

1 Highlights

2 **Optimizing Construction Schedules: Nonparametric Modeling of** 3 **Weather-Induced Delays**

4 Mateus Borges Comito, Rafael Izbicki, Paulo do Canto Hubert Junior, Filipe
5 Moura

- 6 • A flexible non-parametric model was designed to estimate the proba-
7 bility distribution of construction project completion times by incorpo-
8 rating meteorological information.
- 9 • A mathematically rigorous tool for model comparison was proposed,
10 enabling hyperparameter optimization and the selection of the most
11 suitable forecasting model.
- 12 • Using climate data from the United Kingdom, the proposed model
13 demonstrated superior performance in estimating the probability dis-
14 tribution of construction project completion times.
- 15 • The model can be used for forecasting optimal start dates, estimating
16 project completion times, establishing contractual limits for expected
17 weather-related delays, and analyzing critical paths within the project.

18 Optimizing Construction Schedules: Nonparametric
19 Modeling of Weather-Induced Delays

20 Mateus Borges Comito^a, Rafael Izbicki^b, Paulo do Canto Hubert Junior^c,
21 Filipe Moura^d

^a*Department of Statistics, Federal University of São Carlos, São Carlos, São Paulo, Brazil*

^b*Department of Statistics, Federal University of São Carlos, São Carlos, São Paulo, Brazil*

^c*Technology and Data Science Department, FGV - EAESP, Rio de Janeiro, Rio de Janeiro, Brazil*

^d*CTO and Co-founder, EHab.co, United Kingdom*

22 **Abstract**

23 Effective management of construction projects encounters significant chal-
24 lenges, particularly due to frequent delays influenced by climatic variables.
25 Anticipating weather-induced delays is crucial, and while various methods
26 such as stochastic generators, productivity impact models, and machine
27 learning tools are employed to model such gap, there remains a notable gap.
28 Specifically, there is a lack of direct approaches that utilize easily accessi-
29 ble historical meteorological data to model delays. This paper proposes a
30 flexible non-parametric model designed to estimate the probability distribu-
31 tion of construction project completion times by incorporating meteorological
32 information. The proposed model enhances PERT-based approaches by in-
33 corporating climatic effects to improve delay distribution estimation. This
34 enhancement involves the direct modeling of productivity, coupled with the
35 use of Monte Carlo simulation. A commercial version of the algorithm pre-
36 sented is currently part of a solution utilized in over 70 projects across the
37 UK. The results demonstrate the model's utility in forecasting optimal start
38 dates, estimating project completion times, establishing contractual limits
39 for expected weather-related delays, and analyzing critical paths within the
40 project. Additionally, a mathematically rigorous tool for model comparison
41 is presented, allowing for hyperparameter optimization and the selection of
42 the most suitable prediction model.

43 *Keywords:* Weather, Project completion times, Distribution estimation,

45 **1. Introduction**

46 Construction delays can arise from various factors, including incompetence, material and equipment shortages, and adverse weather conditions
47 [1, 2, 3]. These delays negatively impact both owners and contractors, leading
48 to contractual penalties and reduced productivity. Among these factors,
49 weather events are particularly significant due to their direct effect on productivity
50 [4, 5]. Modeling weather-induced delays is challenging because
51 weather events vary in time, location, and the specific conditions required
52 for different construction tasks. Consequently, effective project management
53 must accurately account for these conditions to minimize negative impacts
54 on productivity and deadlines.
55

56 This paper introduces a model designed to estimate the uncertainty in
57 a project's completion time due to weather conditions. By estimating the
58 probability distribution of Z , the project's duration in days, this approach
59 enhances decision-making throughout the construction process. It helps determine
60 the optimal start date and estimate the completion date by calculating
61 distribution quantiles. Additionally, the model can be updated during
62 the project to provide real-time estimates of the completion date and assess
63 whether weather conditions were atypical, relying solely on historical meteorological
64 data. A commercial version of the algorithm presented in this
65 paper is currently part of a solution utilized in over 70 projects across the
66 UK, serving more than 15 clients.

67 Furthermore, a mathematically rigorous tool is proposed for optimizing
68 the hyperparameters of our model, as well as those used in other models.
69 The tool also facilitates the comparison of different models based on their
70 performance in predicting delay times.

71 *1.1. Contributions*

72 The uncertainty surrounding Z is crucial for assertive decision-making;
73 however, most studies focus on obtaining point predictions. Therefore, the
74 main contribution of this work is to present a new non-parametric model to
75 estimate the probability distribution of Z , capable of incorporating weather
76 information to skillfully model delays.

77 The proposed approach directly models the probability of executing each
78 task on a given day, combining this information to estimate Z . This involves

79 modeling the temporal dynamics of task performance throughout the year,
80 considering climatic factors. The execution of a task on any given day is
81 represented by a Bernoulli random variable with a daily execution probability.
82 This entails a stochastic process, described by a Markov Chain in discrete
83 time. Historical meteorological data is then used to estimate the parameters
84 of the chain, and the distribution of Z is obtained through Monte Carlo
85 simulation.

86 The model has hyperparameters that control the learning process. As
87 a second independent contribution of this work, a mathematically rigorous
88 tool for selecting the hyperparameter values is presented. This tool can also
89 be used to compare and select hyperparameters of *any* models, not only the
90 ones proposed in this paper.

91 To demonstrate the practical utility of our method, Figure 1 showcases a
92 byproduct of the model that estimates project durations as a function of the
93 starting date. The results highlight the model’s effectiveness in providing
94 accurate and actionable insights for project scheduling. The comparison
95 between the model’s estimates and the actual project durations, represented
96 by the observed lines at the bottom, underscores the method’s precision and
97 reliability. Further details about the figure can be found in the section that
98 presents real-world examples.

99 1.2. Relation to Other Work

100 1.2.1. Estimating Z

101 Numerous studies investigate delays in construction projects using vari-
102 ous variables and datasets. A conventional two-phase approach involves using
103 climate generation models to simulate future climatic values day by day and
104 subsequently projecting their impact on the construction schedule [5]. The
105 first phase can be conducted using stochastic weather models [14, 6, 7, 8, 9]
106 or Machine Learning methods (ML) [29]. The second phase assesses how
107 the generated climate affects productivity, with substantial literature ded-
108 icated to understanding the impact of weather and other factors on labor
109 productivity [10].

110 The two-phase approach can be effective in estimating if reliable stochas-
111 tic models are available, especially for short-term projects. However, this
112 approach differs from the one proposed in this work, which directly mod-
113 els productivity. The model selection tool proposed in this work can be
114 used to compare any set of models, including those based on the two-phase

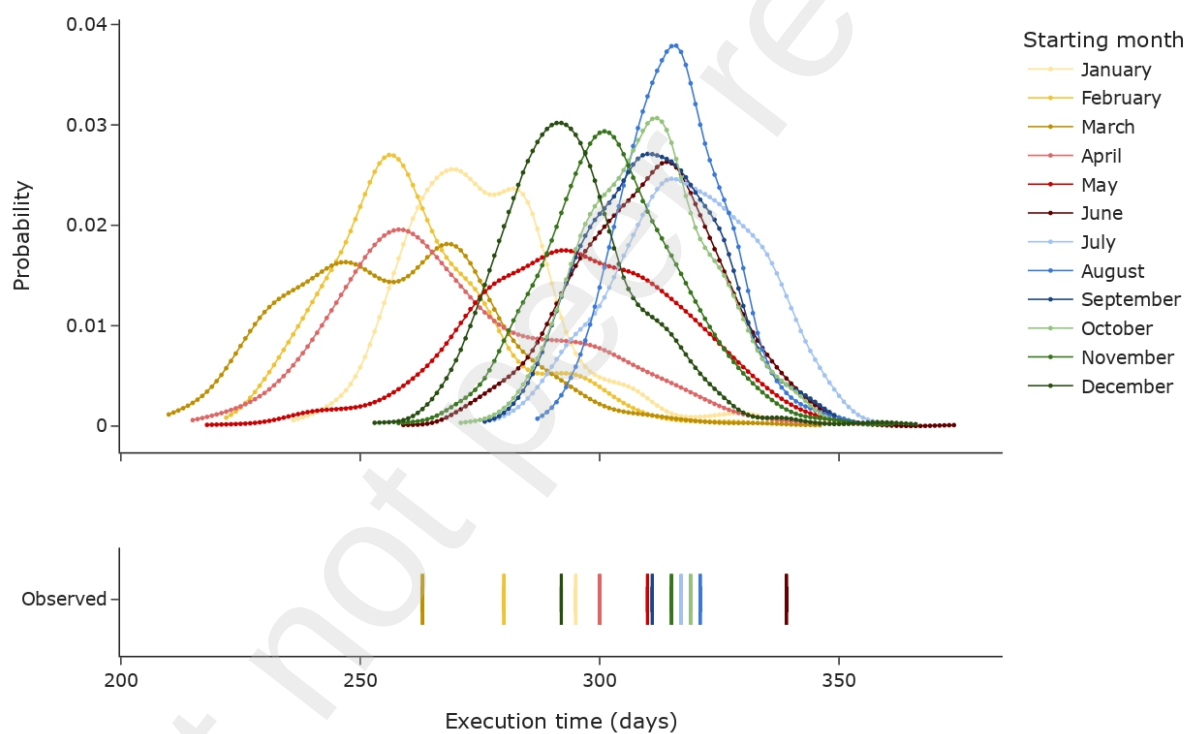


Figure 1: Illustrative application of the method using real-world data: the model estimates the distribution of project durations as a function of its starting date, providing valuable insights for optimal project scheduling. The observed lines at the bottom represent actual project durations, demonstrating the high accuracy of the model's estimates.

115 approach. Finally, it is worth noting that there are studies focused exclu-
116 sively on modeling productivity impact, without necessarily predicting delays
117 [30, 31, 32, 33, 10, 11, 12]. In these cases, researchers make inferences with
118 the objective of interpreting the parameters involved in the model. While
119 these impact-centric studies provide valuable insights into productivity loss
120 and mitigation strategies but do not aim to model overall project delays.

121 ML methods can also be used to directly predict the total project du-
122 ration [13, 18, 19, 20, 21]. Generally, these studies identify general factors
123 impacting delays, measure their impact primarily through surveys or histori-
124 cal data from other projects, and train algorithms based on this information.
125 The impact of weather is often represented by a variable that captures ex-
126 ternal factors. The dataset used in these studies significantly differs from
127 the one utilized here, as only historical meteorological data is employed here.
128 Although it is possible to provide uncertainties, ML methods are typically
129 employed to offer point predictions. Recently, some studies have used models
130 that estimate uncertainties to evaluate project costs [22].

131 Another well-established method for analyzing project schedules and un-
132 certainties is the Program Evaluation and Review Technique (PERT) [23].
133 Extensive literature proposes various extensions and case studies [24, 25, 26,
134 15, 27]. PERT requires parametric distributions of activity durations, which
135 are provided by experts or derived from other projects. The construction
136 schedule is simulated using values from these distributions, and the estimate
137 of Z is obtained through Monte Carlo simulations.

138 Ballesteros-Pérez et al. [34] developed a model as an extension of PERT
139 where task durations are determined not by parametric distributions but
140 by Bernoulli experiments that assess the feasibility of performing tasks on
141 each day of the year until the planned duration is achieved. In this model,
142 uncertainty arises directly from meteorological data, and daily productivity
143 follows a sinusoidal pattern throughout the year, making it parametric.

144 The model proposed here extends this sinusoidal model by directly mod-
145 eling working days using historical meteorological data. While it includes the
146 sinusoidal wave model, it also offers flexibility to choose other methods, such
147 as a non-parametric version and the use of Markov chains to introduce depen-
148 dence between days. Non-parametric approaches generally perform better
149 with larger training datasets [28], so it is expected that the non-parametric
150 version will excel when abundant historical meteorological data are available
151 or when the data do not conform to the expected parametric behavior.

1.2.2. Model Selection

Validation of delay prediction models in construction projects is often conducted through case studies [17, 7, 6, 8, 14, 16, 15]. Although these studies provide valuable insights into the models, they offer limited information on how different approaches compare. The proposed tools complement case studies by providing a mathematically rigorous framework for comparing models that estimate the distribution of Z .

An alternative approach to comparing models involves using cross-validation or data splitting to calculate machine learning (ML) metrics that assess the accuracy of predictions [13]. However, rather than relying on traditional ML metrics like mean squared error, which are intended for point prediction comparisons, the loss functions introduced here are specifically adapted to evaluate the performance of estimates of the distribution of Z .

2. Materials and methods

2.1. Overview

The steps required to estimate the execution time of a construction project and measure the predictive quality of the model are summarized in Figure 2 and detailed in the following sections.

First, the construction schedule is defined. This schedule outlines the tasks needed to produce the deliverables, including their sequence, durations, and resource requirements [43]. Figure 3 provides an example of a fictitious schedule with tasks carried out sequentially, without accounting for weather-related delays.

Next, the necessary climatic conditions for each task must be determined. Table 1, adapted from Ballesteros-Pérez et al. [34], shows the climatic thresholds that determine when specific tasks cannot be performed. For example, "earthworks" cannot be done if the average daily temperature is below 0°C. In this study, it is assumed that climatic limits are assumed for the performance of tasks and that their impact does not extend to subsequent days.

Using historical weather data and the defined climatic conditions, the days in past years is identified when each task could have been performed. This helps estimate the likelihood of completing each task on any given day of the year. Monte Carlo simulation is then used to generate samples of the distribution of Z , allowing us to estimate the distribution and identify important details about delays, such as quantiles and variances. The next section explains this estimation process in detail.

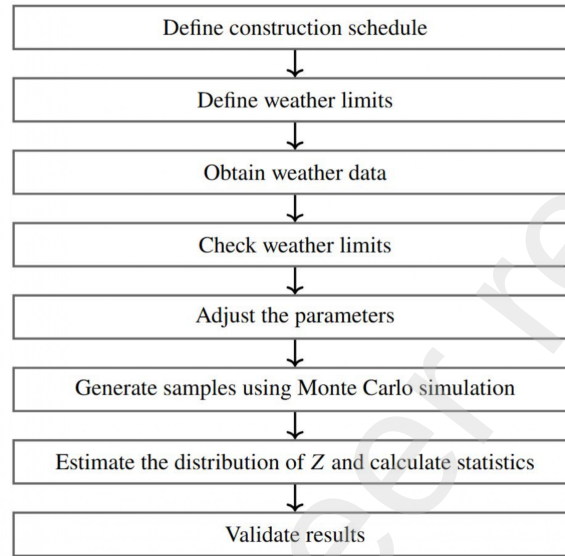


Figure 2: Sequence of steps to estimate the construction project and measure the predictive quality of the model.

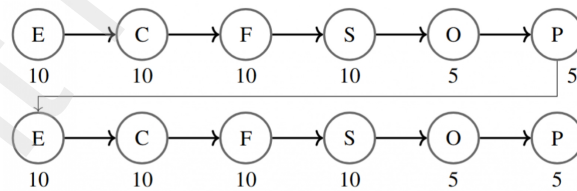


Figure 3: Construction schedule represented by a finite network of arcs, where the nodes represent the activities of earthworks (E), concrete (C), formwork (F), steelwork (S), outdoor painting (O), and pavements (P). For each task, a planned duration is associated below the node.

Weather variable	E	C	F	S	O	P
Minimum temperature $\leq 0^{\circ}\text{C}$		x				x
Average temperature $\leq 0^{\circ}\text{C}$	x			x	x	
Maximum temperature $\geq 40^{\circ}\text{C}$		x		x		x
Precipitation ≥ 1 mm					x	x
Precipitation ≥ 10 mm	x	x			x	x
Precipitation ≥ 30 mm	x	x		x	x	x
Wind gusts ≥ 30 knots		x	x	x	x	

Table 1: Weather limits for type of tasks: earthworks (E), concrete (C), formwork (F), steelwork (S), outdoor painting (O), and pavements (P). “x” denotes conditions under which the task cannot be performed.

2.2. Decomposing the project execution time

Consider a project composed of a set of p tasks, represented by $t = \{1, \dots, p\}$. The schedule of Figure 3 contains $p = 12$ tasks. Although the tasks do not overlap in this schedule, the method can be easily applied when they do intersect. The project execution time, Z , is defined as the number of days until all p tasks are completed, taking delays into account. The objective is to model the distribution of Z by considering the execution time of each task, including delays due to climate factors.

To achieve this, let the random variable $X_{i,y}^{(t)}$ indicate whether task t was executed on day i of year y considering climatic factors:

$$X_{i,y}^{(t)} = \begin{cases} 1 & \text{if task } t \text{ was executed on day } i \text{ of year } y, \\ 0 & \text{otherwise.} \end{cases}$$

It is assumed that the annual behavior repeats each year, so $X_{i,y}^{(t)}$ and $X_{i,y'}^{(t)}$ share the same distribution for all y, y' . Two models are provided: the first assumes that, for each task t , $X_{i,y}^{(t)}$ are independent random variables. The second uses a discrete-time Markov chain.

More concretely, the first model assumes $X_{i,y}^{(t)} \sim \text{Bernoulli}(\theta_i^{(t)})$, where $\theta_i^{(t)}$ is the probability of completing task t on the i -day of a given year, with $X_{i,y}^{(t)}$'s independent for each fixed t . The second model uses a first-order Markov chain with transition probabilities:

$$\mathbb{P}(X_{i,y}^{(t)} = 1 | X_{i-1,y}^{(t)} = s) = P_i^{(t)}(s).$$

Notice that, given a starting day and year, Z is a deterministic function of $X_{i,y}^{(t)}$'s. Therefore, after $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, \dots, \theta_{365}^{(t)})$, $\mathbf{P}^{(t)}(\mathbf{0}) = (P_1^{(t)}(0), \dots, P_{365}^{(t)}(0))$, or $\mathbf{P}^{(t)}(\mathbf{1}) = (P_1^{(t)}(1), \dots, P_{365}^{(t)}(1))$ have been estimated (see the next section for details), Z can be sampled from this model by using Monte Carlo simulation to sample $X_{i,y}^{(t)}$'s with the estimated parameters. This method easily incorporates task dependencies (see Supplementary Material S1 for details). For multiple tasks on the same day, each task is checked independently.

The Monte Carlo samples can be used to approximate α -quantiles and the mean of Z . They can also estimate Z 's distribution via kernel density estimation or discrete versions [41, 42].

2.3. Estimating the model parameters

This section describes the methodology for estimating $\theta_i^{(t)}$, $P_i^{(t)}(0)$ and $P_i^{(t)}(1)$. The datasets $\mathbf{H}^{(1)}, \dots, \mathbf{H}^{(\mathfrak{p})}$ represent three-dimensional binary matrices for each task. The element $H_{l,y,i}^{(t)} \in \{0, 1\}$ indicates whether task t was executable at location l and on the i -th day of year y .

To estimate $\theta_i^{(t)}$, a smoothing function $f^{(t)}$ is applied to $\mathbf{H}^{(t)}$:

$$\hat{\theta}_i^{(t)} = f^{(t)}(\mathbf{H}^{(t)}).$$

The function $f^{(t)}$ is a weighted average, using weights w_{loc} , w_{doy} , and w_{year} for location, day of the year, and year, respectively:

$$\hat{\theta}_i^{(t)} = \frac{\sum_{l \in \mathcal{L}} \sum_{i'=1}^{365} \sum_{y \in \mathcal{Y}} w_{\text{loc}}(l_0, l) w_{\text{doy}}(i, i') w_{\text{year}}(y_0, y) H_{l,y,i'}^{(t)} \mathbb{I}(H_{l,y,i'}^{(t)} \neq \emptyset)}{\sum_{l \in \mathcal{L}} \sum_{i'=1}^{365} \sum_{y \in \mathcal{Y}} w_{\text{loc}}(l_0, l) w_{\text{doy}}(i, i') w_{\text{year}}(y_0, y) \mathbb{I}(H_{l,y,i'}^{(t)} \neq \emptyset)}.$$

The weights measure proximity, ensuring the sum equals 1 and ignoring null values. This non-parametric approach does not assume specific data behavior.

The vector $\hat{\boldsymbol{\theta}}^{(t)}$ represents the probabilities of executing task t throughout the year. After smoothing, a parametric model $f_s^{(t)}$ (e.g., sinusoidal waves) can be fitted to $\hat{\boldsymbol{\theta}}^{(t)}$.

2.3.1. Choosing $w_{\text{year}}(y_0, y)$ - Exponential decay

Assuming the distribution of climatic conditions over the years is stationary can be plausible. However, weighting years differently can help capture temporal trends. If climate change affects conditions or recent years are more representative, weights with exponential decay can be used:

$$w_{\text{year}}(y_0, y) = \alpha^{y_0 - y},$$

where y_0 is the project start year and $\alpha > 0$ defines the decay degree, with $\alpha = 1$ indicating a stationary distribution. This method allows for capturing climate change effects by giving more weight to recent years. For instance, if a task depends on average temperatures that are rising, it may be suitable to assign greater weight to recent years.

2.3.2. Choosing $w_{\text{doy}}(i, i')$ - Moving average

Adopting a moving average approach, weights can be defined as:

$$w_{\text{doy}}(i, i') = \begin{cases} 1 & , \text{ if } (|i - i'| \bmod [365 - k]) < k \\ 0 & , \text{ other wise} \end{cases}$$

where k represents the size of the moving window. Note that the end and the beginning of the year are connected and close. If $k = 0$, only the information from that day is used, resulting in an estimator with high variance; $k \rightarrow \infty$, this variance is reduced, but the bias is increased.

2.3.3. Choosing $w_{\text{loc}}(l_0, l)$ - Euclidian distance

To model the weights associated with the construction location, distance measures or interpolation methods can be used. For example, greater weights can be assigned to locations closer to location l_0 , where the project will be executed. This can be done as follows:

$$w_{\text{loc}}(l_0, l) = \frac{1}{d(l_0, l)},$$

where $d(l_0, l)$ represents the distance between locations l_0 and l . The Euclidean distance between latitude and longitude is commonly used, but other measures, such as the geodesic distance, which better considers the globe's topology, can be adopted as needed. When $|\mathcal{L}| = 1$, weights are not needed, so $w_{\text{loc}}(l_0, l) = 1$.

2.3.4. Choosing $f_s^{(t)}$ - Senoidal wave

A parametric behavior in the data can be considered, such as sinusoidal wave expressions [15]:

$$\tilde{\theta}_i^{(t)'} = f_s^{(t)}(\tilde{\theta}_i^{(t)}) = K^{(t)} + A^{(t)} \cos \left(2\pi f^{(t)} \left[\frac{j}{365} - \varphi^{(t)} \right] \right),$$

where $K^{(t)}$ is the vertical offset, $A^{(t)}$ is the wave $f^{(t)}$ is the frequency (with $f^{(t)} = 1$ corresponding to a complete oscillation), and $\varphi^{(t)}$ corresponds to the day of the year when the sinusoidal wave reaches its maximum value. In this case, the values of $K^{(t)}$, $A^{(t)}$ and $f^{(t)}$ are obtained by fitting the model to $\hat{\theta}^{(t)}$ using least squares.

2.3.5. Markov Chain

The values of $\hat{P}_i^{(t)}(s)$ can be estimated similarly to $\hat{\theta}_i^{(t)}$ by using only the elements of the matrix $\mathbf{H}^{(t)}$ that correspond to the same states in the transition matrix. Elements that do not match the state sequence of the transition matrix are set to null ($\emptyset$). In other words, only information from the previous day that matches the same state $s \in \{0, 1\}$ is utilized, while all other elements are treated as null.

2.3.6. Hyperparameters

Table 2 lists the model's hyperparameters that need to be selected for the method presented here. In the next section, loss functions will be presented that can be used to select the best combination of hyperparameters for the dataset. An application of hyperparameter selection using simulated data is available in Supplementary Material S2.

Hyperparameter	Range	Description
<i>alpha</i>	$[0, 1]$	Degree of exponential decay.
<i>smoothing</i>	$\mathbb{N}^+$	Size of the moving window.
<i>cosine</i>	Boolean	If True, a sinusoidal wave is fitted.
<i>mc</i>	Boolean	If True, a Markov chain is used.

Table 2: List of hyperparameters used in the model, with their respective ranges and detailed descriptions.

2.4. Model selection

The dataset is temporally split into two parts: training and validation. Using the training set, n estimates for $p(z|j, y)$ are created: $p_1(z|i, y), \dots, p_n(z|j, y)$, where j represents the day of the year and y the year of project implementation. Each estimate may originate from different configurations of tuning parameters of a given model, or they may be entirely different models.

The goal of this section is to select the model that best approximates the true f . To achieve this, variations of the validation datasets, denoted

as $(Z_1, I_1, Y_1), \dots, (Z_m, I_m, Y_m)$, are created using various temporal splits of the original validation dataset. Specifically, different starting dates (I_k, Y_k) are randomly chosen. The Z_k values are then obtained by applying the construction schedule to data after the chosen start date, reflecting actual observed weather conditions.

Given a loss function L , the risk of an estimate $\hat{f}$ can then be estimated via

$$\hat{R}(f, \hat{f}) = \frac{1}{m} \sum_{k=1}^m L(Z_k, \hat{p}(z|I_k, Y_k)).$$

Two loss functions were considered: one measuring $\hat{p}$'s performance directly and another focusing on estimating the upper tails of the distribution $p(z)$. More details and comparisons are available in Supplementary Material S3.

2.4.1. Density loss

Inspired by the CDE loss for continuous problems [37, 38, 39, 40], the loss function L directly compares the estimated distribution with the true distribution,

$$L(p, \hat{p}) = \sum_z (p(z|i, y) - \hat{p}(z|i, y))^2,$$

where L depends on the true distribution $p(z)$, making it impossible to calculate. Fortunately, $\hat{R}$ can be decomposed using the expression

$$L(p, \hat{p}) = \sum_z \hat{p}^2(z|i, y) - 2\hat{p}(Z|i, y) + K,$$

where K is a constant that does not depend on $\hat{p}$. Thus, K does not need to be calculated to rank the model $\hat{p}$ according to L .

The term $\sum_z \hat{p}^2(z|i, y)$ takes into account the shape of the distribution, penalizing distributions that are too concentrated. On the other hand, the term $\hat{p}(Z|i, y)$ represents the probability associated with the observed value, and it is desirable that this probability is high for the observed value. It is important to note that if the density is poorly estimated and does not contain the observed value, only the shape of the distribution is considered in the loss function.

311 2.4.2. Weighted pinball loss

312 The proposed loss is based on the pinball loss $L_\alpha(z, \hat{z}_\alpha)$, commonly used
313 in α -quantile regression [35, 36]:

$$L_\alpha(z, \hat{z}_\alpha) = \begin{cases} (z - \hat{z}_\alpha)\alpha & \text{if } z \geq \hat{z}_\alpha \\ (\hat{z}_\alpha - z)(1 - \alpha) & \text{otherwise} \end{cases}$$

314 where z is the actual construction duration and $\hat{z}_\alpha$ is the estimated quantile.
315 The approach presented here extends the standard pinball loss by combining
316 multiple α values to estimate model uncertainty, using a weighting function
317 w to assign weights to each α . The weighted loss $L^w(Z, \hat{p})$ is defined as:

$$L^w(Z, \hat{p}) = \int w(\alpha) L_\alpha(Z, Q(\alpha)) d\alpha$$

318 where $Q(\alpha)$ is the α -quantile of $\hat{p}$.

319 The weighting function $w(\alpha)$ can be chosen to emphasize different parts
320 of the distribution based on the application. In construction projects, es-
321 timating the upper tails of $p(z)$ is crucial to ensure the observed value lies
322 within the density with high probability. Thus, weighting the right tail, e.g.,
323 $w(\alpha) \propto \alpha^2$, is appropriate, making higher α values more significant. Alterna-
324 tively, other weights can adjust the focus on different quantiles. For instance,
325 the left tail can be emphasized using $w(\alpha) \propto (1 - \alpha)^2$, or both tails equally
326 with $w(\alpha) \propto (\alpha - 0.5)^2$.

327 3. Results and discussion

328 Daily meteorological data from the UK, following criteria similar to Ballesteros-
329 Pérez et al. [34], were used. The MIDAS dataset [44] includes daily records
330 of precipitation, temperature, and wind speed from 1985 to 2020. Similar
331 inclusion and exclusion criteria for weather stations were applied. Missing
332 values were imputed with a 7-day moving average if at least three observa-
333 tions were available. Stations with significant data gaps in the validation and
334 test sets were excluded, resulting in 68 stations selected for analysis.

335 The schedule from Figure 3 and the climatic limits from Table 1 are
336 used. The fixed project duration is 160 days, disregarding regional climatic
337 conditions. Due to many missing values for hail, snow, and thunderstorms, a
338 reduced set of variables is used. The probability of workable days is estimated
339 using only days with complete data for all selected variables. The codes used
340 in this study are openly available on Github [48].

3.1. Test dataset

The climate in the UK is changing, with recent decades being warmer, wetter, and sunnier than the 20th century [45]. In the analysis, 2018 was identified as an outlier due to a shorter project execution time. To highlight this, project durations were compared by climate station and month against previous years. The real quantile was calculated by comparing each value with historical values for the same station and month. Figure 4 shows the expected and observed quantiles from 2017 to 2020, highlighting the atypical nature of 2018. According to Kendon et al.[46], the average temperature in the UK for 2018 was 9.5°C, which is 0.6°C above the long-term average of 1981–2010, making it the seventh warmest year in the UK series since 1884. Additionally, rainfall was below average, with only 92% of the 1981–2010 average, making it the driest year since 2010. In contrast, the amount of sunshine was 114% of the 1981–2010 average, making 2018 the third sunniest year since 1929. Notably, the summer of 2018 was one of the hottest, driest, and sunniest summers experienced by the UK in over 100 years, being the hottest summer since 1884 and the driest since 1995. The combination of these climatic factors contributed to a faster-than-expected project execution time, explaining the real quantiles being significantly lower than the expected quantiles. Due to this, the selected test set includes the years 2017, 2019, and 2020.

3.2. Hyperparameter tuning

Hyperparameter optimization was performed weekly for each weather station from 2013 to 2016, excluding holidays and weekends. The hyperparameters considered were:

- $\alpha = \{0.9, 0.95, 1\}$;
- $\text{smoothing} = \{0, 7\}$;
- $\text{cossine} = \{False, True\}$;
- $\text{mc} = \{False, True\}$.

Figure 5 shows the count of stations for each hyperparameter combination. The results indicated that $\alpha = 1$ performed best in 65% (44) of the stations, while other stations prioritized more recent years. In 81% (55) of the stations, Markov chains improved the model. In 47% (32) of the stations,

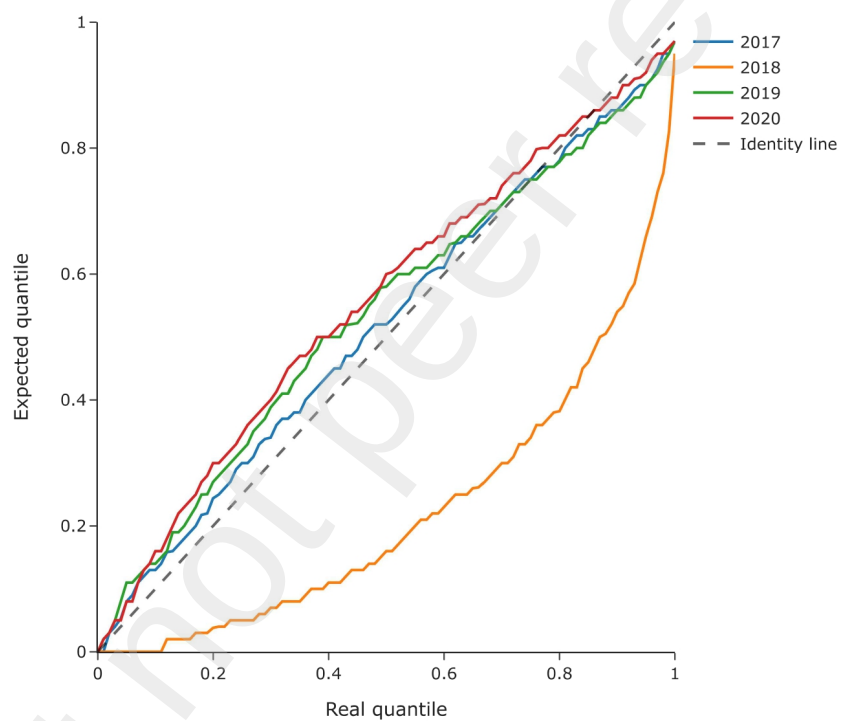


Figure 4: Comparison of expected quantile with real quantile for years 2017, 2018, 2019, and 2020.

374 sinusoidal waves were fitted after smoothing, indicating the data's paramet-
375 ric behavior. In 37% (25) of the stations, sinusoidal waves combined with
376 Markov chains showed that conditional probabilities have annual seasonali-
377 ties. These findings allow for a rigorous comparison of estimated distributions
378 and the best model selection considering all uncertainties.

379 3.3. Model selection

380 The proposed model was compared with a conventional two-step ap-
381 proach, which uses a weather generation model to simulate daily climatic
382 values and project the impact on the construction schedule. The model by
383 Apipattanavis et al. [47], employing a semiparametric weather generator,
384 was chosen for comparison. This model uses a first-order Markov chain to
385 model precipitation states and the k-NN method to approximate the condi-
386 tional PDF. Neighbors are selected based on the Markov chain sequence, and
387 climatic variables are simulated using normalized weights.

388 The primary selection criterion was superior performance in the right-tail
389 pinball loss, focusing on upper quantile estimation. If no significant difference
390 was found, overall pinball loss performance was used to choose a model that
391 estimates the entire distribution well. Density loss was also considered to
392 avoid overly concentrated distributions. The Wilcoxon test was used for
393 sample comparison at a 5% significance level.

394 Results showed no statistical difference in 40% (27) of the cases. The pro-
395 posed model performed better in 49% (33) of the cases, while the compared
396 model performed better in 12% (8) of the cases. The proposed loss functions
397 helped compare models and select the most suitable one for each dataset.

398 3.4. Project Start Date

399 Station 908 was chosen to demonstrate the model's application. Figure 6
400 shows estimated probabilities throughout the year, indicating higher proba-
401 bilities in mid-year. Figure 1 presents the estimated and actual project start
402 dates for each month, ranging from 263 to 339 days. This variation high-
403 lights the importance of selecting an appropriate start date to reduce project
404 duration and costs [15].

405 The distributions for March and February show the shortest execution
406 times, starting near periods of highest task execution probabilities. January
407 and April also benefit from high probability periods mid-year. Densities for
408 the second half of the year are higher due to lower probabilities, resulting in
409 longer execution times.

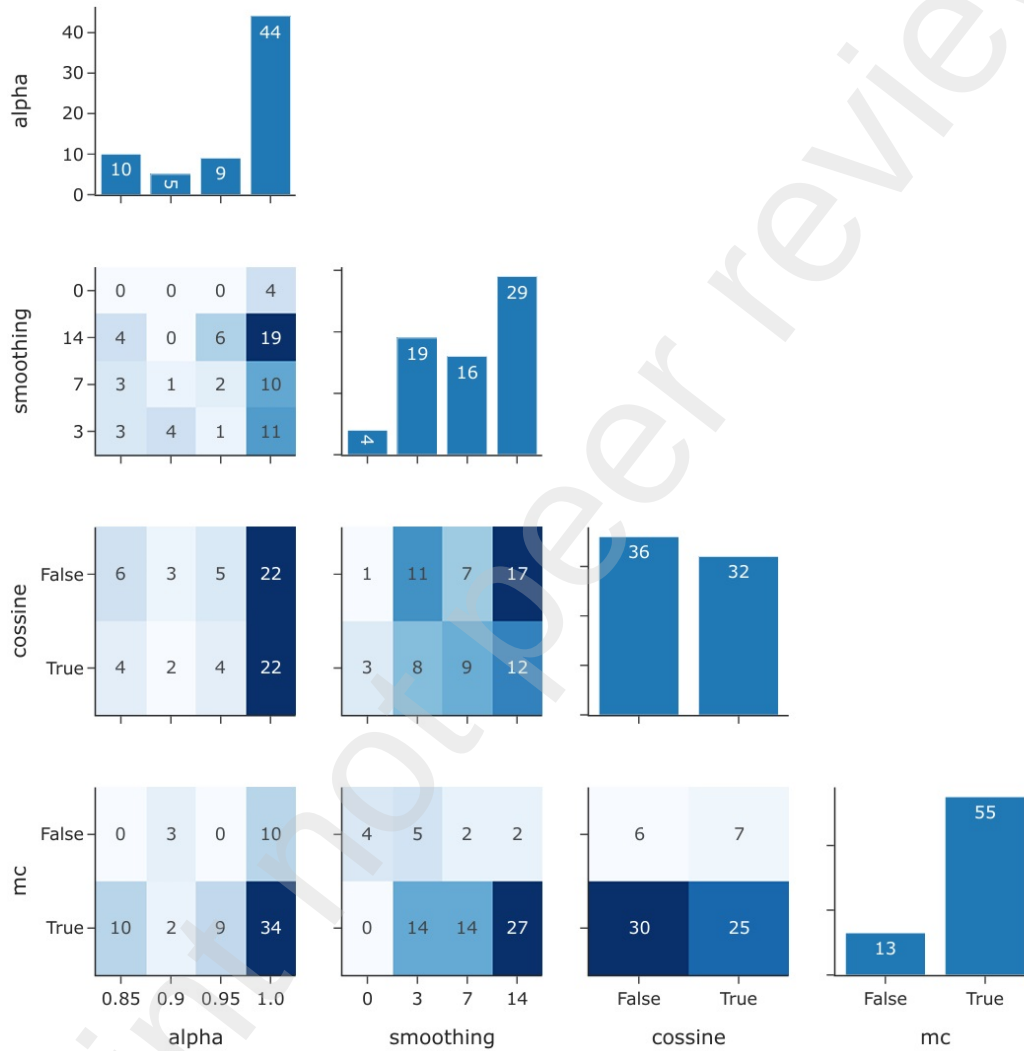


Figure 5: Count of stations for each selected hyperparameter combination, using right-tail weighted pinball loss. The bar charts on the diagonal present the count of stations for each hyperparameter individually, while the heat maps show the count for pairs of hyperparameters.

410 The model can simulate different start dates to identify the most suitable
 411 one. With project duration distributions, productivity losses due to weather
 412 can be anticipated. Quantiles can set contractual limits for expected delays;
 413 for example, the 95th quantile measures maximum expected delay from ad-
 414 verse weather. Post-project, the actual duration quantile can assess weather
 415 impact, aiding compensation negotiations with the contractor.

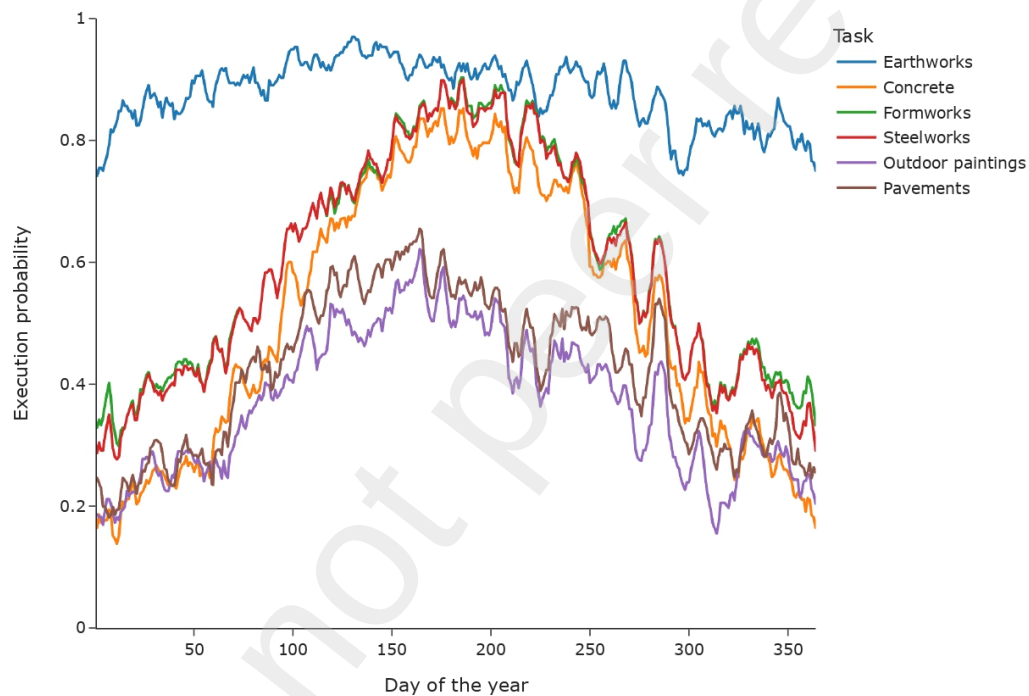


Figure 6: Probability of execution of each task throughout the days of the year, for station 113. Each row represents a specific task.

416 4. Conclusion

417 Construction delays, caused by factors like incompetence, lack of mate-
 418 rials, and adverse weather, negatively impact both owners and contractors,
 419 potentially triggering contractual clauses. This work proposes a method to
 420 model the probability of workable days and estimate project execution time

distributions considering weather conditions. The method uses loss measures to select the best model from a set of candidates, allowing for rigorous comparison of estimated distributions.

Experiments demonstrated the optimization of hyperparameters and showed that weather conditions vary monthly, influencing estimated distributions. The model can help choose the most appropriate start date by simulating different scenarios. It also allows for simulating task start dates, analyzing schedules, identifying critical stretches, and organizing activities to reduce costs.

During execution, the model can be updated with weather data to improve end date estimates and verify the impact of weather on the schedule. Comparing estimated distributions with actual project durations can help mediate negotiations between parties.

Climatic limits are considered to prevent task execution, while meteorological conditions reduce productivity. Tasks are assumed independent; if not, they can be modeled as combined tasks or through climatic limit probabilities. A Markov process can also be used to consider interactions between climatic variables.

The approach can incorporate non-climatic factors, treating planned tasks as random variables using PERT approaches. This includes expert knowledge to add uncertainties to planned durations, integrating climatic effects into simulations. Extensions can introduce resource constraints and other complexities.

This research has broader implications beyond construction, applicable to industries facing unpredictable challenges. The model's adaptability to various project scales can enhance strategic planning, execution, and risk management. In sum, it offers a significant tool for managing weather-related project delays.

5. Acknowledgments

Data for this study come from the Met Office Integrated Data Archive System (MIDAS) for land surface station data, which is archived at the Centre for Environmental Data Analysis for supplied (CEDA). This research was supported by FAPESP (grant 2022/08579-7) and CNPq (grants 422705/2021-7 and 305065/2023-8). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

6. Author contributions: CRediT

Mateus Borges Comito: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing - Review & Editing, Visualization, Writing – original draft. **Rafael Izbicki:** Conceptualization, Methodology, Investigation, Writing - Review & Editing, Writing – original draft. **Paulo do Canto Hubert Junior:** Conceptualization, Methodology, Investigation, Writing - Review & Editing, Writing – original draft. **Filipe Moura:** Writing - Review & Editing.

References

- [1] Durdyev, S., Hosseini, M. R. (2020). Causes of delays on construction projects: a comprehensive list. *International journal of managing projects in business*, 13(1), 20-46.
- [2] Sanni-Anibire, M. O., Mohamad Zin, R., Olatunji, S. O. (2022). Causes of delay in the global construction industry: a meta analytical review. *International Journal of Construction Management*, 22(8), 1395-1407.
- [3] Sepasgozar, S. M., Karimi, R., Shirowzhan, S., Mojtahedi, M., Ebrahimzadeh, S., McCarthy, D. (2019). Delay causes and emerging digital tools: A novel model of delay analysis, including integrated project delivery and PMBOK. *Buildings*, 9(9), 191.
- [4] Assaf, S. A., Al-Hejji, S. (2006). Causes of delay in large construction projects. *International journal of project management*, 24(4), 349-357.
- [5] Schuldt, S. J., Nicholson, M. R., Adams, Y. A., Delorit, J. D. (2021). Weather-related construction delays in a changing climate: a systematic state-of-the-art review. *Sustainability*, 13(5), 2861.
- [6] Apipattanavis, S., Sabol, K., Molenaar, K. R., Rajagopalan, B., Xi, Y., Blackard, B., Patil, S. (2010). Integrated framework for quantifying and predicting weather-related highway construction delays. *Journal of construction engineering and management*, 136(11), 1160-1168.
- [7] Thorpe, D., Karan, E. P. (2008, January). Method for calculating schedule delay considering weather conditions. In *Proceedings of the 24th annual conference of the Association of Researchers in Construction Management (ARCOM 2008)* (Vol. 2, pp. 809-818).

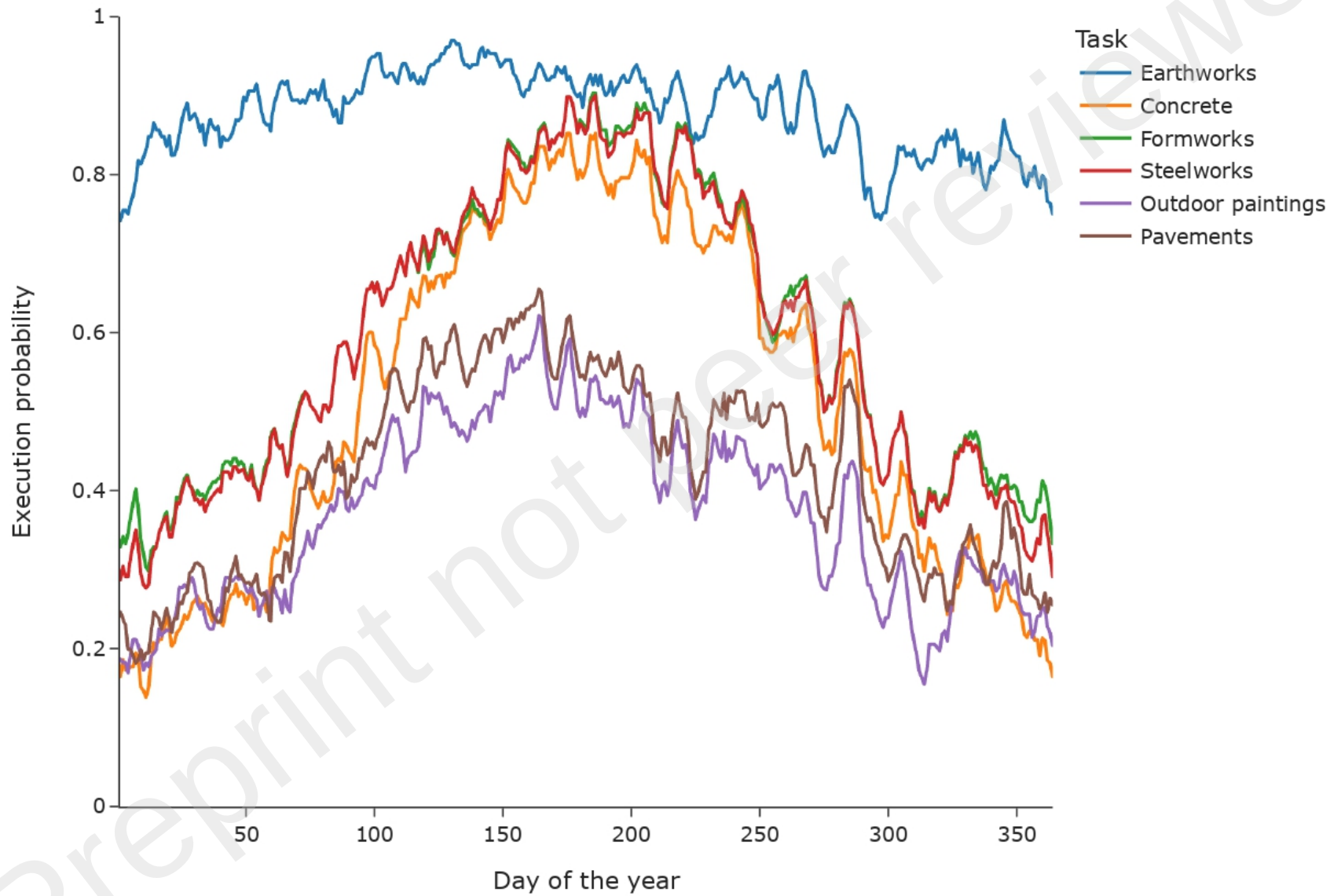
- 489 [8] Shahin, A., AbouRizk, S. M., Mohamed, Y. (2011). Modeling weather-
490 sensitive construction activity using simulation. *Journal of construction*
491 *engineering and management*, 137(3), 238-246.
- 492 [9] Richardson, C. W. (1981). Stochastic simulation of daily precipitation,
493 temperature, and solar radiation. *Water resources research*, 17(1), 182-
494 190.
- 495 [10] Golnaraghi, S., Zangenehmadar, Z., Moselhi, O., Alkass, S. (2019). Ap-
496 plication of artificial neural network (s) in predicting formwork labour
497 productivity. *Advances in Civil Engineering*, 2019(1), 5972620.
- 498 [11] Li, X., Chow, K. H., Zhu, Y., Lin, Y. (2016). Evaluating the impacts of
499 high-temperature outdoor working environments on construction labor
500 productivity in China: A case study of rebar workers. *Building and*
501 *environment*, 95, 42-52.
- 502 [12] Duffy, G., Woldeesenbet, A., Oberlender, G. D. (2012). Advanced linear
503 scheduling program with varying production rates for pipeline construc-
504 tion projects. *Automation in construction*, 27, 99-110.
- 505 [13] Gondia, A., Siam, A., El-Dakhakhni, W., Nassar, A. H. (2020). Machine
506 learning algorithms for construction projects delay risk prediction. *Jour-
507 nal of Construction Engineering and Management*, 146(1), 04019085.
- 508 [14] Jung, M., Park, M., Lee, H. S., Kim, H. (2016). Weather-delay simula-
509 tion model based on vertical weather profile for high-rise building con-
510 struction. *Journal of construction engineering and management*, 142(6),
511 04016007.
- 512 [15] Ballesteros-Pérez, P., Rojas-Céspedes, Y. A., Hughes, W., Kabiri, S.,
513 Pellicer, E., Mora-Melià, D., del Campo-Hitschfeld, M. L. (2017).
514 Weather-wise: A weather-aware planning tool for improving construc-
515 tion productivity and dealing with claims. *AUTOMATION IN CON-
516 STRUCTION*, 84, 81-95.
- 517 [16] Guo, S. J., Chen, J. H., Chiu, C. H. (2017). Fuzzy duration forecast
518 model for wind turbine construction project subject to the impact of
519 wind uncertainty. *Automation in Construction*, 81, 401-410.

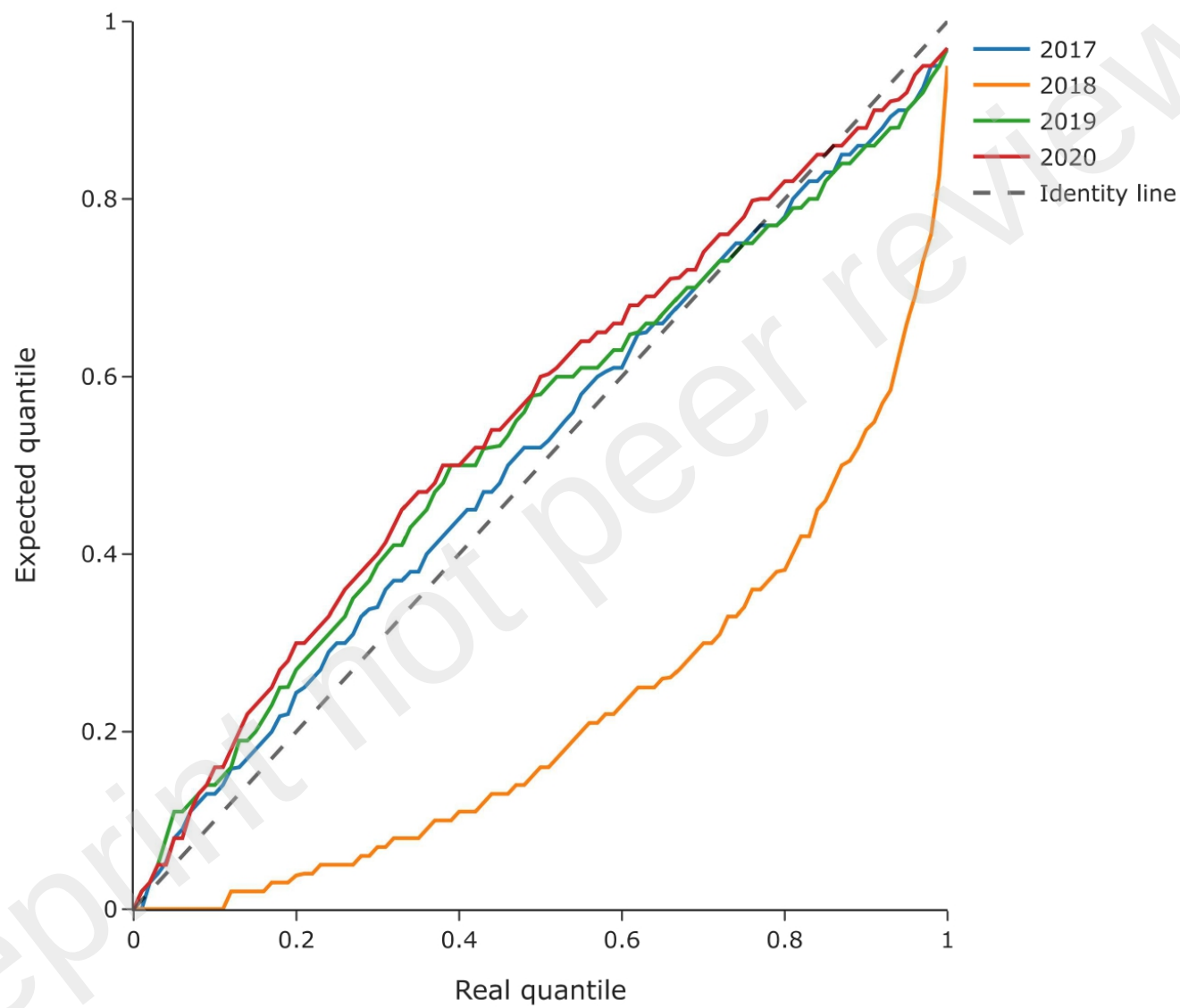
- [17] El-Rayes, K., Moselhi, O. (2001). Impact of rainfall on the productivity of highway construction. *Journal of construction engineering and management*, 127(2), 125-131.
- [18] Yaseen, Z. M., Ali, Z. H., Salih, S. Q., Al-Ansari, N. (2020). Prediction of risk delay in construction projects using a hybrid artificial intelligence model. *Sustainability*, 12(4), 1514.
- [19] Egwim, C. N., Alaka, H., Toriola-Coker, L. O., Balogun, H., Sunmola, F. (2021). Applied artificial intelligence for predicting construction projects delay. *Machine Learning with Applications*, 6, 100166.
- [20] Wauters, M., Vanhoucke, M. (2014). Support vector machine regression for project control forecasting. *Automation in Construction*, 47, 92-106.
- [21] Pewdum, W., Rujirayanyong, T., Sooksatra, V. (2009). Forecasting final budget and duration of highway construction projects. *Engineering, Construction and Architectural Management*, 16(6), 544-557.
- [22] Chakraborty, D., Elhegazy, H., Elzarka, H., Gutierrez, L. (2020). A novel construction cost prediction model using hybrid natural and light gradient boosting. *Advanced Engineering Informatics*, 46, 101201.
- [23] Van Slyke, R. M. (1963). Monte Carlo methods and the PERT problem. *Operations Research*, 11(5), 839-860.
- [24] Rezaie, K., Amalnik, M. S., Gereie, A., Ostadi, B., Shakhseeniaee, M. (2007). Using extended Monte Carlo simulation method for the improvement of risk management: Consideration of relationships between uncertainties. *Applied Mathematics and Computation*, 190(2), 1492-1501.
- [25] Sadeghi, N., Fayek, A. R., Pedrycz, W. (2010). Fuzzy Monte Carlo simulation and risk assessment in construction. *Computer-Aided Civil and Infrastructure Engineering*, 25(4), 238-252.
- [26] Kokkaew, N., Wipulanusat, W. (2014). Completion delay risk management: A dynamic risk insurance approach. *KSCE Journal of Civil Engineering*, 18, 1599-1608.
- [27] Ballesteros-Pérez, P., Rojas-Céspedes, Y. A., Hughes, W., Kabiri, S., Pellicer, E., Mora-Melià, D., del Campo-Hitschfeld, M. L. (2017).

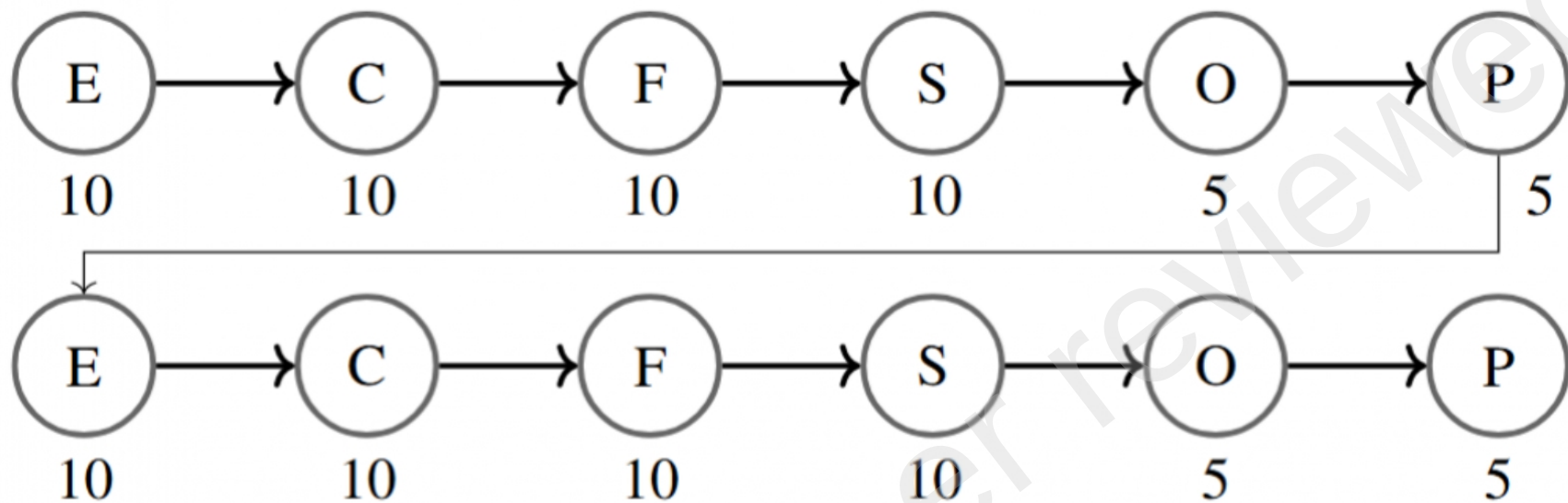
- 551 Weather-wise: A weather-aware planning tool for improving construc-
552 tion productivity and dealing with claims. *Automation in construction*,
553 84, 81-95.
- 554 [28] Hastie, T., Tibshirani, R., Friedman, J. (2017). The elements of statis-
555 tical learning: data mining, inference, and prediction.
- 556 [29] Hanoon, M. S., Ahmed, A. N., Zaini, N. A., Razzaq, A., Kumar, P.,
557 Sherif, M., El-Shafie, A. (2021). Developing machine learning algorithms
558 for meteorological temperature and humidity forecasting at Terengganu
559 state in Malaysia. *Scientific Reports*, 11(1), 18935.
- 560 [30] Heravi, G., Eslamdoost, E. (2015). Applying artificial neural networks
561 for measuring and predicting construction-labor productivity. *Journal*
562 *of Construction Engineering and Management*, 141(10), 04015032.
- 563 [31] Song, L., AbouRizk, S. M. (2008). Measuring and modeling labor pro-
564 ductivity using historical data. *Journal of construction engineering and*
565 *management*, 134(10), 786-794.
- 566 [32] Chinnadurai, J., Venugopal, V., P, K. (2016). Influence of occupational
567 heat stress on labour productivity—a case study from Chennai, India.
568 *International Journal of Productivity and Performance Management*,
569 65(2), 245-255.
- 570 [33] Ibbs, W., Sun, X. (2017). Weather's effect on construction labor pro-
571 ductivity. *Journal of legal affairs and dispute resolution in engineering*
572 *and construction*, 9(2), 04517002.
- 573 [34] Ballesteros-Pérez, P., Smith, S. T., Lloyd-Papworth, J. G., Cooke, P.
574 (2018). Incorporating the effect of weather in construction scheduling
575 and management with sine wave curves: application in the United King-
576 dom. *Construction management and economics*, 36(12), 666-682.
- 577 [35] Koenker, R., Hallock, K. F. (2001). Quantile regression. *Journal of eco-*
578 *nomic perspectives*, 15(4), 143-156.
- 579 [36] Steinwart, I., Christmann, A. (2011). Estimating conditional quantiles
580 with the help of the pinball loss.

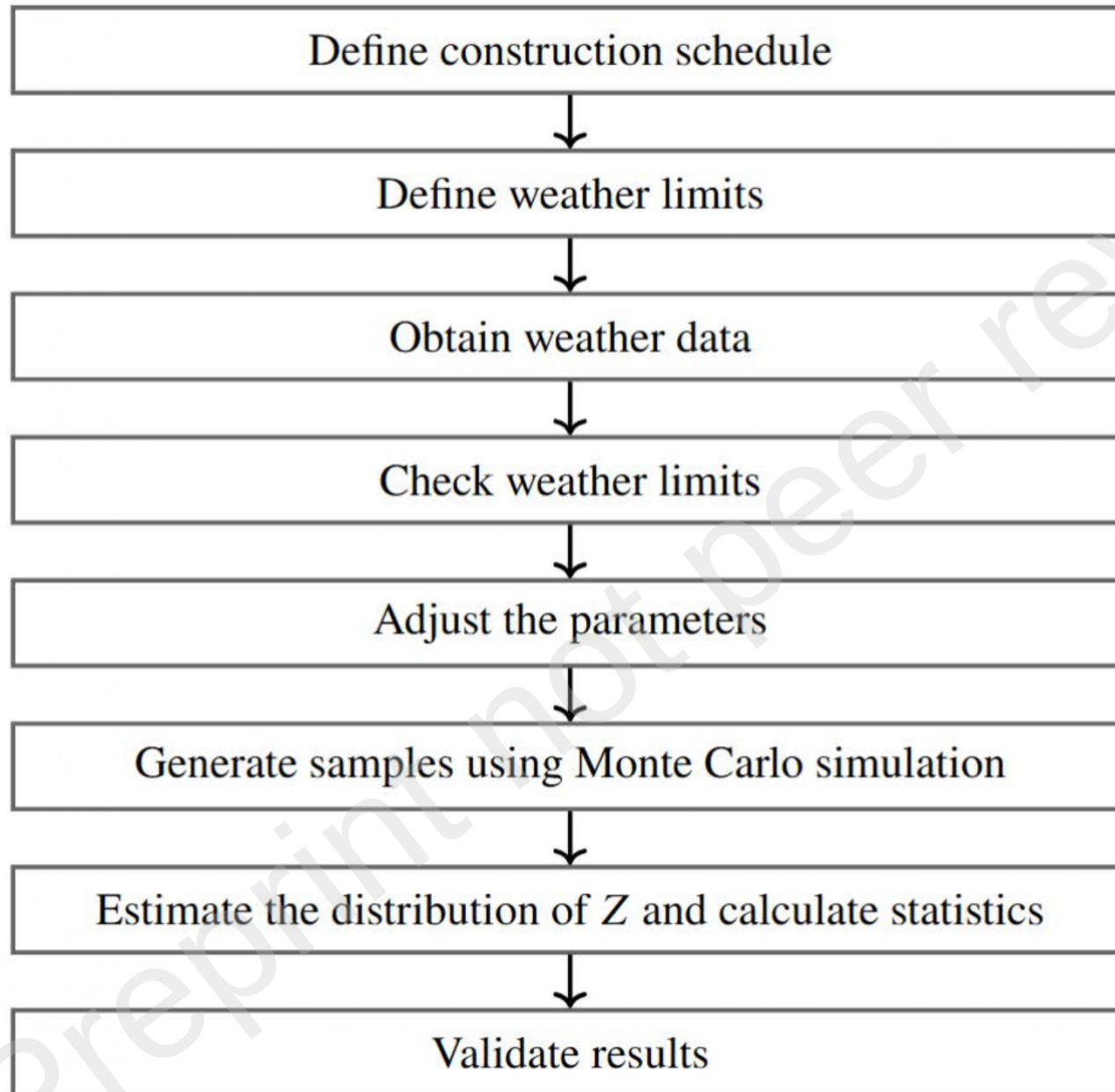
- [37] Izbicki, R., B. Lee, A. (2017). Converting high-dimensional regression to high-dimensional conditional density estimation.
- [38] Freeman, P. E., Izbicki, R., Lee, A. B. (2017). A unified framework for constructing, tuning and assessing photometric redshift density estimates in a selection bias setting. *Monthly Notices of the Royal Astronomical Society*, 468(4), 4556-4565.
- [39] Izbicki, R., Lee, A. B., Freeman, P. E. (2017). Photo-z estimation: An example of nonparametric conditional density estimation under selection bias.
- [40] Dalmaso, N., Pospisil, T., Lee, A. B., Izbicki, R., Freeman, P. E., Malz, A. I. (2020). Conditional density estimation tools in python and R with applications to photometric redshifts and likelihood-free cosmological inference. *Astronomy and Computing*, 30, 100362.
- [41] Nadaraya, E. A. (1965). On non-parametric estimates of density functions and regression curves. *Theory of Probability & Its Applications*, 10(1), 186-190.
- [42] Rajagopalan, B., Lall, U. (1995). A kernel estimator for discrete distributions. *Journal of Nonparametric Statistics*, 4(4), 409-426.
- [43] Guide, P. M. B. O. K. (2008). *A guide to the project management body of knowledge*.
- [44] Met Office. (2019). Met Office MIDAS open: UK land surface stations data (1853-current). *Centre for Environmental Data Analysis*, 2024.
- [45] Kendon, M., McCarthy, M., Jevrejeva, S., Matthews, A., Sparks, T., Garforth, J. (2021). State of the UK Climate 2020. *International Journal of Climatology*, 41, 1-76.
- [46] Kendon, M., McCarthy, M., Jevrejeva, S., Matthews, A., Legg, T. (2019). State of the UK climate 2018. *International Journal of Climatology*, 39(Suppl 1), 1-55.
- [47] Apipattanavis, S., Podestá, G., Rajagopalan, B., Katz, R. W. (2007). A semiparametric multivariate and multisite weather generator. *Water Resources Research*, 43(11).

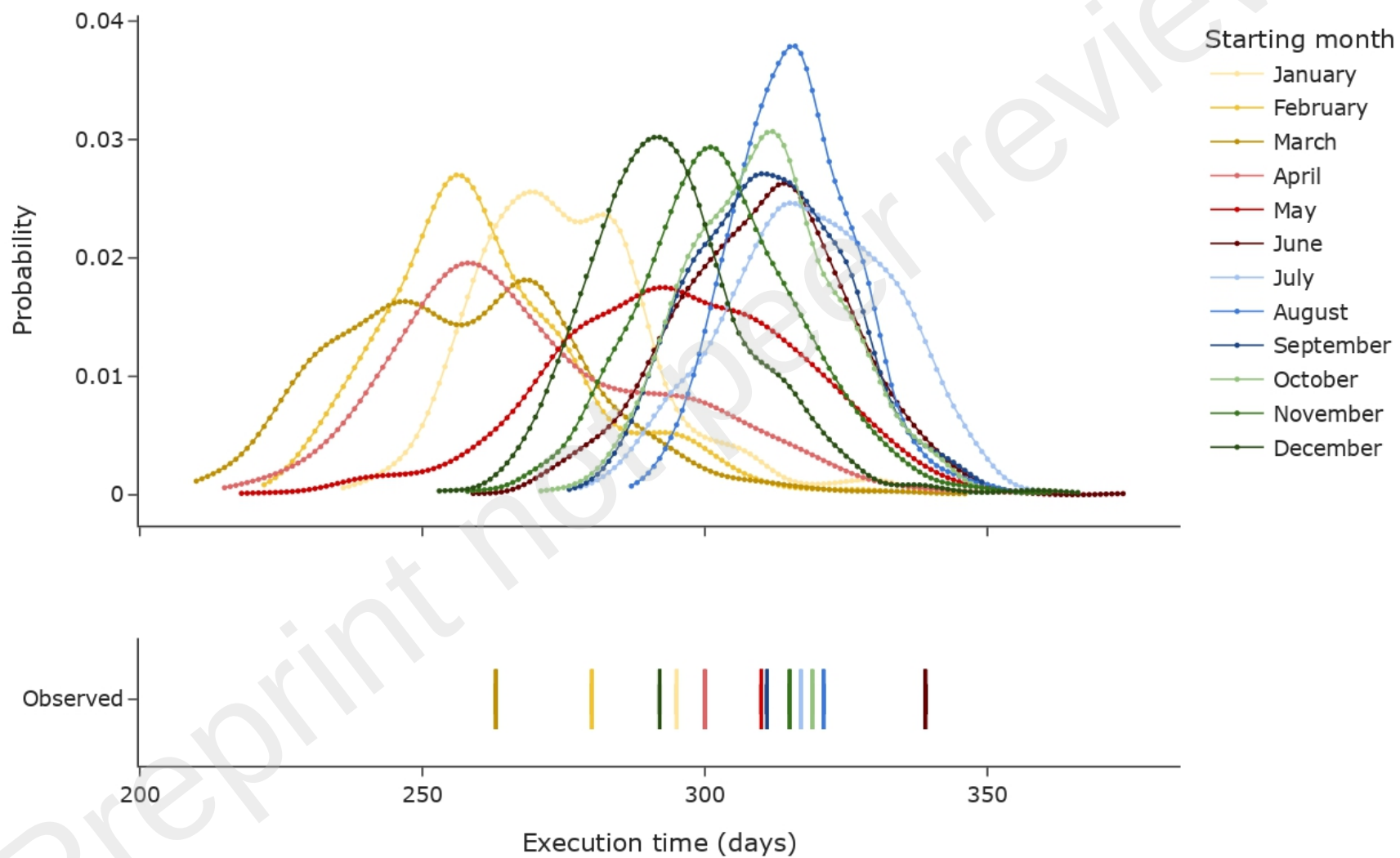
- 612 [48] Borges, M., Izbicki, R., HUBERT, P. (2024). Nonparamet-
613 ric Modeling of Weather-Induced Delays [Computer software].
614 [https://github.com/mborgesss/nonparametric-modeling-weather-](https://github.com/mborgesss/nonparametric-modeling-weather-delays)
615 [delays](https://github.com/mborgesss/nonparametric-modeling-weather-delays)

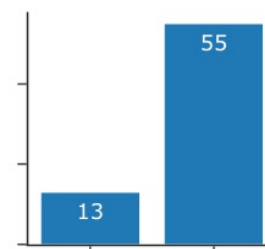
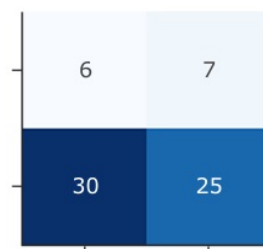
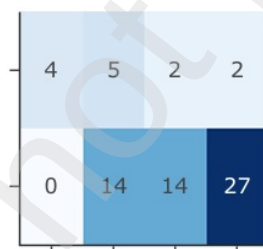
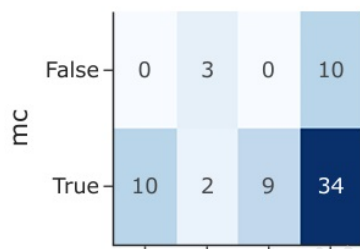
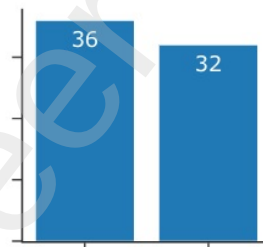
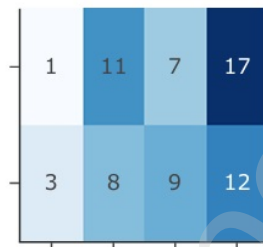
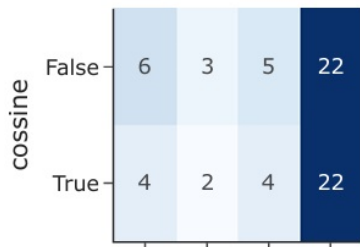
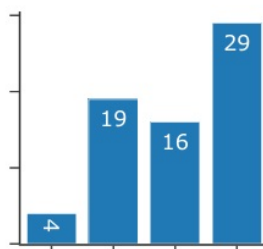
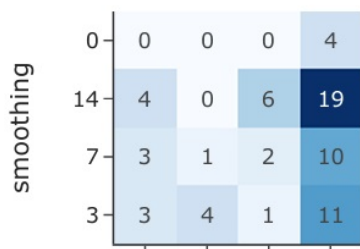
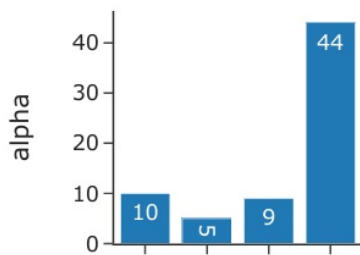












0.85 0.9 0.95 1.0

0 3 7 14

False True

False True

alpha

smoothing

cossine

mc