – loss curves reveal that sparse models with orthogonal bases efficiently populate a Pareto front, balancing complexity and accuracy.
(d) Determined coefficients consistent over models with different capacities - coefficient plots for models built on orthogonal bases show the progressive inclusion of higher-order components without disrupting lower-order structure, in contrast to models using simple regression and ordinary polynomials.
(e) Achieved successful generalization – training the model over only 20% of the data, while testing was performed over the other 80%, highlights successful generalization. The results are also robust under bootstrapping.
(f) Essentially decomposed the UTCI in a Fourier expansion with a Legendre-polynomial basis, with parameters scaling as expected. Thus, we are arguably close to the theoretical optimum results for a robust approximation in the $L_2$ metric (or least squares).

Sparse orthogonal regression provides an effective framework for constructing accurate and numerically stable polynomial approximations of the UTCI. Our main contribution is therefore not methodological novelty in sparse regression itself, but the use of an orthogonal polynomial basis as a practical approximation strategy with favorable numerical properties, including order-by-order consistency and stable low-order truncations. In addition, the results obtained from random train–test splits, together with their robustness under bootstrapping,

show that using only 20% of the data for training is not a requirement of the method, but a deliberately stringent test of generalization. The comparable performance on the remaining 80% of the data indicates that the approach remains accurate, robust, and efficient even under a severe limitation in the number of training data points, while remaining well suited for practical applications that require portability and ease of implementation.

We have also prepared an easy-to-use Python function for the new approximation (please refer to the Code and data availability section on how to obtain the code). The code relies only on basic mathematical operations, which makes it easy to adapt to other programming languages, such as Fortran or C++. We also implemented a check to see if the environmental state falls within the domain of validity of the approximation. If this is not the case, the code produces a warning that the resulting UTCI values could have large errors or be unrealistic.

## Funding

This publication is supported by the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Postdoctoral Fellowship Programme, SMASH co-funded under the grant agreement No. 101081355. The operation (SMASH project) is co-funded by the Republic of Slovenia and the European Union from the European Regional Development Fund. The authors acknowledge the financial support of the Slovenian Research Agency via the Gravity project *AI for Science*, GC-0001 and of the Slovenian Research And Innovation Agency (research core funding No. P1-0188).

## Author Contributions

S.R. - Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing (original draft preparation), G.S. - Conceptualization, Resources, Validation, Software, Writing (review and editing), L.T. - Conceptualization, Methodology, Project administration, Supervision, Validation, Writing (review and editing), S.D. - Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing (review and editing).

## Code and data availability

The Python function code for the calculation of the new approximation as well as all the code needed to reproduce the results of the analysis is published on Zenodo (Roman, 2025a).

## Conflict of Interest Statement

The authors declare no conflicts of interest.

# Bibliography


Atanasova, N., Recknagel, F., Todorovski, L., Džeroski, S., and Kompare, B.: Computational assemblage of ordinary differential equations for chlorophyll-a using a lake process equation library and measured data of Lake Kasumigaura, Ecological Informatics: Scope, Techniques and Applications, pp. 409–427, 2006a.

Atanasova, N., Todorovski, L., Džeroski, S., Remec, Š. R., Recknagel, F., and Kompare, B.: Automated modelling of a food web in lake Bled using measured data and a library of domain knowledge, Ecological Modelling, 194, 37–48, 2006b.

Atanasova, N., Todorovski, L., Džeroski, S., and Kompare, B.: Application of automated model discovery from data and expert knowledge to a real-world domain: Lake Glumsø, ecological Modelling, 212, 92–98, 2008.

Atanasova, N., Džeroski, S., Kompare, B., Todorovski, L., and Gal, G.: Automated discovery of a model for dinoflagellate dynamics, Environmental Modelling & Software, 26, 658–668, 2011.

Blazejczyk, K., Epstein, Y., Jendritzky, G., Staiger, H., and Tinz, B.: Comparison of UTCI to selected thermal indices, International Journal of Biometeorology, 56, 515–535, https://doi.org/10.1007/s00484-011-0453-2, 2012.

Brence, J., Todorovski, L., and Džeroski, S.: Probabilistic grammars for equation discovery, arXiv preprint arXiv:2012.00428, 2020.

Brence, J., Džeroski, S., and Todorovski, L.: Dimensionally-consistent equation discovery through probabilistic attribute grammars, Information Sciences, 632, 742–756, 2023.

Bridewell, W., Asadi, N. B., Langley, P., and Todorovski, L.: Reducing overfitting in process model induction, in: Proceedings of the 22nd international conference on Machine learning, pp. 81–88, 2005.

Brimicombe, C., Napoli, C. D., Quintino, T., Pappenberger, F., Cornforth, R., and Cloke, H. L.: Thermofeel: A python thermal comfort indices library, SoftwareX, 18, 101 005, https://doi.org/10.1016/j.softx.2022.101005, 2022.

Bröde, P.: Issues in UTCI Calculation from a Decade's Experience, pp. 13–21, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-030-76716-7_2, 2021a.

Bröde, P.: UTCI-Test-Data, https://doi.org/10.5281/zenodo.5503967, 2021b.

Bröde, P., Fiala, D., Błażejczyk, K., Holmér, I., Jendritzky, G., Kampmann, B., Tinz, B., and Havenith, G.: Deriving the operational procedure for the Universal Thermal Climate Index (UTCI), International Journal of Biometeorology, 56, 481–494, https://doi.org/10.1007/s00484-011-0454-1, URL http://link.springer.com/10.1007/s00484-011-0454-1, 2012.

Brunton, S. L., Proctor, J. L., and Kutz, J. N.: Discovering governing equations from data by sparse identification of nonlinear dynamical systems, Proceedings of the national academy of sciences, 113, 3932–3937, 2016.

Błażejczyk, K.: BioKlima - Universal tool for bioclimatic and thermophysiological studies, URL https://www.igipz.pan.pl/bioklima-crd.html, [Accesed date: 10.10.2025.], 2025.

Błażejczyk, K. and Kuchcik, M.: UTCI applications in practice (methodological questions), Geographia Polonica, 94, https://doi.org/10.7163/GPol.0198, 2021.

Čerepnalkoski, D., Taškova, K., Todorovski, L., Atanasova, N., and Džeroski, S.: The influence of parameter fitting methods on model structure selection in automated modeling of aquatic ecosystems, Ecological Modelling, 245, 136–165, 2012.

Di Napoli, C., Messeri, A., Novák, M., Rio, J., Wieczorek, J., Morabito, M., Silva, P., Crisci, A., and Pappenberger, F.: The Universal Thermal Climate Index as an Operational Forecasting Tool of Human Biometeorological Conditions in Europe, in: Applications of the Universal Thermal Climate Index UTCI in Biometeorology, pp. 193–208, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-030-76716-7_10, URL https://link.springer.com/10.1007/978-3-030-76716-7_10, 2021a.

Di Napoli, C., Barnard, C., Prudhomme, C., Cloke, H. L., and Pappenberger, F.: ERA5-HEAT: A global gridded historical dataset of human thermal comfort indices from climate reanalysis, Geoscience Data Journal, 8, https://doi.org/10.1002/gdj3.102, 2021b.

Džeroski, S., Langley, P., and Todorovski, L.: Computational discovery of scientific knowledge, in: Computational discovery of scientific knowledge: Introduction, techniques, and applications in environmental and life sciences, pp. 1–14, Springer, 2007.

Fiala, D., Havenith, G., Brode, P., Kampmann, B., et al.: UTCI- Fiala multinode model of human heat transfer and temperature regulation, Int J Biometeorol., 56, https://doi.org/10.1007/s00484-011-0424-7, 2012.

Jeraj, M., Džeroski, S., Todorovski, L., and Debeljak, M.: Application of machine learning methods to palaeoecological data, Ecological modelling, 191, 159–169, 2006.

Kuzmanović, D., Banko, J., and Skok, G.: Improving the operational forecasts of outdoor Universal Thermal Climate Index with post-processing, International Journal of Biometeorology, 68, 965–977, https://doi.org/10.1007/s00484-024-02640-6, 2024.

Mežnar, S., Džeroski, S., and Todorovski, L.: Efficient generator of mathematical expressions for symbolic regression, Machine Learning, 112, 4563–4596, 2023.

Omejc, N., Gec, B., Brence, J., Todorovski, L., and Džeroski, S.: Probabilistic grammars for modeling dynamical systems from coarse, noisy, and partial data, Machine Learning, 113, 7689–7721, 2024.

Pappenberger, F., Jendritzky, G., Staiger, H., Dutra, E., Di Giuseppe, F., Richardson, D. S., and Cloke, H. L.: Global forecasting of thermal health hazards: the skill of probabilistic predictions of the Universal Thermal Climate In-

dex (UTCI), International Journal of Biometeorology, 59, 311–323, https://doi.org/10.1007/s00484-014-0843-3, URL http://link.springer.com/10.1007/s00484-014-0843-3, 2015.

Reid, S., Tibshirani, R., and Friedman, J.: A study of error variance estimation in lasso regression, Statistica Sinica, pp. 35–67, 2016.

Roman, S.: Historical dynamics of the Chinese dynasties, Heliyon, 7, 2021.

Roman, S.: Theories and models: Understanding and Predicting Societal Collapse, in: The Era of Global Risk: An Introduction to Existential Risk Studies, pp. 27–54, Open Book Publishers, URL https://doi.org/10.11647/OBP.0336.02, 2023.

Roman, S.: Code for Approximating the universal thermal climate index (UTCI) using sparse regression with orthogonal polynomials, https://doi.org/10.5281/zenodo.16880382, 2025a.

Roman, S.: Maximum Entropy Models for Unimodal Time Series: Case Studies of Universe 25 and St. Matthew Island, in: International Conference on Discovery Science, pp. 32–44, Springer, 2025b.

Roman, S. and Bertolotti, F.: A master equation for power laws, Royal Society open science, 9, 220 531, 2022.

Roman, S. and Bertolotti, F.: Global history, the emergence of chaos and inducing sustainability in networks of socio-ecological systems, Plos one, 18, e0293 391, 2023.

Roman, S. and Palmer, E.: The Growth and Decline of the Western Roman Empire: Quantifying the Dynamics of Army Size, Territory, and Coinage, Cliodynamics, 10, 2019.

Simidjievski, N., Todorovski, L., and Džeroski, S.: Learning ensembles of population dynamics models and their application to modelling aquatic ecosystems, Ecological Modelling, 306, 305–317, 2015.

Simidjievski, N., Todorovski, L., and Džeroski, S.: Modeling dynamic systems with efficient ensembles of process-based models, PloS one, 11, e0153 507, 2016.

Smits, G. F. and Kotanchek, M.: Pareto-front exploitation in symbolic regression, in: Genetic programming theory and practice II, pp. 283–299, Springer, 2005.

Stein, E. M. and Shakarchi, R.: Fourier analysis: an introduction, vol. 1, Princeton University Press, 2011.

Steinmann, P., Verstegen, J., Van Voorn, G., Roman, S., and Ligtenberg, A.: Scenario search: finding diverse, plausible and comprehensive scenario sets for complex systems, Socio-Environmental Systems Modelling, 7, 18 823–18 823, 2025.

Tanevski, J., Todorovski, L., and Džeroski, S.: Learning stochastic process-based models of dynamical systems from knowledge and data, BMC systems biology, 10, 1–17, 2016a.

Tanevski, J., Todorovski, L., and Džeroski, S.: Process-based design of dynamical biological systems, Scientific reports, 6, 34 107, 2016b.

Tanevski, J., Todorovski, L., and Džeroski, S.: Combinatorial search for selecting the structure of models of dynamical systems with equation discovery, Engineering Applications of Artificial Intelligence, 89, 103 423, 2020.

Tartarini, F. and Schiavon, S.: pythermalcomfort: A Python package for thermal comfort research, SoftwareX, 12, 100 578, https://doi.org/10.1016/j.softx.2020.100578, 2020.

Termonia, P., Fischer, C., Bazile, E., Bouyssel, F., Brozkova, R., Bénard, P., Bochenek, B., Degrauwe, D., Derková, M., Khatib, R., Hamdi, R., Mašek, J., Pottier, P., Pristov, N., Seity, Y., Smolikova, P., Španiel, O., Tudor, M., Wang, Y., and Joly, A.: The ALADIN System and its canonical model configurations AROME CY41T1 and ALARO CY40T1, Geoscientific Model Development, 11, https://doi.org/10.5194/gmd-11-257-2018, 2018.

Todorovski, L. and Džeroski, S.: Declarative bias in equation discovery, in: Proceedings of the International Conference on Machine Learning, pp. 376–384, 1997.

Todorovski, L. and Džeroski, S.: Theory revision in equation discovery, in: International Conference on Discovery Science, pp. 389–400, Springer, 2001.

Todorovski, L. and Džeroski, S.: Integrating knowledge-driven and data-driven approaches to modeling, Ecological Modelling, 194, 3–13, 2006.

Todorovski, L., Džeroski, S., and Kompare, B.: Modelling and prediction of phytoplankton growth with equation discovery, Ecological Modelling, 113, 71–81, 1998.