TIME-SERIES FORECASTING OF OCCUPATIONAL INJURY RISK

GIUSEPPE COCOMAZZI

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

547014

COMMITTEE

dr. Stijn Rotman
dr. Travis Wiltshire

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

January 11, 2026

WORD COUNT

8475

ACKNOWLEDGMENTS

I would like to thank the previous students in the program, whose theses were a great source of inspiration while writing this work.
My wife, Elisabeth, helped me train on real-life 'family spikes', while my daughter, Emilia, continually tested my ability to capture unpredictable behavior.
Wienand delivered an impromptu Sunday data science masterclass, reminding me that machine learning models ultimately have to be deployed.
Finally, my supervisor, Stijn Rotman, provided insightful feedback following the first submission.

TIME-SERIES FORECASTING OF OCCUPATIONAL INJURY RISK

GIUSEPPE COCOMAZZI

Abstract

Forecasting the temporal evolution of severe occupational injuries can potentially provide valuable insights for safety management stakeholders to adopt proactive measures and allocate preventive resources more effectively. While previous work in the occupational safety literature has mainly focused on descriptive analyses and classification tasks, this thesis addresses this gap by framing severe injury risk as a multi-step time series forecasting problem. The analysis is conducted using the Occupational Safety and Health Administration's Severe Injury Report (SIR) dataset, which is structured as panel data, raising the methodological question of how to best pool information across grouping factors.

Therefore, we investigate how the forecasting performance of linear models (Ridge, Lasso, Elastic Net, and Partial Least Squares) and tree-based ensemble models (CatBoost, XGBoost, and LightGBM) varies across two temporal aggregations (monthly and weekly), multiple forecast horizons, and the inclusion of exogenous features.

The results show that all machine learning models consistently outperform seasonal and statistical baselines, with similar forecasting performance across model families. Forecast accuracy, as measured by RMSE, MAE, and MASE, is largely driven by autoregressive features rather than by model complexity. Pooling information across states does not substantially improve average forecasting accuracy but reduces forecast variability.

1 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

The OSHA's SIR dataset is publicly available at Severe Injury Dashboard | Occupational Safety and Health Administration. The data is anonymised. Work on this thesis did not involve collecting data from human participants or animals. The original owner of the data and code used in this thesis retains ownership of the data and code during and after the completion of this thesis.

All the figures belong to the author. In terms of writing, a generative language model, OpenAI's ChatGPT 5.1, was used to improve the author's original content, for paraphrasing, spell checking and grammar, and help debug LaTex syntax. The same model assisted in writing and documenting the code, included the function to fetch data from the U.S. Bureau of Statistics. Other typesetting tools or services are listed in Appendix A.

2 PROBLEM STATEMENT & RESEARCH GOAL

2.1 *Context*

Occupational safety and health (OSH) organizations across the world help foster the well-being of workers by preventing workplace accidents, promoting safe practices, and enforcing regulatory standards. Employers are legally responsible for ensuring a healthy working environment, as employees may be exposed to a wide range of physical, biological, and psychosocial hazards (Howe et al., 2024; Vassiley, Barratt, Dayaram, & Burgess, 2025; Vitrano & Micheli, 2024).

According to the International Labour Organization (ILO), an estimated 364 million workers globally suffer non-fatal occupational accidents each year, with substantial geographical disparities in incidence rates. Injured workers often face long-term health complications, reduced employability, and financial hardship.

Beyond its societal implications, workplace safety has a considerable economic dimension. It is widely recognized that improving workplace safety can reduce economic losses at both firm and national levels. ILO estimates suggest that almost an annual 4% loss in global Gross Domestic Product is accountable to occupational accidents and diseases (Organization, 2020).

Traditional prevention measures to reduce occupational injury risk include compliance with safety standards and regulatory frameworks; inspections and audits; environmental controls and maintenance schedules; safety education programs, and post-incident analysis. Despite the efforts, manual reporting increasingly struggles to keep up with both rapidly evolving working conditions and the increasing volume of historical safety data. Post-hoc interventions are no longer sufficient for effective prevention.

For these reasons, real-time responsiveness has become crucial to further mitigate workplace injury risks. The integration of the Internet of Things technologies into the workplace allows for a continuous streaming of data coming from sensors, wearables, and smart equipment, supporting constant monitoring of safety conditions and triggering timely preventive actions (Michaels & Wagner, 2025).

From a scientific point of view, this complex data flow may pose a challenge to traditional time series methods such as Error, Trend, Seasonal models (ETS), autoregressive integrated moving average (ARIMA) and its seasonal extension (SARIMA), which have long served as standard baselines for forecasting tasks in the public-health domain. These models provide statistically grounded forecasts that often perform well in settings with clear seasonal structure and relatively smooth temporal dynamics (Spiliotis, 2023). On the other hand, Machine learning (ML) methods can leverage rich and dynamical data sources to capture complex, non-linear patterns, thereby marking a shift from a reactive to a proactive safety management strategy (Kontopoulou, Panagopoulos, Kakkos, & Matsopoulos, 2023; Masini, Medeiros, & Mendes, 2023). Because the temporal factor will be the main leading axis in our analysis, we will first explore the data to extract components such as trend and seasonality to guide our model selection.

While most studies have focused on classification tasks, the area of time series forecasting for occupational injury prevention remains still underexplored, despite its potential to anticipate future risk patterns and support more targeted interventions.

2.2 *Research strategy*

As outlined in subsection 2.1, knowing *when* an intervention may be required is crucial for increasing the likelihood of alleviating workers' exposure to severe injuries. If meaningful patterns in historical data can be reliably captured by forecasting algorithms, stakeholders can make more informed decisions about how to prioritize preventive resources, and about which types of data need continuous monitoring. As different temporal granularities can impact the accuracy of the generated forecasts (Rostami-Tabar, Goltsos, & Wang, 2023), the first question we aim to explore is to what extent different levels of temporal resolution can help smooth out an intermitting time series at the original lead time. Moreover, this question gains a further dimension when the data is organised according to a grouping factor, such as having time series across multiple states (Qu, Timmermann, & Zhu, 2024). The model choice will additionally take into account the explainability of the models.

RQ1: *How does the forecasting performance of a set of linear and tree-based ensemble models vary at two temporal aggregations (monthly and weekly) in multi-step forecasting of severe injuries risk, as evaluated by RMSE, MAE and MASE?*

Weekly aggregation increases the number of observations but it may introduce noise and variability, while monthly aggregation may produce coarser predictions. Understanding this trade-off is essential for designing reliable predictive pipelines (Barhrhouj, Ananou, & Ouladsine, 2025).

Given the absence of a comparable supervised regression setting in prior work on occupational injury forecasting, we will train two different families of regressors based on the exploratory initial step and on the main findings from related datasets. The choice for linear models with penalty terms (Ridge, Lasso, Elastic Net) allows us to assess how regularization can affect feature selection in a strongly seasonal data structure across a balanced panel (Han, Zhou, Sun, & Li, 2025). Among tree-based ensemble models, CatBoost is deemed particularly suited for our panel data, as it can handle categorical features (Hall & Rasheed, 2025). XGBoost and LightGBM appear to consistently be the most accurate models in recent forecasting benchmarks and machine-learning competitions (Makridakis, Spiliotis, & Assimakopoulos, 2022).

Evaluation metrics will be compared against a naive seasonal baseline and a stronger statistical model, namely the additive Holt-Winters exponential smoothing method (ETS). As stated by Makridakis, Spiliotis, and Assimakopoulos (2018), this classical statistical method still shows remarkable accuracy.

Finally, the sample size and a low-dimensional feature matrix allowed us to compare a wide variety of models with negligible computational cost.

SQ1.1: How does the inclusion of exogenous categorical variables affect the forecasts of the ML models?

We extend the initial setup to include three exogenous features, aptly aggregated to fit our forecasting task. The purpose of this subquestion is to assess whether the accuracy of the models is mostly driven by autoregressive features or whether selected exogenous variables can account for part of the variability in the forecasts.

RQ2: *How does the predictive accuracy change when estimated as a Global Forecasting Model (across all states jointly) rather than as a Local Forecasting Model (state-specific), for both monthly and weekly temporal resolutions?*

With this question we evaluate whether pooling information across states enhances the predictive performance relative to modeling each state independently. A global model may exploit shared temporal knowledge and increase the available sample size, while a local model may better

capture state-specific patterns. This question is closely related to the literature on panel data discussed in Section 3.3.

For example, some industry sectors may explain better a given amount of the variability in the distribution of hospitalizations, as severe injuries occur more often in e.g. construction sites. To answer this question, we will explore feature permutation to quantify how much each feature contributes to the predictive power of the model.

2.3 *Societal Relevance*

OSH agencies must strategically allocate their resources to maximize injury reduction. Severe injuries, in particular, can pose an even higher burden for both employees and the society at large. Forecasting the relative risk of these life-altering events in the area where they will most likely occur can provide actionable insights to guide OSH agencies on how to better target their interventions.

Recent evidence underscores the potential of such data-driven targeting strategies. As Johnson, Levine, and Toffel (2023) suggest, a machine learning approach to prioritize inspections could have averted twice as many injuries compared to the current OSHA prevention programs. Building on this premise, the present study focuses on forecasting relative risk trajectories across U.S. states to support more informed and proactive resource deployment.

2.4 *Scientific Relevance*

A debated topic in the forecasting literature concerns to what extent global models – trained on pooled cross-sectional time-series – can outperform local models, which are trained independently for each unit (e.g., state, firm, or region). We systematically test whether a global pooling strategy improves forecasting accuracy compared to a local strategy when applied to a real-world panel dataset. Multiple ML regressors will be trained and compared under both frameworks, providing empirical evidence on the benefits and limitations, if any, of pooling information across states. Additionally, by framing the problem as a supervised regression task, we incorporate lagged features, rolling statistics, and exogenous variables, in line with the best practice in time series forecasting with ML regressors (Bojer, 2022; Makridakis et al., 2022).

To the best of our knowledge, this is the first study to comparatively assess global versus local forecasting strategies for occupational injury risk prediction. More specifically, although time series forecasting is well established in adjacent health-related domains such as epidemiology and

biostatistics, the regression-based framing of our study extend methodological practices common in these scientific fields to occupational safety research (Reich et al., 2019).

3 RELATED WORK

In this section, we present the relevant literature with a particular focus on studies that inform the design choice. To structure the review, we use the CRISP-DM framework as a conceptual reference. The CRISP-DM framework is a widely adopted process methodology that illustrates the main stages in the data mining workflow (Kurgan & Musilek, 2006).

To make the review more readable, Table 1 maps the most relevant studies to each step in the CRISP-DM framework and to the generated design choice specific to this study.

3.1 *Occupation Injury Risk Prediction*

The field of occupation injury risk prediction using ML models has been largely dominated by classification tasks. In their comprehensive survey, Vivian et al. (2025) identified three main research directions: (i) identifying workplace injury risk factors, (ii) predicting return-to-work rates, and (iii) analyzing sociodemographic features across different worker populations. Most studies rely on business-level datasets or focus on specific injury types such as traumatic brain injuries (Van Deynse et al., 2023) or foot and ankle injuries.

The ML models most frequently used in this field are selected for their ability to capture patterns in heterogeneous sets of categorical variables. While no single model clearly outperforms the others, Random Forest, Support Vector Machines, Decision Trees, and Linear Regression have achieved the best results across evaluation metrics such as accuracy, sensitivity, and F-1 score, underscoring the importance of tailoring model selection to data characteristics (Vivian et al., 2025; Yuan, Varathan, Suhaimi, & Ling, 2023). However, regression-based approaches that model the temporal evolution of occupational risk remain virtually absent from the literature.

The work most closely aligned with our research is Cerqua, Giannantoni, et al. (2024). The authors develop a supervised regression framing for the time-series forecasting of workplace fatalities in Italy, focusing on spatial heterogeneity and ex-ante allocation of regulatory interventions. The methodology explicitly addresses how to apply machine learning pipelines to panel data, implementing a rolling forecasting origin approach for hyperparameter tuning, and evaluating model performance on a one-year hold-out test set. They selected a mix of models: linear, tree-based,

Table 1: Mapping data processing stages to related literature and design choices.

Data process	Related work	Design choice
Business understanding	Vivian, Bauder, and Khoshgoftaar (2025); Cerqua, Giannantoni, Letta, and Pinto (2024)	Identify research gap; frame as regression task; define forecasting objective.
Data understanding	Gomes, Parasram, Collins, and Socias-Morales (2023); Williams and Marc (2024); Coulombe, Marcellino, and Stevanovic (2025); Montero-Manso and Hyndman (2021); Rostami-Tabar et al. (2023)	Model seasonality; normalize the target variable; treat the dataset as panel data; aggregation strategy.
Data Preparation	Taieb, Bontempi, Atiya, and Sorjamaa (2011)	Multi-step forecasting strategy.
Modeling	Makridakis et al. (2022); Khairuddin et al. (2022)	Identify state of the art; select relevant exogenous features.
Evaluation	Hewamalage, Bergmeir, and Bandara (2022); Cerqueira, Torgo, and Mozetič (2020)	Select RMSE, MAE and MASE; choose out-of-sample approach.
Deployment	Michaels and Wagner (2025)	Inform discussion section.

and the Long Short-Term Memory neural network, ultimately identifying Partial Least Squares (PLS) as the best performing model across different sets of predictors.

3.2 *Studies on OSHA's SIR dataset*

Data mining applications on the OSHA's SIR dataset are limited to four studies.

On the descriptive side, a notable study examining the temporal aspects of occupational injury data is Gomes et al. (2023). The authors analysed the

Severe Injury Report (SIR) dataset provided by the Occupational Safety and Health Administration (OSHA), using time series decomposition models to uncover seasonal effects and general trends. Moreover, they evaluated industry and injury-type breakdowns, revealing that manufacturing and construction are the most affected industries.

Importantly, they identified systematic temporal patterns in the occupational injuries, such as the predominance of injuries on weekdays compared to weekends, and a higher injury incidence in the summer months. However, Williams and Marc (2024) did not find a statistically significant difference between cold and warm seasons ($p = 0.09$). Their study introduced population adjustments for normalizing injury counts by industry employment size, using 2-digit NAICS codes to compute injuries per 100,000 workers.

We adopt a similar adjustment for the number of hospitalization, accounting only for state workforce size. In this way, we predict the incidence of new hospitalized employees over a specific time and at the stale level. While both studies provided valuable descriptive insights into the temporal features of the SIR data, they did not extend to predictive or regression-based modeling.

On the classification side, Khairuddin et al. (2022) trained five machine learning models on the SIR dataset to predict hospitalization and amputation outcomes, finding that Random Forest achieved the highest performance on metrics such as accuracy and F1-score. After inspecting the feature importance, they found Nature of Injury and Type of Event to be the two most relevant variables. Based on this finding, we encode those two variables to fit our regression problem (see Section 4.4).

Building on this work, Hasan Khalleefah Hassan and Khalifa (2025) evaluated a different set of machine learning models, and reported that the AdaBoost Classifier showed the best performance on accuracy and F1-score.

Both studies exploit the industry-wide properties of the OSHA's SIR dataset to develop a model that may better capture patterns across heterogeneous industries. We follow their approach with the supplementary motivation that increasing the observations across state-level time series may benefit a global approach to our regression task. Similarly, we employ a comparative approach by evaluating different models. However, as both studies aimed to categorize past events, their results and model selection can not serve as a benchmark for our work, but only as a point of reference.

3.3 *Data Understanding*

When observations are collected over time for multiple entities—such as states, firms, or individuals—the resulting structure is referred to as *time-series cross-sectional data*, or more simply, *panel data*. Since the SIR dataset includes reports from multiple U.S. states across the same time period, a crucial question concerns how to model panel data, in particular the time-invariant components. In their theoretical review, Baltagi (2013) summarize several studies in which macroeconomic panel data are modeled using fixed-effects, mixed-effects, or random-effects specifications to forecast continuous outcomes at multiple horizons, with performance evaluated on RMSE.

In a *fixed-effects* model, each unit is assigned a fixed intercept that captures all characteristics that do not vary over time. In this framework, unobserved heterogeneity across units is not explicitly modeled; in other words, the effects of time-invariant variables—such as structural differences between states—are not estimated. Conversely, a *mixed-effects* (or *random-effects*) model treats unit-specific effects as random variables drawn from a common normal distribution. This allows the model to estimate the variance of these effects, thereby capturing how the time-invariant characteristics of one unit relate to other units. Both approaches can be integrated within a *Hierarchical Linear Modeling* (HLM) framework, where fixed effects represent the average impact of predictors across the entire population, and random effects account for unit-specific deviations from this overall trend (Baltagi, 2013; Chen, 2021).

In this study, monthly relative risks are aggregated across all states to estimate the overall risk of hospitalization, while still allowing for state-level deviations by computing risks relative to each state's workforce rather than the national total. Although hierarchical modeling provides a more flexible framework, explicitly modeling between-state variation falls beyond the scope of the present research.

Exploratory analysis of the SIR dataset reveals substantial heterogeneity in hospitalization risk across OSHA states, with smaller states exhibiting higher volatility. In such settings, model choice and the treatment of cross-sectional information become critical. Coulombe et al. (2025) show that, for a panel of U.S. state-level fiscal variables, forecasting performance improves when flexible models are combined with cross-sectional pooling, even in the presence of pronounced heterogeneity. Their study compares a range of machine learning models—including Ridge, Lasso, Sparse-Group Lasso, Random Forests, Boosted Trees, and Neural Networks—under alternative panel structures, namely no pooling, local estimation, global pooling, and clustered pooling. The results indicate that nonlinear models are especially

effective at forecasting volatile outcomes and at capturing cross-sectional nonlinearities, while pooling substantially stabilizes predictions relative to purely local approaches.

We adopt a similar modeling perspective. However, because our application does not involve mixed-frequency predictors and relies on a more limited set of covariates, we retain regularized linear models alongside nonlinear alternatives. In this context, cross-sectional pooling is expected to benefit linear models as well, by reducing estimation variance and improving generalization across states.

The question whether to implement a global as opposed to a local forecasting model has been extensively addressed by Montero-Manso and Hyndman (2021). In their study, they found that global methods show good results not only in groups of similar time series, but even in heterogeneous groups. When a large number of time series is available, global models may generalize better with fewer parameters. Importantly, global models can afford much larger memory in terms of lags, being able to capture long-memory patterns that local models could learn only if manually engineered. We empirically test these findings by training different models locally – fitting a single function to each cross-sectional time-series – and globally – by pooling all state-level series into a single learning problem.

As shown by Wilms, Cupelli, and Monti (2018), incorporating exogenous variables can particularly benefit forecasting settings where nonlinear interactions are present. While the SIR dataset includes several categorical features, many are either not relevant to the forecasting task or must be treated as compositional data. In particular, the variables Nature and Type of Event can not be fed into the models as they are, but they must be transformed into numerical values that respect the compositional constraint that category shares sum to one.

A further design choice involves the aggregation of observations that are collected at a higher frequency than that used for modeling. As the exploratory data analysis (EDA) shows, maintaining the daily resolution of the SIR dataset results in a too sparse feature space.

As argued by Rostami-Tabar et al. (2023), aggregation choices are often forced by data availability, and the temporal resolution of the data frequently aligns with the resolution at which forecasts are required. In this study, we evaluate both monthly and weekly aggregations, and later justify why a monthly resolution may be preferable for deployment.

A more specific design choice is whether to aggregate in such a way that data are divided into separate buckets (non-overlapping approach) or to create a temporal aggregation by sliding a moving window that equals the aggregation level (overlapping approach). We adopt a non-overlapping

approach, as it simplifies the handling of compositional features during data preparation.

3.4 *Data Preparation*

When preparing the data for forecasting, an important consideration is to choose a forecasting strategy. In their seminal review, Taieb et al. (2011) formalized five strategies commonly used for multi-step forecasting: (i) recursive, (ii) direct, (iii) direct-recursive hybrids, (iv) multi-input multi-output, and (v) direct multi-output.

Under the recursive strategy, a model generates a one-step ahead forecast which is fed back to the historical window, and used to recompute all the features, before the model can generate a new one-step ahead forecast. This process is repeated until the final time point in the horizon is reached. under the direct approach, a separate model is independently trained to predict a single output for each horizon. For instance, forecasting three horizons (e.g., $h = 1, 3, 6$) requires fitting three separate models, one per horizon. Unlike the recursive strategy, the direct approach does not use forecasts as inputs in the historical window. As a consequence, errors do not compound at each next step, overcoming the risk of bias accumulation over longer horizons.

The most relevant strategy for answering our research questions is the direct strategy. Because our models are built under different assumptions - linear, tree-based, and artificial networks - the direct approach allows for comparability by deploying the same pipeline under the same evaluation paradigm.

3.5 *Modeling*

As outlined in the Context section, a major shift from statistical to machine learning models has led to the adoption of machine learning algorithms in recent competitions on time series forecasts. In particular, Makridakis et al. (2022) reviewed the winning methods for the M5 competition, noting that the many of the best submissions use gradient boosted trees as implemented in LightGBM for its ability to handle multiple features of various types. Another tree-based ensemble method, XGBoost, is regarded as state of the art algorithm in time series forecasting (Fang, Yang, Lv, An, & Wu, 2022; Obasi, Cheng, Varianou-Mikellidou, Dimopoulos, & Boustras, 2026).

Models based on neural networks, such as Long Short-Term Memory (LSTM) architectures, and ensemble forecasting approaches are increasingly common in the recent forecasting literature. These methods are particularly effective in settings characterized by high-dimensional inputs and large

training samples, where complex nonlinear dynamics and long-range temporal dependencies can be learned directly from the data (Spiliotis, 2023).

In the context of the SIR dataset, the available time series are relatively short and show heterogeneous dynamics across states, especially for smaller states with sparse observations. Under such conditions, highly parameterized neural network models risk overfitting and may offer limited gains over simpler approaches. We confirmed the exclusion of approaches based on Recurrent Neural Network after evaluating the results from the experiments conducted to answer Research Question 1.

3.6 *Evaluation*

As discussed in Section 3.3, the most commonly used evaluation metrics in time-series forecasting are the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE), followed by scale-dependent alternatives such as the Mean Absolute Percentage Error (MAPE) and the Mean Absolute Scaled Error (MASE). These metrics are widely adopted in both statistical and machine-learning–based forecasting studies due to their interpretability and robustness across a broad range of applications (Hewamalage et al., 2022).

Because our models are evaluated at two different temporal resolutions, we rely on MASE to facilitate meaningful performance comparisons across frequencies. MASE scales the mean absolute forecast error by the in-sample mean absolute error of a naive (typically seasonal) benchmark (Hyndman & Koehler, 2006). Values of MASE below one indicate that a model outperforms the seasonal naive forecast, while values above one imply inferior performance. This property makes MASE suitable for panel and multi-frequency forecasting settings, where scale differences and aggregation effects can otherwise confound error comparisons.

To test for out-of-sample generalizability in time series, appropriate handling of the panel structure during the temporal split of the data into a train and a test set can help prevent a wide range of potential data leakage (Cerqua, Letta, & Pinto, 2024). Because in a time series observations are temporally correlated with each other, the i.i.d assumption is violated and common cross-validation strategies are not applicable. In this study we use a prequential approach, where the time series is divided into sequential folds. In the first iteration, the first n folders are used to train the model, and the subsequent folder for testing. In the following iterations, the testing folder is merged with the training blocks, and the next folder is now used for testing, until all blocks are tested. Results from Cerqueira et

al. (2020) suggest that out-of-sample and prequential approaches show the best performance in non-stationary series.

3.7 *Literature Gaps & Contribution*

From the above review we can conclude that the SIR dataset has been used to gain descriptive insights into the temporal dimension of the injury report as well as to classify injury severity based on multiple categorical features. However, much remains unanswered when it comes to how severe injuries evolve over time and across states from a forecasting perspective. A truly forecasting framing is thus worthy of being developed, as it could improve risk monitoring and enable decision makers on a proactive management of prevention resources. Moreover, we contribute to the discussion in the forecasting community about global versus local modeling of panel data by providing an empirical comparison of both approaches in a novel domain, using a unified feature set and consistent evaluation metrics.

4 METHODS

Figure 1 provides a pictorial overview of the workflow adopted in this study, which is more extensively detailed in the present section.

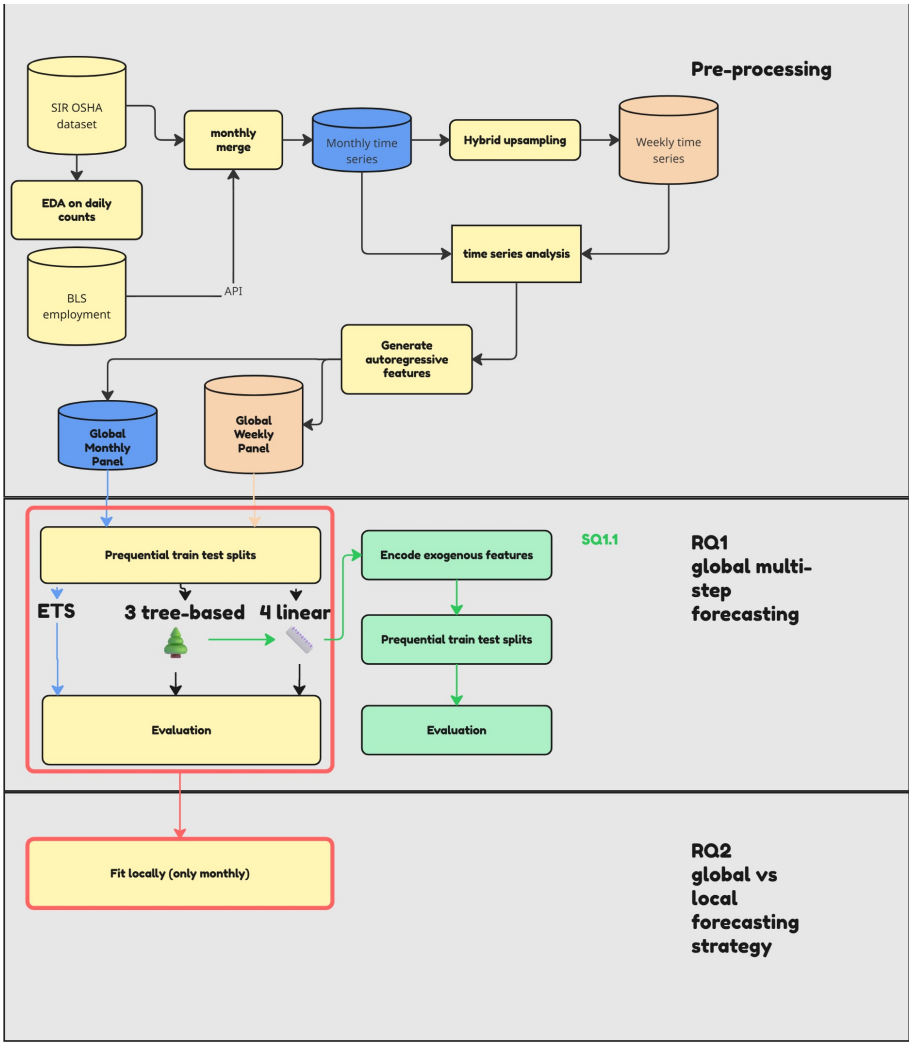


Figure 1: Workflow

4.1 *Dataset Description*

The dataset is publicly available through the OSHA Severe Injury Dashbord [1]. OSHA collects reports about severe injuries occurred within U.S. states

[1] https://www.osha.gov/severe-injury-reports. The dataset was retrieved on August 28, 2025.

under its jurisdiction. These reports form the Severe Injury Reports (SIR) dataset. There are 34 states with a federal plan under OSHA jurisdiction, although the SIR dataset also contains reports from non-federal-plan states when the injured employees are federal workers. In addition, seven states have state plans covering only state and local government workers, while private-sector workers are covered by federal plans. To ensure a fair comparison between states, we included only states operating under a federal OSHA jurisdiction, together with the seven states mentioned above, as all reports from those states are collected from the private sector.

The dataset contains 98,801 entries, spanning from January 1, 2015 to February 28, 2025. Each entry represents a report submitted by an employer. A single report may mention multiple severe injuries (e.g. one hospitalization and one amputation), as employers are required to specify *"the number of employees who suffered a fatality, in-patient hospitalization, amputation, or loss of an eye"*[2]. This means that the feature used to construct the target series (Hospitalization) is a count variable, and that we can not aggregate the data at the report level.

To validate the target feature, we cross-tabulated Hospitalization and Amputation counts in Table 2 and we performed a qualitative analysis of some edge cases. In particular, we studied the *Final Narrative* feature for a sample of accidents where the employer reported both an amputation and a hospitalization. Even in these cases, the record involves only one employee; whereas accidents with multiple hospitalizations do refer to multiple employees. Even if this reporting inconsistency does not significantly impact the hospitalization size, we have to explicitly define our outcome variable as the *incidence of reported inpatient hospitalizations*.

Table 2: Cross-tabulation of hospitalization and amputation indicators

Amputation	**Hospitalized**						
	0	1	2	3	4	5	6
0	41	70,152	550	11	2	2	2
1	18,448	7,154	6	0	1	0	0
2	6	5	6	0	0	0	0

The dataset provides 26 features, ranging from geographical information about the city, the address, and the state where the accident occurred, to details about the nature of the injury, the body part affected, the employer, and the business sector.

To make our model more transparent for decision makers, we derived a new target variable from the feature Hospitalization. As explained

[2] https://www.osha.gov/laws-regs/regulations/standardnumber/1904/1904.39

in Section 3.2, we normalize the raw count of hospitalization with the workforce for each state and for each month included in the dataset.

To achieve interstate comparability, data from the Bureau of Labor Statistics (BLS) were fetched about seasonally adjusted, monthly employment at a state level. Employment is estimated on nonfarm, payroll jobs by the Current Employment Statistics (CES) survey. The covered time period spans from January 1, 2015 to December 31, 2024. Three states, namely American Samoa, Guam, and Northern Mariana Islands, were excluded from the panel since they are not covered by the BLS. After inspecting the testing set, one more U.S. territory, Virgin Islands, was excluded because of few data points.

4.2 *Data Cleaning and Preparation*

To evaluate for forecastability, we computed the percentage of days with no reports over the total temporal range for each state (mean value across all states=0,62). Given the data sparsity at the daily level, we aggregated the data to two lower resolutions using a non-overlapping approach. After merging the two datasets, some missing values originated from a mismatch between the SIR time period (ending in February 2025) and the BLS time period (ending in December 2024). These entries were removed from the panel.

For the weekly forecasting task, the state employment data are upsampled from monthly to weekly frequency using a hybrid strategy. Missing weekly values are interpolated with the exception of the period from March to September 2020. During this interval, the COVID-19 pandemic caused an abrupt drop in employment that the interpolation method was not able to capture properly. For this particular period, each initial monthly value was forward-filled until the next observed monthly value. Upsampling monthly employment data to weekly frequency assumes relatively smooth workforce trends within each month.

After the merging step, we validate the resulting dataset and enforce a balanced panel structure by explicitly filling missing state–date combinations with zeros. These zero values represent structural zeros, corresponding to dates on which no hospitalizations were reported for a given state, rather than missing observations.

No additional missing values were identified in the features used for modeling. We visually checked for inconsistencies introduced during the merging process, such as abnormal spikes that could indicate outliers, and no irregular patterns were found.

4.3 *Exploratory Data Analysis*

A skewed data distribution of the target feature may lead to poor generalization, as the time series can be sensitive to extreme values. Figure 2 shows the counts of hospitalization and amputation before adjusting for the workforce size. The hospitalization distribution is slightly positively skewed.

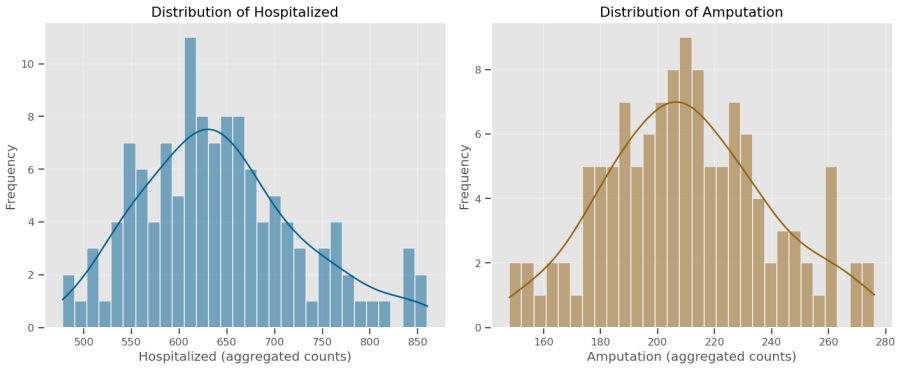


Figure 2: Distributions of hospitalizations and amputations (raw counts)

After adjusting for the workforce, the hospitalization risk distribution is shown in Figure 3 and its normality was assessed with a Shapiro-Wilk test, which rejected the null hypothesis of normality ($W = 0.96$, $p < 0.001$).

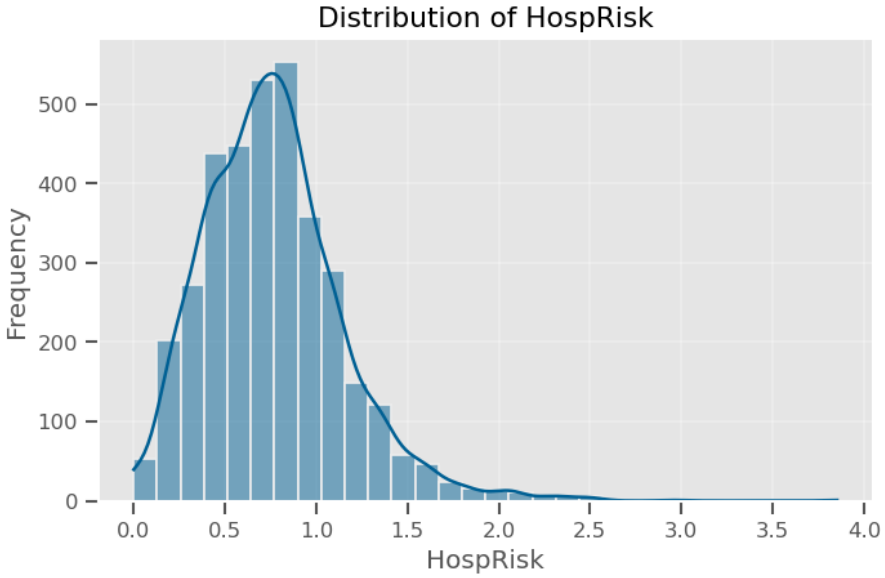


Figure 3: Distribution of hospitalizations adjusted for state workforce

A common approach to normalize a skewed distribution is to apply a log transformation to the target feature. However, because our target is already normalized by workforce size and because the target naturally included zero values, we do not apply further transformations to our target (Benatia, Bellégo, & Pape, 2025).

To assess time-series components, we first visually inspect the mean monthly time series for hospitalization risk across all states in the panel. Figure 4 shows seasonal fluctuations, with regular upwards trends in the summer and a general decline after the COVID-19 period.

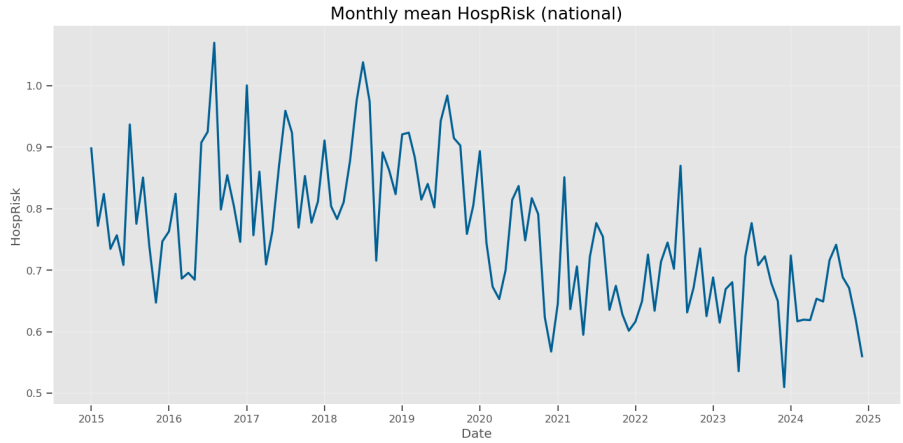


Figure 4: Monthly mean hospitalization risk

A visual inspection further confirms the seasonal components of the time series at a state level, with higher volatility when states with a lower workforce are interested by occasional spikes, as it is clear from Figure 5, which shows three states selected according to their relative workforce size.

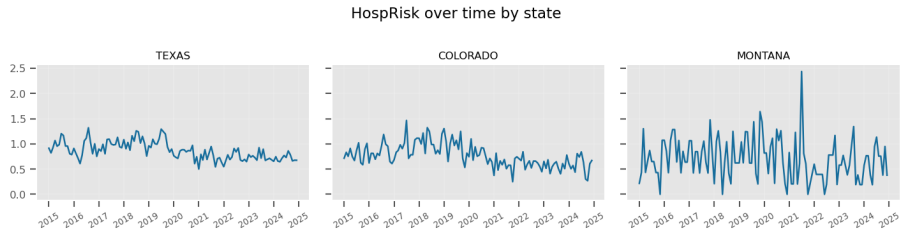


Figure 5: Monthly hospitalization risk for Texas, Colorado, and Montana

A more formal test for temporal components is provided by the seasonal-trend decomposition (STL) shown in Figure 6.

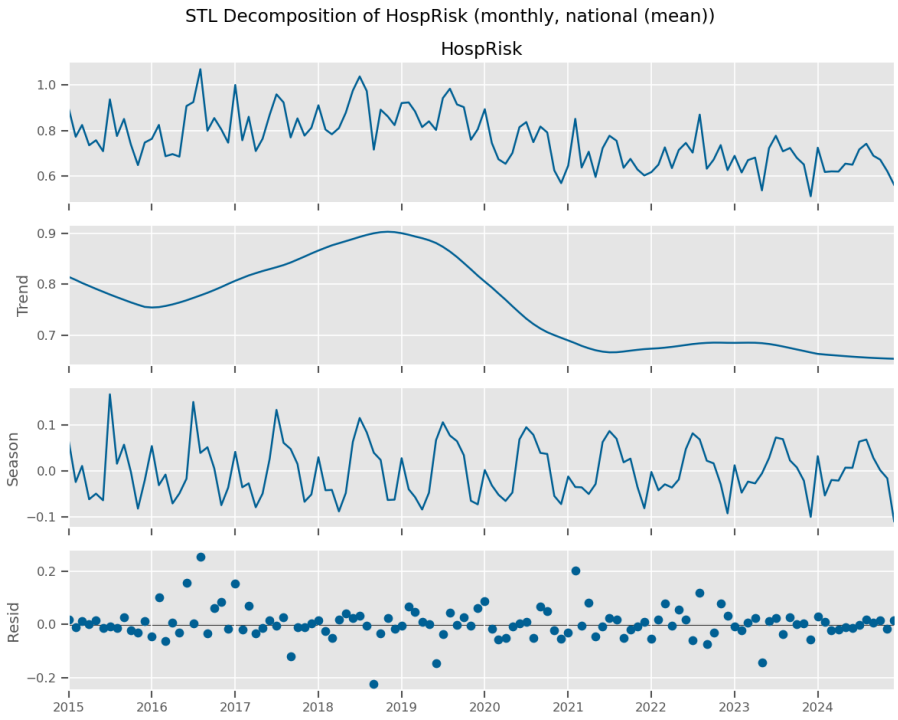


Figure 6: STL decomposition for monthly hospitalization risk at national level

4.4 *Feature Engineering*

Besides deriving a new target feature, some additional steps are necessary to allow machine learning models to learn from temporally ordered data. Table 3 summarizes the autoregressive features needed to frame a time series into a supervised regression problem, next to calendar and exogenous variables.

Table 3: Overview of target variables, autoregressive features, and exogenous features used in the monthly forecasting models.

Category	**Variables**
Target variable	`HospRisk`: hospitalization risk per 100,000 employees.
Calendar and seasonal features	`Year`, `Month`, `Quarter`, `Week of the Year`.
Lagged features	`lag1`, `lag2`, `lag3`, `lag6`, `lag12`.
Rolling features	3-month rolling mean: `roll3`, 6-month rolling mean: `roll6`, 12-month rolling mean: `roll12`, 3-, 6-, 12-month exponential weighted mean: `ewm3`, `ewm6`, `ewm12`.
Exogenous variables	`State` (categorical identifier, one-hot encoded in the models); top k NAICS 2-digit industry: `NAICSmix`; top 3 Nature categories: `share_nature`; top 3 Event categories: `share_event`.

As explained in Section 3.2, the categorical variables `Nature` and `Event` describe, respectively, the nature of injury and the event leading to injury for each reported incident. To incorporate information from these high-cardinality categorical variables into the forecasting models, we constructed aggregate mix features, denoted `share_nature` and `share_event`, which summarize the distribution of category levels within each state–time period.

Specifically, individual category levels (e.g., fractures, falls, transportation incidents) were first mapped to a reduced set of higher-level categories. Because the coding scheme for the `Nature` variable was changed in 2024, causing a strong distortion in the temporal evolution of some category levels visible in Figure 7, the retained levels were selected based on their empirical frequency using data up to the end of 2023.

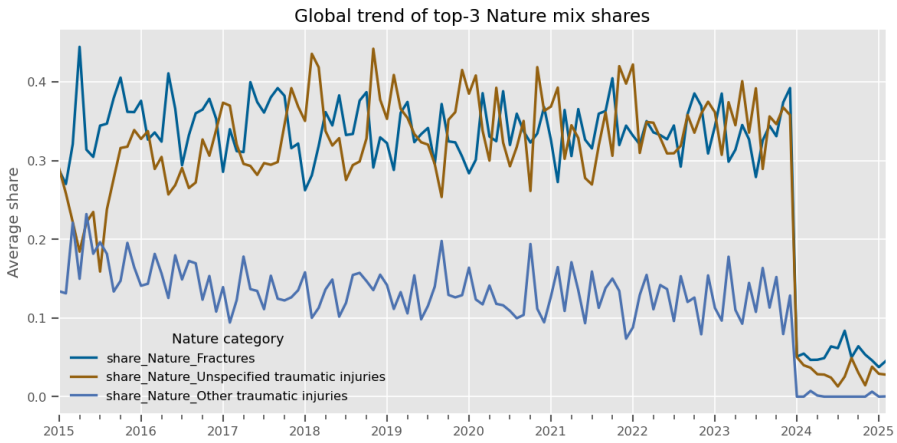


Figure 7: Monthly national trend for the top three levels in the variable `Nature`

For each variable, only the three most frequent category levels were retained, while all remaining levels were grouped into a residual *Other* category. Based on seasonal–trend decomposition, the dominant categorical mix features showed moderate seasonal structure. These features were therefore encoded using lagged 12-period rolling means.

4.5 *Experimental setup*

In this subsection, we describe the experimental setup adopted in each experiment for reproducibility purposes. For the monthly aggregation, each original time series was reframed as a supervised regression problem by constructing a matrix composed of the autoregressive and temporal features summarized in Table 3. Importantly, the exogenous features were used only for answering SQ1.1.

Each observation is identified by a group index (State) and a timestamp corresponding to the beginning of the month. To prevent data leakage, predictors at time t never use information from time $t + h$ or later. This constraint is enforced by shifting the autoregressive features according to the forecasting horizon. The correctness of the temporal split is further validated by explicitly checking the alignment between each train end and the corresponding test start date.

We adopt a prequential validation strategy to obtain robust estimates of out-of-sample predictive performance. Since model accuracy may depend on the specific choice of the training period, evaluating models using a single train–test split could lead to results that are overly sensitive to a particular time window. In each split, the training set includes all

observations up to a given train end date, while the test set consists of the subsequent 12 months only.

Table 4 reports an overview of the resulting splits.

Table 4: Rolling-origin training and test splits (monthly frequency)

Train end	Test start	Test end	Train weeks	Train rows
2016-12-01	2017-01-01	2017-12-01	12	360
2017-12-01	2018-01-01	2018-12-01	24	720
2018-12-01	2019-01-01	2019-12-01	36	1080
2019-12-01	2020-01-01	2020-12-01	48	1440
2020-12-01	2021-01-01	2021-12-01	60	1800
2021-12-01	2022-01-01	2022-12-01	72	2160
2022-12-01	2023-01-01	2023-12-01	84	2520
2023-12-01	2024-01-01	2024-12-01	96	2880

To ensure that each model is trained on a sufficiently long historical window, we discard early splits with limited training data and retain the six train ends corresponding to the years 2018–2023. Each split therefore represents an independent forecasting experiment, in which models are trained using information available up to year t and evaluated on year $t+1$.

An overview of each model's hyperparameters can be found in Appendix A. Linear models use default settings, while tree-based ensemble models share a common set of fixed hyperparameters chosen using simple heuristics rather than formal tuning. In particular, shallow trees, low learning rates, moderate ensemble sizes, and mild subsampling were adopted to account for short training windows and predominantly autoregressive features, favoring stable and comparable performance across rolling-origin splits.

All models are tested on one year of data and evaluated at three forecasting horizons ($h = 1, 3, 6$) using a direct multi-step strategy.

Models are compared against a Seasonal Naive baseline, which generates forecasts by repeating the observed value from the same calendar month in the previous year. This constitutes a strong benchmark in our setting, as the exploratory data analysis and time-series decomposition revealed pronounced annual seasonality and a relatively smooth underlying trend, implying that a large fraction of the predictable variation can already be captured by seasonal persistence alone. After training and testing all models, we implemented an ETS model based on the Holt-Winters method to define a stronger baseline and to assess to what extent such a model could capture level, trend, and seasonal components. Since it is

not possible to fit an ETS model to a panel data, a separate ETS model was fitted for each state using the same rolling-origin splits as the machine learning models.

For the weekly resolution, we adopted the same experimental setup with some changes to accommodate a weekly aggregation. In particular, each period is defined to end on Mondays, and employment data were upsampled using the hybrid technique described in section 4.2. All models are evaluated at three forecasting horizons ($h = 1, 4, 13$) corresponding to one week, one month and one quarter ahead. Autoregressive features are generated with weekly periods.

We decided not to fit an ETS model for the weekly aggregation after inspecting the STL decomposition and performing an error analysis on the fitted models. Since the weekly national time series shows sharp weekly spikes and high volatility, the assumptions underlying exponential smoothing models are defied.

TO WRITE: ets; difference global and local strategy implementations; permutation importance; state encoding; yaojohnson transf.

5 RESULTS

This section presents the results from three experiments. The same metrics, namely RMSE, MAE, and MASE are used to evaluate all the experiments. 1. aggregate metrics for the monthly panel data, globally pooled, against the weekly panel data 2. aggregate metrics for the monthly panel data with exogenous variables 3. aggregate metrics for the monthly panel data locally estimated

A visual analysis of the errors and a comparison of the models performance will follow. The feature permutation will be the last step. The final results from the monthly panel data, aggregated across six folds, are summarized in Table 5 for the short-term (1 month), in Table 6 for the medium-term (3 months) and in Table 7. As it is clear from a quick inspection, all models performed similarly, with no significant difference across linear and tree-based families.

Across the three forecast horizons, all models show stable performance as measured by RMSE, MAE, and MASE. In all cases, the evaluated models outperform the seasonal baseline. On the long-term horizon, performance relative to the baseline deteriorates slightly for all models, indicating either increased forecast uncertainty or a slightly more predictable seasonal trend in the baseline. While average errors stay comparable across models, the standard deviation of the error metrics is higher for XGBoost and LightGBM, a potential sign of greater variability across the test folds and a tendency toward overfitting. Because the MAE does not vary across

Table 5: Monthly experiment (globally pooled, no exogenous features). Forecast horizon $h = 1$. Scores are reported as mean ± standard deviation across prequential test folds.

Model	RMSE	MAE	MASE
CatBoost	0.283 ± 0.022	0.206 ± 0.022	0.716 ± 0.082
ElasticNet	0.283 ± 0.022	0.207 ± 0.019	0.721 ± 0.065
Lasso	0.283 ± 0.021	0.207 ± 0.019	0.720 ± 0.065
LightGBM	0.284 ± 0.022	0.208 ± 0.022	0.724 ± 0.081
PLS	0.284 ± 0.022	0.208 ± 0.020	0.724 ± 0.067
Ridge	0.284 ± 0.022	0.207 ± 0.019	0.721 ± 0.065
Seasonal Naive	0.390 ± 0.037	0.276 ± 0.027	0.941 ± 0.083
XGBoost	0.289 ± 0.025	0.210 ± 0.024	0.728 ± 0.083

Table 6: Monthly experiment (globally pooled, no exogenous features). Forecast horizon $h = 3$. Scores are reported as mean ± standard deviation across prequential test folds.

Model	RMSE	MAE	MASE
CatBoost	0.282 ± 0.027	0.205 ± 0.026	0.713 ± 0.090
ElasticNet	0.282 ± 0.023	0.205 ± 0.019	0.715 ± 0.060
Lasso	0.282 ± 0.022	0.204 ± 0.019	0.713 ± 0.061
LightGBM	0.287 ± 0.029	0.209 ± 0.028	0.727 ± 0.095
PLS	0.283 ± 0.023	0.205 ± 0.019	0.718 ± 0.061
Ridge	0.283 ± 0.023	0.205 ± 0.019	0.718 ± 0.058
Seasonal Naive	0.388 ± 0.036	0.275 ± 0.027	0.938 ± 0.088
XGBoost	0.290 ± 0.034	0.209 ± 0.030	0.730 ± 0.101

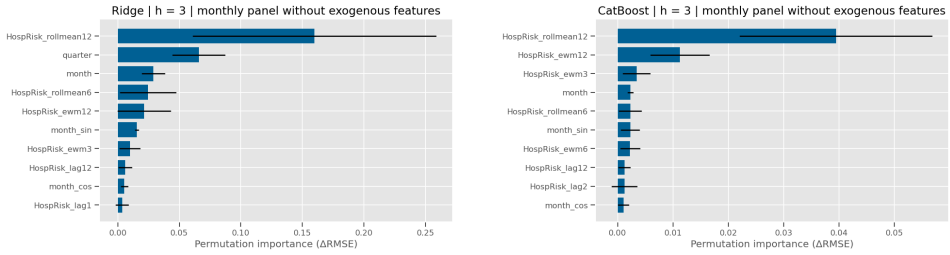
Table 7: Monthly experiment (globally pooled, no exogenous features). Forecast horizon $h = 6$. Scores are reported as mean ± standard deviation across prequential test folds.

Model	RMSE	MAE	MASE
CatBoost	0.350 ± 0.028	0.254 ± 0.025	0.772 ± 0.058
ElasticNet	0.350 ± 0.028	0.254 ± 0.025	0.772 ± 0.058
Lasso	0.350 ± 0.028	0.254 ± 0.025	0.772 ± 0.058
LightGBM	0.351 ± 0.029	0.255 ± 0.026	0.775 ± 0.062
PLS	0.350 ± 0.028	0.254 ± 0.025	0.772 ± 0.058
Ridge	0.350 ± 0.028	0.254 ± 0.025	0.772 ± 0.058
Seasonal Naive	0.397 ± 0.040	0.279 ± 0.029	0.873 ± 0.066
XGBoost	0.357 ± 0.031	0.259 ± 0.028	0.789 ± 0.066

the models, the MASE values show limited variation. This may indicate that model choice has only a marginal impact on forecast accuracy once seasonality and autoregressive structure are accounted for.

The results for the weekly aggregation are reported in the Appendix A because they do not reveal any significant difference compared to the monthly setting. Error metrics are only slightly more stable across folds, suggesting that more observations did not benefit one model above the others.

These findings suggest that all models learn from similar patterns in the data, and that forecast accuracy is largely driven by a small number of dominant features. To answer this question, permutation importance was computed and stored for each model. Permutation importance quantifies the increase in forecast error when the values of a feature are shuffled. This measurement is model-agnostic (Fisher, Rudin, & Dominici, 2019) and therefore more suitable for comparing models that carry different assumptions, such as linear and tree-based ones. Figure 8 illustrates the permutation importance at horizon $h = 3$ for two representative models.



(a) Ridge (b) CatBoost

Figure 8: Permutation importance for Ridge (left) and CatBoost (right) for horizon $h = 3$. Results are aggregated across prequential test folds.

The 12-month rolling mean is clearly the dominant predictor across all models. Permuting this features leads to a deterioration in forecast accuracy, confirming that long-term trend information is the driver of predictive performance. This may also explain the stability in error metrics across short-, medium-, and long-term horizons. For the Ridge regressor, permuting the 12-month rolling mean increases RMSE by approximately 0.16, whereas for CatBoost the increase is about 0.04. This may suggest that linear models rely more heavily on this single feature. However, since rolling means and exponentially weighted moving averages are correlated autoregressive features and they also rank among the most important features for the CatBoost model, the apparent difference may be much smaller.

To test whether machine learning models may perform any better compared to classical statistical models, an additive ETS model was fitted separately on each single state time series and forecasts were aggregated across states. Results for the monthly panel data, reported in Table 8, indicate that the ETS benchmark has higher RMSE and MAE than all machine learning models, with the exception of the seasonal baseline.

Table 8: Performance of the ETS model after aggregating state-level predictions. Results are reported as mean ± standard deviation across prequential test folds.

Model	RMSE	MAE
ETS ($h = 1$)	0.314 ± 0.030	0.225 ± 0.024
ETS ($h = 3$)	0.319 ± 0.037	0.228 ± 0.030
ETS ($h = 6$)	0.325 ± 0.045	0.231 ± 0.035

After performing Wilcoxon signed-rank tests on fold-level RMSE values. For example, we found a statistically significant difference ($p = 0.03$) between CatBoost and the ETS benchmark at each forecast horizon. While providing evidence that machine learning models tend to outperform a default ETS, this result should be interpreted with caution, given the small number of test folds and the mild dependence induced by rolling-origin evaluation.

The error analysis of the residuals for a Lasso model (again, the results are very similar across all models) in Figure 9 shows a regression-to-the-mean behavior, systematically underestimating high values and overestimating low values.

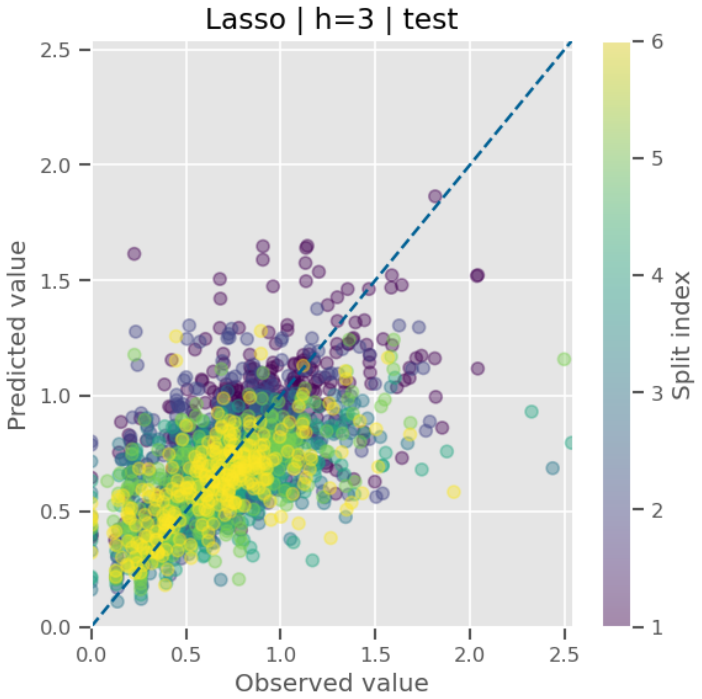


Figure 9: Scatter plot of predicted against true values for a Lasso model, $h = 3$

More pronounced differences emerge when a time series aggregated across the 30 states included in the dataset is visualized as the line plot shown in Figure 10 and Figure 11.

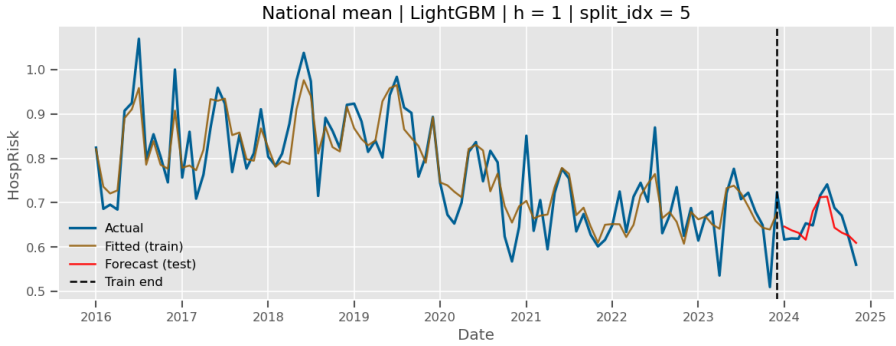


Figure 10: Fitted and predicted values against true values for a LightGBM model, $h = 1$, monthly mean across all selected states

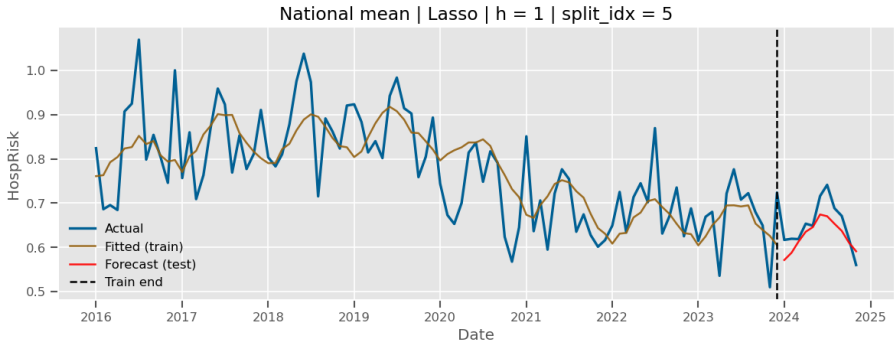
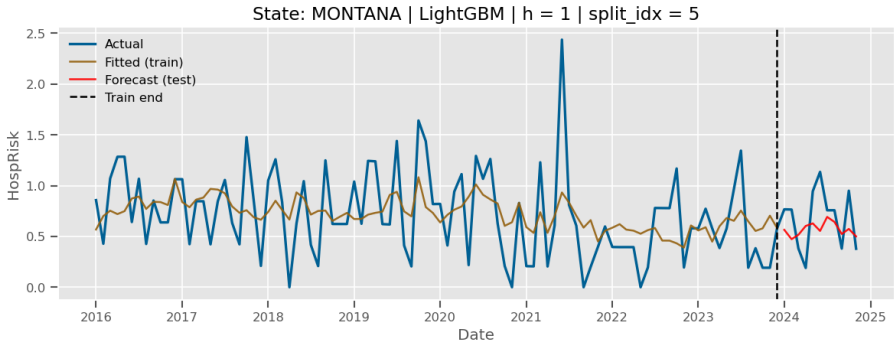


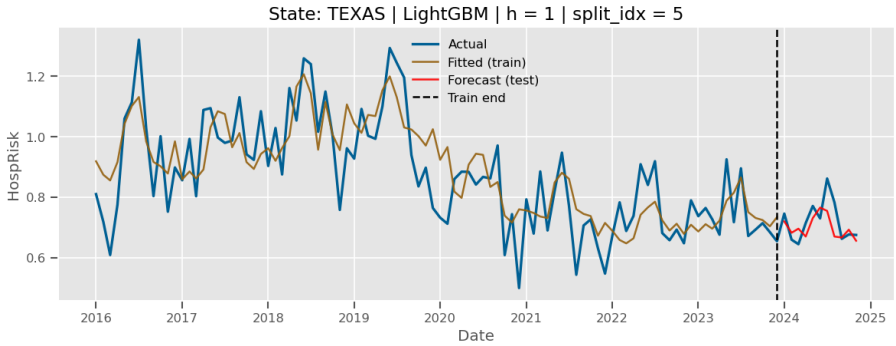
Figure 11: Fitted and predicted values against true values for a Lasso model, $h = 1$, monthly mean across all selected states

LightGBM, as a representative tree-based ensemble model, is able to capture some of the spikes in the training set, but shows reduced generalization capacity on the test set, suggesting overfitting. In contrast, the Lasso model, representing linear approaches, produces smoother fitted values that dampen such spikes, while achieving comparable accuracy on the test set. Both models appear to struggle in capturing downward spikes.

When zooming in on individual states, heterogeneity in hospitalization risk values becomes evident. States with a smaller workforce tend to show more volatile and less predictable trajectories. While LightGBM yields fitted and forecasted paths for Texas that are broadly consistent with the cross-state average, it is less able to capture the high variability observed in a small state such as Montana, as shown in Figure 12.



(a) Montana



(b) Texas

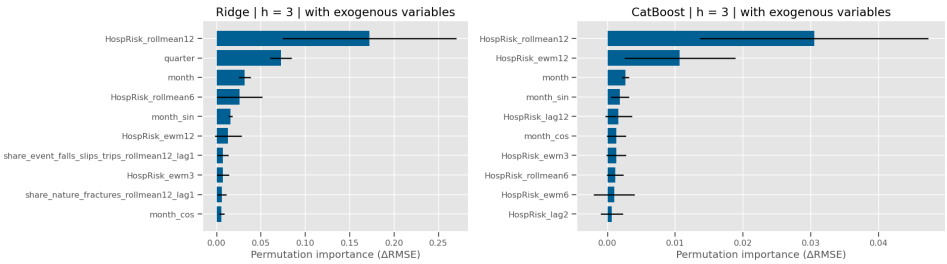
Figure 12: Observed values, fitted values on the training set, and one-step-ahead forecasts for the LightGBM model ($h = 1$).

Moving to the feature matrix expanded with exogenous features (NAICS industry mix, Event mix, and Nature mix), Table 9 summarizes the metrics for $h = 3$.

Table 9: Forecasting performance of models with exogenous variables. Results are reported as mean ± standard deviation across prequential test folds.

Model	RMSE	MAE	MASE
CatBoost	0.289 ± 0.030	0.211 ± 0.027	0.732 ± 0.091
ElasticNet	0.286 ± 0.025	0.209 ± 0.021	0.729 ± 0.067
Lasso	0.285 ± 0.025	0.208 ± 0.021	0.724 ± 0.067
LightGBM	0.293 ± 0.030	0.215 ± 0.029	0.753 ± 0.104
PLS	0.286 ± 0.026	0.209 ± 0.022	0.730 ± 0.072
Ridge	0.287 ± 0.026	0.210 ± 0.022	0.735 ± 0.068
Seasonal Naive	0.396 ± 0.035	0.280 ± 0.027	0.951 ± 0.091
XGBoost	0.296 ± 0.032	0.216 ± 0.029	0.751 ± 0.100

While all metrics show a pattern similar to the one for the feature matrix without the exogenous features, the general performance seems to be slightly deteriorated. The compositional exogenous features do barely appear in the permutation importance of the Ridge model and are absent from the CatBoost, as shown in Figure 13.



(a) Ridge (b) CatBoost

Figure 13: Permutation importance for Ridge (left) and CatBoost (right) for horizon $h = 3$, with exogenous variables. Results are aggregated across prequential test folds.

Finally, we compare the forecast metrics of the global models with that of the locally fitted ones. Table 10 reports the metrics for the local strategy at horizon $h = 6$, where some major differences are visible both across models within the local framework and when contrasting local and global strategies.

Three main findings can be highlighted. First, LightGBM achieves the best performance among all local models, outperforming the second-best

Table 10: Local forecasting performance (horizon $h = 6$). Values are mean ± std across rolling-origin splits.

Model	RMSE	MAE	MASE
CatBoost	0.283 ± 0.158	0.235 ± 0.133	0.810 ± 0.244
ElasticNet	0.300 ± 0.194	0.249 ± 0.167	0.843 ± 0.361
Lasso	0.297 ± 0.193	0.247 ± 0.167	0.833 ± 0.355
LightGBM	0.278 ± 0.159	0.229 ± 0.134	0.784 ± 0.229
PLS	0.301 ± 0.194	0.250 ± 0.168	0.853 ± 0.381
Ridge	0.289 ± 0.181	0.240 ± 0.154	0.818 ± 0.331
Seasonal Naive	0.337 ± 0.192	0.272 ± 0.154	0.926 ± 0.266
XGBoost	0.292 ± 0.162	0.243 ± 0.138	0.845 ± 0.278

competitor, CatBoost, with a statistically significant margin according to a Wilcoxon signed-rank test ($p = 0.02$). Second, the local LightGBM significantly outperform the global counterpart, indicating that state-specific fitting can be beneficial at longer horizons. Third, the standard deviation of forecast errors is markedly higher under the local approach than under the global approach. This increased dispersion suggests that while local models can produce better median performance, their gains are less stable across time, highlighting a trade-off between accuracy and robustness.

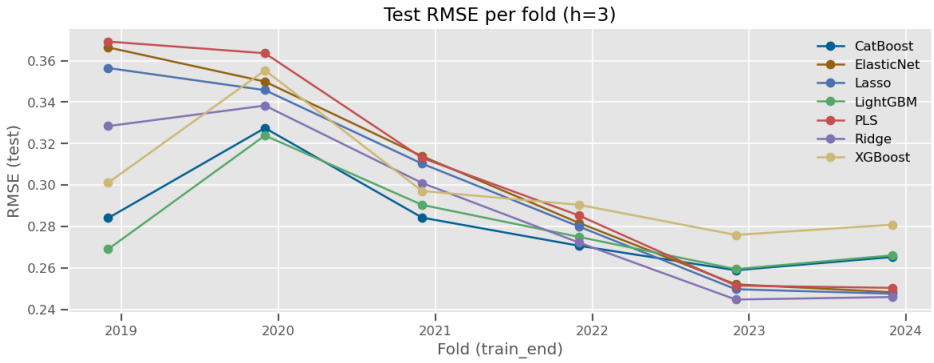


Figure 14: RMSE values for the locally fitted models across six splits, $h = 3$.

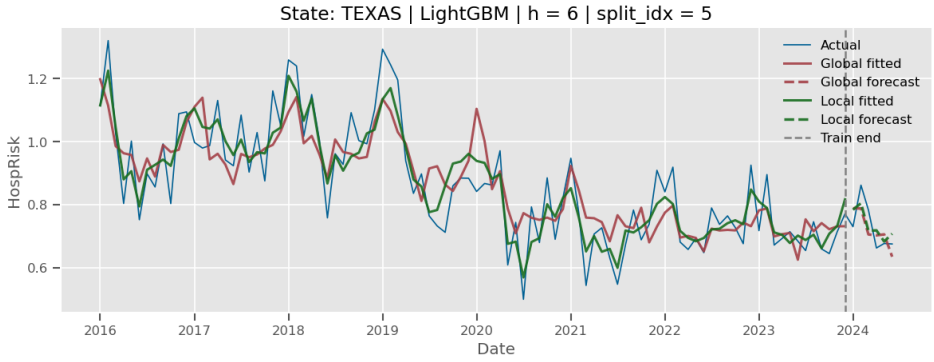


Figure 15: Fitted and predicted values against true values for a LightGBM model, $h = 1$, local (green line) against global model (red line), Texas.

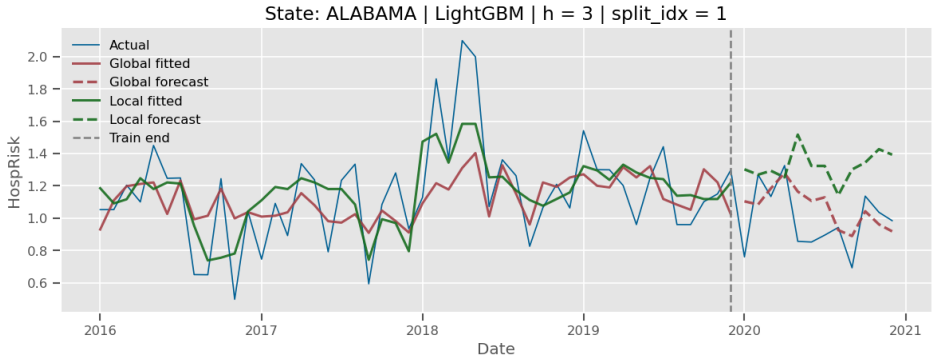


Figure 16: Fitted and predicted values against true values for a LightGBM model, $h = 3$, local (green line) against global model (red line), Alabama.

6 DISCUSSION

6.1 *Answering Research Questions*

RQ1

RQ1: *How does the forecasting performance of a set of linear and tree-based ensemble models vary at two temporal aggregations (monthly and weekly) in multi-step forecasting of severe injuries risk, as evaluated by RMSE, MAE and MASE?*

Overall, the forecasting performance of linear models (ElasticNet, Lasso, PLS, and Ridge) and tree-based ensemble models (CatBoost, LightGBM, and XGBoost) is similar, both at monthly and weekly resolutions, and across the three forecasting horizons considered. The increased sample size in the weekly setting led the models converge to even more stable

predictions. While Cerqua, Giannantoni, et al. (2024) found PLS to be the best performing algorithm in terms of Mean Squared Forecast Error, its performance in our study was comparable with the other algorithms. Direct comparison with their results should be made with caution, as their analysis relied on a wide set of covariates to model fatalities at a finer geographical granularity.

Although the performance metrics did not show significant differences between linear and tree-based ensemble models, visual inspection of the fitted training values highlighted a tendency for tree-based ensembles to overfit more strongly. This behavior is consistent with previous findings.

SQ1.1

SQ1.1: How does the inclusion of exogenous categorical variables affect the forecasts of the ML models?

The inclusion of exogenous categorical variables slightly deteriorated the performance metrics and did not benefit the interpretability of the features. Quantitatively, this finding somehow reflect the finding of Cerqua, Giannantoni, et al. (2024), who found that models performed better on a smaller set of covariates.

RQ2

RQ2: *How does the predictive accuracy change when estimated as a Global Forecasting Model (across all states jointly) rather than as a Local Forecasting Model (state-specific), for both monthly and weekly temporal resolutions?*

We did find only small evidence of LightGBM yielding lower RMSE on the locally fitted approach compared to the global one at longer monthly forecasting horizons ($h = 6$).

6.2 *Limitations and further research*

This study presents two main kinds of limitations: those arising from modeling and design choices, and those inherent to the nature of the dataset and to the way reports are collected.

6.2.1 *Design limitations*

A key empirical finding was that forecasts are mainly driven by autoregressive features across all models, namely the rolling means. Including compositional exogenous variables did not improve the model's ability to capture volatility. The assumption that higher hospitalization risks in small

states may be associated with specific industrial sectors is not reflected in the recorded nature and event classifications leading to hospitalization. The compositional features were encoded in a simplex space, with values bounded in the [0, 1] interval. Although alternative transformations exist that map compositional data to an unbounded space (e.g., Zhang et al. (2024) review several log-ratio transformations), the limited contribution of those features suggests that the effect will probably be negligible, potentially improving interpretability rather than performance.

Temporal aggregation to coarser resolutions was necessary to reduce the sparsity present at the daily frequency. However, these aggregations constrained the potential richness of the dataset, in particular in terms of the set of usable covariates. As a result, potentially informative features (such as the `Final Narrative`) were excluded. Again, it is possible to argue that including more features could lead to marginal improvements in performance at the cost of interpretability. Moreover, complex, customized preprocessing pipelines are prone to implementation mistakes.

A structure change following the COVID-19 pandemic may have had the effect of stabilizing the hospitalization risk. After an initial adjustment phase, forecasting performance generally improved from 2021 onward. However, this improvement may be simply due to longer train periods rather than to an increased capacity of the models to capture structural patterns. The rolling-origin evaluation approach, while standard in the forecasting literature, makes it difficult to disentangle performance gains due to longer training sets from those resulting from a genuine capacity of generalizing on out-of-sample test data. Future research could address this limitation by adopting an alternative evaluation strategy – such as fixed-length rolling windows – to better isolate structural effects from sample-size effects (Cerqueira et al., 2020).

Finally, while our design choice included some basic calendar features –such as month, quarter, and year– integrating other calendar features is still possible. In particular, encoding of public holidays may help models better capture intra-seasonal downward spikes (around the end of the year). Additionally, indicators of policy changes may provide some guidance in modeling unexplained shifts in hospitalization risks.

6.2.2 *Dataset limitations*

As highlighted in Section 4.1, inconsistent reporting schemes make it harder to encode the hospitalization risk. Moreover, changes in the categorization of the *Nature* variable limited the forecasting analysis to data up to 2023, unless one can reliably relabel the category levels.

More importantly, as Cerqua, Giannantoni, et al. (2024) argued in their discussion of target selection, occupational injury reports suffer from

underreporting, which is the systematic omission of injury events from official records. Consequently, the true number of severe injuries may be higher than what it is observed in the SIR dataset.

Two broader limitations concern the nature of the target variable. First, hospitalization risk is a lagging indicator, and as such it can only be computed *after* an accident has occurred. In contrast, leading indicators –such as near-miss reports and safety inspections — are more proactive safety measures (Yapi, Latouche, Guillin, & Bailly, 2025) and can provide early signals of hazardous conditions. Future research could augment the SIR dataset with complementary sources reporting on these leading indications, such as the OSHA Enforcement Data.

Second, severe injuries are extremely rare events. Since few positive events at higher resolutions occur, it becomes hard to distinguish signal from noise. As it is clear from the standard deviation values in RMSE, the global panel approach mitigated the fluctuations relying on smoothing features learned across the states, such as the rolling mean. However, framing the problem as a continuous regression task may not be optimal in the presence of rare events.

An alternative modeling strategy would be to explicitly account for the discrete nature of severe injuries by first forecasting the probability that at least one severe injury occurs within a given period, and then modeling the expected number of occurrences conditional on an event. This approach leads to a two-stage modeling framework, combining a classification component for event occurrence with a count regression model for event intensity. A similar approach can be found in the binary time series modeling proposed by Yapi et al. (2025).

Finally, the relevance of hierarchical and panel-based forecasting structures in current forecasting research is illustrated by the design of the M5 Forecasting Competition, where participants were required to generate forecasts across multiple aggregation levels of sales data (Makridakis et al., 2022). Although a common geographical framework was preferred in our study, geostatistical approaches may reveal a more effective clustering structures of the states included in the SIR dataset. Identifying regional clusters and combining them with hierarchical forecasts may improve performance while still preserving model interpretability (Wickramasuriya, Athanasopoulos, & Hyndman, 2019).

7 CONCLUSION

This study helped bridge a research gap in the current literature on forecasting severe occupational injuries by comparing a set of linear and tree-based ensemble models to forecast the hospitalization risk across selected U.S.

states. Models performed similarly at different temporal resolutions, and across three chosen forecasting horizons. All models consistently outperformed seasonal and statistical baselines, showing that a machine learning approach yields lower prediction errors even with a small sample size.

Since model interpretability is crucial for actionable insights, inspecting the permutation importance highlighted that rolling means played a fundamental role in stabilizing the predictions, with linear models achieving more robust predictions in terms of standard deviations. These findings suggest that, in panel settings with rare events, pooling strategies and feature design may be more influential than model complexity.

Overall, this work contributes some empirical evidence that simple, interpretable machine learning models can provide reliable forecasts of severe injury risk when combined with appropriate aggregation and pooling strategies. Future research may extend these results by integrating leading indicators, exploring hierarchical forecasting structures, and assessing generalizability across related occupational safety datasets.

REFERENCES

Baltagi, B. H. (2013). Chapter 18 - Panel Data Forecasting. In G. Elliott & A. Timmermann (Eds.), *Handbook of Economic Forecasting* (Vol. 2, pp. 995–1024). Elsevier. Retrieved from https://www.sciencedirect.com/science/article/pii/B978044462731500018X (ISSN: 1574-0706) doi: https://doi.org/10.1016/B978-0-444-62731-5.00018-X

Barhrhouj, A., Ananou, B., & Ouladsine, M. (2025). Assessing the Impact of Temporal Data Aggregation on the Reliability of Predictive Machine Learning Models. In V. Julian et al. (Eds.), *Intelligent Data Engineering and Automated Learning – IDEAL 2024* (Vol. 15346, pp. 481–492). Cham: Springer Nature Switzerland. Retrieved 2025-12-22, from https://link.springer.com/10.1007/978-3-031-77731-8_43 (Series Title: Lecture Notes in Computer Science) doi: 10.1007/978-3-031-77731-8_43

Benatia, D., Bellégo, C., & Pape, L. (2025, September). *Dealing with Logs and Zeros in Regression Models.* arXiv. Retrieved 2025-12-30, from http://arxiv.org/abs/2203.11820 (arXiv:2203.11820 [econ]) doi: 10.48550/arXiv.2203.11820

Bojer, C. S. (2022, October). Understanding machine learning-based forecasting methods: A decomposition framework and research opportunities. *International Journal of Forecasting*, *38*(4), 1555–1561. Retrieved 2025-11-17, from https://linkinghub.elsevier.com/retrieve/pii/S0169207021001771 doi: 10.1016/j.ijforecast.2021.11.003

Cerqua, A., Giannantoni, C., Letta, M., & Pinto, G. (2024, December). *'Dead Man Working': A Place-based Approach to Workplace Fatalities* [SSRN Scholarly Paper]. Rochester, NY: Social Science Research Network. Retrieved 2025-09-08, from https://papers.ssrn.com/abstract=5040905 doi: 10.2139/ssrn.5040905

Cerqua, A., Letta, M., & Pinto, G. (2024). *On the (Mis)Use of Machine Learning with Panel Data.* SSRN. Retrieved 2025-11-07, from https://www.ssrn.com/abstract=5014594 doi: 10.2139/ssrn.5014594

Cerqueira, V., Torgo, L., & Mozetič, I. (2020, November). Evaluating time series forecasting models: an empirical study on performance estimation methods. *Machine Learning*, *109*(11), 1997–2028. Retrieved 2026-01-07, from https://doi.org/10.1007/s10994-020-05910-7 doi: 10.1007/s10994-020-05910-7

Chen, J. M. (2021, February). An Introduction to Machine Learning for Panel Data. *International Advances in Economic Research*, *27*(1), 1–16. Retrieved 2025-11-07, from https://doi.org/10.1007/s11294-021-09815-6 doi: 10.1007/s11294-021-09815-6

Coulombe, P. G., Marcellino, M., & Stevanovic, D. (2025). Panel Machine Learning with Mixed-Frequency Data: Monitoring State-Level Fiscal Variables.

Fang, Z.-g., Yang, S.-q., Lv, C.-x., An, S.-y., & Wu, W. (2022, July). Application of a data-driven XGBoost model for the prediction of COVID-19 in the USA: a time-series study. *BMJ Open*, *12*(7), e056685. Retrieved 2026-01-06, from https://bmjopen.bmj.com/lookup/doi/10.1136/bmjopen-2021-056685 doi: 10.1136/bmjopen-2021-056685

Fisher, A., Rudin, C., & Dominici, F. (2019). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of machine learning research : JMLR*, *20*, 177. Retrieved 2026-01-08, from https://pmc.ncbi.nlm.nih.gov/articles/PMC8323609/

Gomes, H., Parasram, V., Collins, J., & Socias-Morales, C. (2023, September). Time series, seasonality and trend evaluation of 7 years (2015–2021) of OSHA severe injury data. *Journal of Safety Research*, *86*, 30–38. Retrieved 2025-09-14, from https://www.sciencedirect.com/science/article/pii/S0022437523000798 doi: 10.1016/j.jsr.2023.06.005

Hall, T., & Rasheed, K. (2025, January). A Survey of Machine Learning Methods for Time Series Prediction. *Applied Sciences*, *15*(11), 5957. Retrieved 2025-11-17, from https://www.mdpi.com/2076-3417/15/11/5957 (Publisher: Multidisciplinary Digital Publishing Institute) doi: 10.3390/app15115957

Han, C., Zhou, Y., Sun, J., & Li, Z. (2025, November). An optimized ridge regression for forecasting time series with a fixed pe-

riod. *Pattern Recognition Letters*, *197*, 274–281. Retrieved 2025-12-22, from https://www.sciencedirect.com/science/article/pii/S0167865525002971 doi: 10.1016/j.patrec.2025.08.017

Hasan Khalleefah Hassan, M., & Khalifa, W. M. S. (2025). Work Place Safety: Machine Learning Techniques for Assessing Workplace Incident Severity. *IEEE Access*, *13*, 34211–34226. Retrieved 2025-09-10, from https://ieeexplore.ieee.org/abstract/document/10891569 doi: 10.1109/ACCESS.2025.3543136

Hewamalage, H., Bergmeir, C., & Bandara, K. (2022, April). Global models for time series forecasting: A Simulation study. *Pattern Recognition*, *124*, 108441. Retrieved 2025-11-07, from https://www.sciencedirect.com/science/article/pii/S0031320321006178 doi: 10.1016/j.patcog.2021.108441

Howe, A. S., Tan, J., Yuen, B., Saini, H., Saade-Cleves, N., Obeidat, D., ... Nowrouzi-Kia, B. (2024, August). Physical and Psychosocial Correlates of Occupational Physical Injury in the Global Construction Industry: A Scoping Review. *Environmental Health Insights*, *18*, 11786302241270371. Retrieved 2025-12-24, from https://pmc.ncbi.nlm.nih.gov/articles/PMC11345736/ doi: 10.1177/11786302241270371

Hyndman, R. J., & Koehler, A. B. (2006, October). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. Retrieved 2026-01-06, from https://www.sciencedirect.com/science/article/pii/S0169207006000239 doi: 10.1016/j.ijforecast.2006.03.001

Johnson, M. S., Levine, D. I., & Toffel, M. W. (2023, October). Improving Regulatory Effectiveness through Better Targeting: Evidence from OSHA. *American Economic Journal: Applied Economics*, *15*(4), 30–67. Retrieved 2025-11-06, from https://pubs.aeaweb.org/doi/10.1257/app.20200659 doi: 10.1257/app.20200659

Khairuddin, M. Z. F., Lu Hui, P., Hasikin, K., Abd Razak, N. A., Lai, K. W., Mohd Saudi, A. S., & Ibrahim, S. S. (2022, October). Occupational Injury Risk Mitigation: Machine Learning Approach and Feature Optimization for Smart Workplace Surveillance. *International Journal of Environmental Research and Public Health*, *19*(21), 13962. Retrieved 2025-09-08, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9653932/ doi: 10.3390/ijerph192113962

Kontopoulou, V. I., Panagopoulos, A. D., Kakkos, I., & Matsopoulos, G. K. (2023, August). A Review of ARIMA vs. Machine Learning Approaches for Time Series Forecasting in Data Driven Networks. *Future Internet*, *15*(8), 255. Retrieved 2025-11-29, from https://www.mdpi.com/1999-5903/15/8/255 (Publisher: Multidisci-

plinary Digital Publishing Institute) doi: 10.3390/fi15080255

Kurgan, L. A., & Musilek, P. (2006, March). A survey of Knowledge Discovery and Data Mining process models. *The Knowledge Engineering Review*, *21*(1), 1–24. Retrieved 2026-01-10, from https://www.cambridge.org/core/journals/knowledge-engineering-review/article/survey-of-knowledge-discovery-and-data-mining-process-models/368D6AFE435EB5E30378398D34D61C17 doi: 10.1017/S0269888906000737

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018, March). Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PLOS ONE*, *13*(3), e0194889. Retrieved 2025-12-24, from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0194889 (Publisher: Public Library of Science) doi: 10.1371/journal.pone.0194889

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022, October). M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, *38*(4), 1346–1364. Retrieved 2025-11-29, from https://www.sciencedirect.com/science/article/pii/S0169207021001874 doi: 10.1016/j.ijforecast.2021.11.013

Masini, R. P., Medeiros, M. C., & Mendes, E. F. (2023, February). Machine learning advances for time series forecasting. *Journal of Economic Surveys*, *37*(1), 76–111. Retrieved 2025-11-07, from https://onlinelibrary.wiley.com/doi/10.1111/joes.12429 doi: 10.1111/joes.12429

Michaels, D., & Wagner, G. R. (2025, April). OSHA Injury Data: An Opportunity for Improving Work Injury Prevention. *American Journal of Public Health*, *115*(4), 588–595. Retrieved 2025-09-04, from https://ajph.aphapublications.org/doi/full/10.2105/AJPH.2024.307934 (Publisher: American Public Health Association) doi: 10.2105/AJPH.2024.307934

Montero-Manso, P., & Hyndman, R. J. (2021, March). *Principles and Algorithms for Forecasting Groups of Time Series: Locality and Globality*. arXiv. Retrieved 2025-11-05, from http://arxiv.org/abs/2008.00444 (arXiv:2008.00444 [cs]) doi: 10.48550/arXiv.2008.00444

Obasi, I. C., Cheng, P., Varianou-Mikellidou, C., Dimopoulos, C., & Boustras, G. (2026, March). Machine learning for occupational accident analysis: Applications, challenges, and future directions. *Journal of Safety Science and Resilience*, *7*(1), 100250. Retrieved 2025-12-22, from https://www.sciencedirect.com/science/article/pii/S2666449625000842 doi: 10.1016/j.jnlssr.2025.100250

Organization, I. L. (2020). *Safety + health for all : an ILO Flagship Programme : key facts and figures (2016-2020) - International Labour Organization.*

Retrieved 2025-11-29, from https://labordoc.ilo.org/discovery/fulldisplay/alma995108293102676/41ILO_INST:41ILO_V2

Qu, R., Timmermann, A., & Zhu, Y. (2024, July). Comparing forecasting performance with panel data. *International Journal of Forecasting*, *40*(3), 918–941. Retrieved 2025-11-08, from https://www.sciencedirect.com/science/article/pii/S0169207023000766 doi: 10.1016/j.ijforecast.2023.08.001

Reich, N. G., Brooks, L. C., Fox, S. J., Kandula, S., McGowan, C. J., Moore, E., ... Shaman, J. (2019, February). A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(8), 3146–3154. Retrieved 2025-11-29, from https://pmc.ncbi.nlm.nih.gov/articles/PMC6386665/ doi: 10.1073/pnas.1812594116

Rostami-Tabar, B., Goltsos, T. E., & Wang, S. (2023, February). Forecasting for lead-time period by temporal aggregation: Whether to combine and how. *Computers in Industry*, *145*, 103803. Retrieved 2025-12-20, from https://www.sciencedirect.com/science/article/pii/S0166361522001993 doi: 10.1016/j.compind.2022.103803

Spiliotis, E. (2023). Time Series Forecasting with Statistical, Machine Learning, and Deep Learning Methods: Past, Present, and Future. In M. Hamoudia, S. Makridakis, & E. Spiliotis (Eds.), *Forecasting with Artificial Intelligence: Theory and Applications* (pp. 49–75). Cham: Springer Nature Switzerland. Retrieved 2025-12-24, from https://doi.org/10.1007/978-3-031-35879-1_3 doi: 10.1007/978-3-031-35879-1_3

Taieb, S. B., Bontempi, G., Atiya, A., & Sorjamaa, A. (2011, August). *A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition.* arXiv. Retrieved 2025-11-13, from http://arxiv.org/abs/1108.3259 (arXiv:1108.3259 [stat]) doi: 10.48550/arXiv.1108.3259

Van Deynse, H., Cools, W., De Deken, V.-J., Depreitere, B., Hubloue, I., Kimpe, E., ... Putman, K. (2023, October). Predicting return to work after traumatic brain injury using machine learning and administrative data. *International Journal of Medical Informatics*, *178*, 105201. Retrieved 2026-01-02, from https://www.sciencedirect.com/science/article/pii/S1386505623002198 doi: 10.1016/j.ijmedinf.2023.105201

Vassiley, A., Barratt, T., Dayaram, K., & Burgess, J. (2025, April). Psychosocial workplace hazards and industrial relations: An introduction. *Journal of Industrial Relations*, *67*(2), 177–201. Retrieved 2025-12-24, from https://doi.org/10.1177/00221856251326664 (Publisher: SAGE Publications Ltd) doi: 10.1177/00221856251326664

Vitrano, G., & Micheli, G. J. L. (2024, May). Effectiveness of Occupational Safety and Health interventions: a long way to go. *Frontiers in Public Health*, *12*, 1292692. Retrieved 2025-12-24, from https://pmc.ncbi.nlm.nih.gov/articles/PMC11111875/ doi: 10.3389/fpubh.2024.1292692

Vivian, G. A., Bauder, R. A., & Khoshgoftaar, T. M. (2025, July). A comprehensive survey on machine learning for workplace injury analysis: risk prediction, return to work strategies, and demographic insights. *Journal of Big Data*, *12*(1), 167. Retrieved 2025-09-01, from https://doi.org/10.1186/s40537-025-01229-z doi: 10.1186/s40537-025-01229-z

Wickramasuriya, S. L., Athanasopoulos, G., & Hyndman, R. J. (2019, April). Optimal Forecast Reconciliation for Hierarchical and Grouped Time Series Through Trace Minimization. *Journal of the American Statistical Association*, *114*(526), 804–819. Retrieved 2026-01-10, from https://doi.org/10.1080/01621459.2018.1448825 (Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2018.1448825) doi: 10.1080/01621459.2018.1448825

Williams, A. A., & Marc, J. (2024, December). Traumatic workplace injuries: A cross-sectional analysis of OSHA severe injury reports, including the impacts of seasonality and COVID-19 from 2015 to 2022. *Journal of Safety Research*, *91*, 38–49. Retrieved 2025-12-20, from https://www.sciencedirect.com/science/article/pii/S0022437524000999 doi: 10.1016/j.jsr.2024.08.004

Wilms, H., Cupelli, M., & Monti, A. (2018, July). Combining auto-regression with exogenous variables in sequence-to-sequence recurrent neural networks for short-term load forecasting. In *2018 IEEE 16th International Conference on Industrial Informatics (INDIN)* (pp. 673–679). Porto: IEEE. Retrieved 2025-11-07, from https://ieeexplore.ieee.org/document/8471953/ doi: 10.1109/INDIN.2018.8471953

Yapi, A., Latouche, P., Guillin, A., & Bailly, Y. (2025, June). *A new machine learning framework for occupational accidents forecasting with safety inspections integration.* arXiv. Retrieved 2025-11-06, from http://arxiv.org/abs/2507.00089 (arXiv:2507.00089 [cs]) doi: 10.48550/arXiv.2507.00089

Yuan, C. J., Varathan, K. D., Suhaimi, A., & Ling, L. W. (2023, January). PREDICTING RETURN TO WORK AFTER CARDIAC REHABILITATION USING MACHINE LEARNING MODELS. *Journal of Rehabilitation Medicine*, *55*, 2432. Retrieved 2026-01-11, from https://pmc.ncbi.nlm.nih.gov/articles/PMC9838562/ doi: 10.2340/jrm.v55.2432

Zhang, Y., Schluter, J., Zhang, L., Cao, X., Jenq, R. R., Feng, H., … Zhang, L. (2024, December). Review and revamp of composi-

tional data transformation: A new framework combining proportion conversion and contrast transformation. *Computational and Structural Biotechnology Journal*, *23*, 4088–4107. Retrieved 2026-01-10, from https://www.sciencedirect.com/science/article/pii/S200103702400374X doi: 10.1016/j.csbj.2024.11.003

A APPENDIX

TO ADD: daily resolution; weekly metrics tables; plots of extra states; squarified naics mix viz

Table 11: Hyperparameter configurations for all models

Model	**Hyperparameters**
Ridge Regression	$\alpha = 1.0$, random_state = 0
Lasso Regression	$\alpha = 0.001$, max_iter = 10,000, random_state = 0
Elastic Net	$\alpha = 0.001$, l1_ratio = 0.5, max_iter = 10,000, random_state = 0
PLS Regression	n_components = 10
XGBoost	n_estimators = 300, learning_rate = 0.05, max_depth = 3, subsample = 1.0, colsample_bytree = 1.0, reg_lambda = 1.0, objective = reg:squarederror, random_state = 0, n_jobs = -1
LightGBM	n_estimators = 300, learning_rate = 0.05, max_depth = 3, num_leaves = 7, subsample = 1.0, colsample_bytree = 1.0, objective = regression, reg_lambda = 0.0, random_state = 0, n_jobs = -1, verbose = -1
CatBoost	iterations = 300, learning_rate = 0.05, depth = 3, l2_leaf_reg = 3.0, loss_function = RMSE, random_state = 0, verbose = False

B REQUIREMENTS

Table 12: Software environment and library requirements

Component	**Version**
Python	3.13.5
Visual Studio Code	1.107.0
Zotero	7.0.30
TeXworks	0.6.10
catboost	1.2.8
ipython	9.4.0
lightgbm	4.6.0
matplotlib	3.10.8
numpy	2.4.0
pandas	2.3.3
requests	2.32.5
scikit-learn	1.8.0
scipy	1.16.3
seaborn	0.13.2
statsmodels	0.14.6
xgboost	3.1.2