UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
CAMPUS DE GUARATINGUETÁ

VITOR PINTO RIBEIRO

Bayesian modelling of a decision support system for irrigation

Guaratinguetá
2024

Vitor Pinto Ribeiro

Bayesian modelling of a decision support system for irrigation

Thesis presented to the Postgraduation Program in Mechanical Engineering from São Paulo State University "Júlio de Mesquita Filho", UNESP, Guaratinguetá, for obtaining the PhD title in Mechanical Engineering in the area of Transmission and Conversion of Energy.

Advisor: Prof$^{\text{o}}$ Dr. José Antônio Perrella Balestieri
Co-advisor: Prof$^{\text{o}}$ Dr. Carlos Dias Maciel

Guaratinguetá
2024

Ribeiro, Vitor Pinto
R484b Bayesian modelling of a decision support system for irrigation / Vitor Pinto Ribeiro - Guaratinguetá, 2024.
28 f : il.
Bibliografia: f. 26-28

Tese (Doutorado) – Universidade Estadual Paulista, Faculdade de Engenharia e Ciências de Guaratinguetá, 2024.

Orientador: Prof. Dr. José Antônio Perrella Balestieri
Coorientador: Prof. Dr. Carlos Dias Maciel

1. Evapotranspiração. 2. Inferência (Estatística)
3. Sistemas estocásticos. I. Título.
CDU 519.233(043)

Luciana Máximo
Bibliotecária/CRB-8/3595

POTENTIAL IMPACT OF THIS RESEARCH

As this research focuses on optimising water resource allocation for irrigation, it promotes food security (SDG 2), contributes to clean water goals (SDG 6), and allows for sustainable consumption and production practices (SDG 12). The impact is especially relevant given that several Brazilian socioeconomic factors revolve around agriculture.

IMPACTO POTENCIAL DESTA PESQUISA

Esta pesquisa trata da otimização da alocação de recursos hídricos para irrigação, o que promove segurança alimentar (ODS 2), contribui para objetivos de água potável (ODS 6) e práticas sustentáveis de consumo e produção (ODS 12). O impacto é especialmente relevante dado que vários fatores socioeconômicos brasileiros dependem da agricultura.

unesp

UNIVERSIDADE ESTADUAL PAULISTA
CAMPUS DE GUARATINGUETÁ

VITOR PINTO RIBEIRO

ESTA TESE FOI JULGADA ADEQUADA PARA A OBTENÇÃO DO TÍTULO DE
"DOUTOR EM ENGENHARIA MECÂNICA"

PROGRAMA: ENGENHARIA MECÂNICA
CURSO: DOUTORADO

APROVADA EM SUA FORMA FINAL PELO PROGRAMA DE PÓS-GRADUAÇÃO

gov.br Documento assinado digitalmente
MANOEL CLEBER DE SAMPAIO ALVES
Data: 14/03/2024 09:25:05-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Manoel Cléber de Sampaio Alves
Coordenador

BANCA EXAMINADORA:

gov.br Documento assinado digitalmente
CARLOS DIAS MACIEL
Data: 23/02/2024 10:46:57-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. CARLOS DIAS MACIEL
Coorientador – USP

gov.br Documento assinado digitalmente
MICHEL BESSANI
Data: 23/02/2024 10:55:56-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. MICHEL BESSANI
UFMG

gov.br Documento assinado digitalmente
RICARDO ZORZETTO NICOLIELLO VENCIO
Data: 28/02/2024 15:28:42-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. RICARDO ZORZETTO NICOLIELLO VENCIO
USP

gov.br Documento assinado digitalmente
EDMILSON MARMO MOREIRA
Data: 02/03/2024 08:18:42-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. EDMILSON MARMO MOREIRA
UNIFEI

gov.br Documento assinado digitalmente
PALOMA MARIA SILVA ROCHA RIZOL
Data: 28/02/2024 09:17:54-0300
Verifique em https://validar.iti.gov.br

Prof[a]. Dr[a]. PALOMA MARIA SILVA ROCHA RIZOL
UNESP

FEVEREIRO de 2024

CURRICULUM

VITOR PINTO RIBEIRO

BIRTH DATE	15/07/1994 - Rio Claro / SP
AFFILIATION	Geraldo Pinto Ribeiro Júnior Alessandra Cristina Stefani Baptista Pinto Ribeiro
2009 / 2011	Middle school Colégio Puríssimo Coração de Maria
2009 / 2011	Technician in Information Technology Microcamp
2012 / 2018	Graduate in Computer Engineering University of São Paulo (USP)
2019 / –	Ph.D. in Mechanical Engineering São Paulo State University (UNESP)

Dedicated to those who dare try to break any vicious cycle,
whether one of self-loathing, self-doubt, and/or hopelessness.
It does get better.

ACKNOWLEDGEMENTS

I am thankful to Augusto S. P. B. Ferreira, Guilherme N. Fortes, Rubens C. Toledo, Maísa M. Á. D. Maciel, Victor Ordoñez, and Victor P. Silvano. I appreciate my friendship with every one of you, independent of distance and time together – whether physical or virtual. You have provided me with invaluable support before and during this graduate program. I cherish all the moments and interactions, from existential talks, personal fears, and difficulties; to game nights, discussions over a coffee, and random nonsense on a balcony at dawn. Thanks to all for being a part of what I currently am.

I am deeply thankful to my advisors Prof. José Antônio P. Balestieri and Prof. Carlos D. Maciel, also extended to Prof. Ivonete Ávila. I have no words to express how much your support is invaluable to me as a person and a researcher. Each coffee cup, each discussion, argument, personal story, advice, joke, and lyrics. Each meal, holiday, road trip, event, and reception of me into your homes. The (almost) infinite patience and care towards me. All these interactions are meaningful to me. Maybe most important of all, I am thankful to you for acknowledging and insisting on my potential when even I had no confidence in myself.

I am grateful to *Spectrum Line*, especially Mr Adilson W. Chinatto Jr. and Mrs Cynthia C. M. Junqueira. Their support and infrastructure are essential for this project and my personal development.

I am also grateful to the members of the Examination Board who gracefully agreed to postpone my exam and evaluate this thesis delivered past the accepted time frame.

I want to thank all the faculty and staff of the Faculty of Engineering and Sciences (F.E.G./UNESP) for the infrastructure and opportunities provided, particularly to Prof. Daniela C. Mourão for her assistance as my supervisor during my two years as a Lecturer for undergraduate students.

Finally, I thank UNESP's Innovation Agency (A.U.In.) and the National Council for Scientific and Technological Development (C.N.Pq.) for my research scholarship and support with bureaucratic issues and the prompt answers to my questions.

This work was supported by the National Council for Scientific and Technological Development (C.N.Pq.) - grant number 142390/2019-4;
This work was supported by the Coordination for the Improvement of Higher Education Personnel (C.A.P.E.S.) - financing code 001.

“ Quem sabe eu ainda sou uma garotinha
Esperando o ônibus da escola
Sozinha
Cansada com minhas meias três quartos
Rezando baixo pelos cantos
Por ser uma menina má”
(Cazuza e Frejat)

ABSTRACT

The Sustainable Development Goals are a set of seventeen goals proposed by the United Nations as guidelines to promote the sustainable growth of signatory nations. Such objectives include ones related to food security and safe water availability. Brazil is one of the leading countries in agricultural production, which ties most of its consumptive water usage to farming and irrigation. Most farmers activate their irrigation systems before dawn to save on water, electricity and related fees and require planning on the volume of water to supply the crops with enough to ensure optimal productivity. The main component of water loss is evapotranspiration, a complex and non-linear phenomenon that is a function of climatological variables and crop and soil parameters. Many analytical methods to estimate evapotranspiration require hard-to-obtain data or are approximations with a high region specificity. With the development of new computational tools, many researchers began to employ machine learning and remote sensing to improve the assessment of hydro-climatological variables. Bayesian Network is a class of probabilistic models based on graphs and appropriate to deal with stochastic problems and under scenarios of missing data with the main drawback as devolving into an NP-hard optimisation problem. A probabilistic model based on Bayesian Networks called Inference Diagram provides the toolset to develop support decision systems over stochastic scenarios. The objectives of this project are threefold. First, to implement and validate a bio-inspired algorithm to build a Bayesian Network from data. Second, to create a Bayesian Network capable of estimating evapotranspiration from climatological data even under missing data. And third, to propose an Influence Diagram model that accommodates hydro-climatological stochastic variables and provides suggestions on irrigation scheduling. This project's results achieve the first two objectives, each resulting in a published conference paper, while the third goal requires more development time.

KEYWORDS: Bayesian networks; Probabilistic inference; Influence diagrams; Evapotranspiration.

RESUMO

Os Objetivos de Desenvolvimento Sustentável são um conjunto de dezessete objetivos propostos pelas Nações Unidas como diretrizes para promover o crescimento sustentável das nações signatárias. Tais objetivos incluem aqueles relacionados à segurança alimentar e disponibilidade de água potável. O Brasil é um dos países líderes em produção agrícola, que destina a maior parte de seu consumo de água para agropecuária e irrigação. A maioria dos agricultores ativa seus sistemas de irrigação antes do amanhecer para economizar água, eletricidade e taxas relacionadas e para tanto requerem planejamento no volume de água para abastecer as culturas com o suficiente para garantir a produtividade ideal. O principal componente da perda de água é a evapotranspiração, um fenômeno complexo e não linear que é função de variáveis climatológicas e parâmetros da plantação e solo. Muitos métodos analíticos para estimar a evapotranspiração requerem dados difíceis de obter ou são aproximações com alta especificidade de região. Com o desenvolvimento de novas ferramentas computacionais, muitos pesquisadores começaram a empregar aprendizado de máquina e sensoriamento remoto para melhorar a avaliação de variáveis hidroclimatológicas. A Rede Bayesiana é uma classe de modelos probabilísticos baseados em grafos e apropriada para lidar com problemas estocásticos e sob cenários de falta de dados com a principal desvantagem de recair em um problema de otimização NP-difícil. Um modelo probabilístico baseado em Redes Bayesianas chamado Diagrama de Influência fornece o conjunto de ferramentas para desenvolver sistemas de apoio à decisão sobre cenários estocásticos. Os objetivos deste projeto são três. Primeiro, implementar e validar um algoritmo bioinspirado para construir uma Rede Bayesiana a partir de dados. Em segundo lugar, criar uma Rede Bayesiana capaz de estimar a evapotranspiração a partir de dados climatológicos mesmo sob dados ausentes. E terceiro, propôr um modelo de Diagrama de Influência que acomode variáveis estocásticas hidroclimatológicas e forneça sugestões sobre o regime de irrigação. Os resultados deste projeto atingem os dois primeiros objetivos, cada um resultando em um artigo de conferência publicado, enquanto o terceiro objetivo requer mais tempo de desenvolvimento.

PALAVRAS-CHAVE: Redes bayesianas; Inferência estatística; Diagramas de influência; Evapotranspiração.

LIST OF FIGURES

CONTENTS

1 INTRODUCTION

The Sustainable Development Goals (S.D.G.s) are seventeen goals set forth by the United Nations (U.N.) to serve as guidelines for the signatory countries to achieve sustainable development and growth across several levels of social, political, and economic spheres (United Nations Department for Economic and Social Affairs, 2020).

Figure 1 contains the original panel with the seventeen S.D.G.s. Several macro policies for sustainable growth may result from implementing one or more targets for these goals. Examples include social inclusion, food security, and rational and sustainable use of water resources.

Figure 1 – Panel with the seventeen United Nations' Sustainable Development Goals (S.D.G.s). These goals and their underlying targets and indicators can be used to promote and evaluate sustainable development on several key social, political, and economic factors.

Source: Reproduced from United Nations Department for Economic and Social Affairs (2020).

Brazil has a significant role in the global scenario as a major food producer and exporter. According to the Food and Agriculture Organisation of the United Nations (F.A.O.), in the calendar year 2021, Brazil was the biggest producer in the world of soya beans and sugar cane, second in cattle meat, third in maize and chicken meat and fifth in pork meat (FAO, 2021).

As a result of the country's role as a sizeable agricultural producer – with crops that can serve as livestock fodder –, this sector accounts for the highest consumptive use of water resources, focusing on crop irrigation (Fontenelle; Fuckner; Soares, 2021).

The high consumptive use of water for food production establishes the rational and sustainable use of water resources at the crux for achieving the S.D.G.s referencing this issue and the goals relative to food production and security.

There are multiple methods to control and practice sustainable water usage for irrigation. Examples include policies enforced by fees and limited grants for water extraction, adopting irrigation systems with higher water efficiency, and even rainwater collection and utilisation. One of the largest classes of such methods comprises those that monitor a characteristic associated with a crop's water necessity. These methods try to provide the bare minimum of water required to maintain optimal crop development and production. Dobriyal *et al.* (2012) provides a review on this extensive class of methods.

New methods emerged following the onset of new technologies in data gathering and processing in agricultural settings. Examples include using remote sensing and satellite data alongside artificial intelligence and computer vision algorithms (Singh; Kaur, 2022; Biabi; Mehdizadeh; Salmi, 2019).

The advent of the Internet of Things (I.o.T.) allowed the update of classical methods for estimating soil moisture by connecting simple *in-situ* data sensors to more powerful *ex-situ* computers. Examples include water content assessment from soil permittivity via parallel electrodes on the soil (Wu *et al.*, 2022) and ground matric potential via tensiometers (Abdelmoneim *et al.*, 2023).

To comply with regulations and policies that enforce variable water and electricity fees according to the time of day, most farmers exclusively activate their irrigation system before dawn. As such, the amount of water used in each irrigation has to consider the volume lost to the environment during the following day. The four major water loss contributors are percolation, runoff, evaporation from wet soil and plants, and crop transpiration (Anayah; Kaluarachchi, 2019).

Evapotranspiration (ET) is the metric that combines evaporation and transpiration from a cultivated area and is the main component of water loss in agriculture. This element is difficult to measure, non-linear, complex and varies from different crops and within the environment. Despite these factors, a metric of reference evapotranspiration (ET_0) based on a standardised area with a grassy crop provides an initial estimate of evapotranspiration (Allen *et al.*, 1998).

The observance of this element's importance precedes wide computer adoption. This historical precedence led to the proposition of several analytical methods for ET_0 estimates based on regression models and available climatological variables (Hargreaves; Samani, 1985; Linacre, 1977; Benavides; Lopez, 1970; Turc, 1961). Many approximations rely different climatological data sets, and most provide a reliable ET_0 estimate only within a region (Ghiat; Mackey; Al-Ansari, 2021).

Given the wide range of methods to estimate ET_0, the F.A.O. recommended the Penman-Monteith method. This analytical approximation provides a daily estimate based on several climatological variables, some difficult to obtain (Allen *et al.*, 1998). As such, the community considers the Penman-Monteith the gold standard but still employs other methods that rely on fewer variables (Lu *et al.*, 2005).

The advance of hydro-meteorological fields and the development and mass adoption of machine-learning techniques resulted in several computational algorithms to estimate ET_0, most based on deterministic methods such as Neural Networks and Evolutionary Algorithms (Jayashree *et al.*, 2023; Amirashayeri *et al.*, 2023; Lucas *et al.*, 2020; Granata, 2019).

Another class of computational modelling suitable to problems involving stochastic data – such as climatological variables – is the Bayesian Network (Mitchell *et al.*, 2021). This exemplar of a probabilistic model based on graphs encompasses the stochastic relationships between variables in a system, aided by conditional probability distributions (Ianishi *et al.*, 2020).

The Bayesian Network design provides features adequate to estimate ET_0. One is its interpretability, as persons with no computational background can evaluate the resulting model (Hui *et al.*, 2022). Another is its ability to perform inference tasks even under missing data scenarios (IliĆ *et al.*, 2022).

The main drawback of this kind of model is the graph learning task. Finding a graph that has the best adherence to the data devolves into an NP-hard problem with a super-exponential search space (Gross *et al.*, 2019). There are heuristics and algorithms which tackle this issue and provide suitable structures without an optimality guarantee (Dai *et al.*, 2018; Natori *et al.*, 2015).

The decision to activate the irrigation system and for how long is usually based on human judgement, despite rain and evapotranspiration being natural and stochastic phenomena. This action may consider the climatological data, an estimate of soil moisture, the financial cost of water and electrical resources, and even the opportunity cost of not providing water but having a worse harvest.

Therefore, there is interest in computational methods that can accommodate an agent's ability to perform actions based on stochastic variables and any cost attribution – financial or from convenience. One such method is the Influence Diagram, an extension of Bayesian Networks which integrates the cost of actions from a set of agents over a stochastic scenario (Hansen, 2021).

This probabilistic model provides the optimal combination of actions to maximise a set of utility functions under scenarios of uncertainty. Decision support systems can leverage this characteristic to suggest a course of action for the user (Hughes *et al.*, 2021).

This thesis advocates for the adequacy of Bayesian methods to deal with such non-linear and non-stationary phenomena. This essay proposes, implements and discusses procedures related to the framework of a probabilistic decision support system, broadly divided into three components: Bayesian Network structural learning based on synthetic and non-synthetic data, Bayesian model expansion and refinement based on real-world non-linear and non-stationary phenomena, and stochastic decision-making recommendation based on climatological factors alongside criteria from expert knowledge.

This thesis' objectives are: first, implementing a bio-inspired algorithm to obtain a plausible graph structure from a Bayesian Network exclusively from data. This method represents a scientific contribution due to the innate super-exponential nature of the graph search space.

Second, to build a Bayesian Network based on climatological data from five cities across the Brazilian MATOPIBA region (which comprises the states of Maranhão, Tocantins, Piauí and Bahia). This objective is relevant to evaluate this toolset's capabilities in modelling real-world phenomena based on real-world data.

Third, evaluate a technique based on Bayesian regression to obtain the most relevant random variables for building a Bayesian Network. This selection process is crucial to providing the fewest random variables to a Bayesian Network's structural learning algorithm due to the latter's super-exponential nature.

Fourth, employ the Bayesian framework to model a climatological phenomenon based on agrometeorological data and assess the adherence of the model to the expected real-world dynamics, both via analytical comparison to the literature's gold standard and also to indirect in-situ measures via tensiometers reading.

Fifth, leverage the Bayesian regression and Bayesian Network models to provide evapotranspiration estimates under scenarios where one may lack some of the required climatological data. This approach may provide an adaptability layer to estimate evapotranspiration, even absenting mandatory climatological data for the analytical methods.

Lastly, propose and evaluate an Influence Diagram which provides the best course of action for irrigation based on observed stochastic components and under different water requirement scenarios.

Chapter 2 presents a brief literature review for stochastic evapotransporation estimation methods. Chapter 3 contains the theoretical basis for the subjects discussed throughout this thesis. Chapter 4 presents a detailed description of the developed methods and the data sets - synthetic and non-synthetic - used for all the experiments. Chapter 5 provides the results obtained during this thesis execution and the relevant discussions. Chapter 6 closes the main document with the conclusions and final considerations. Additionally, Appendices A through C provide extra data supporting some findings in the main document, and Annexes A through C contain the integral version of the published papers derived from this research.

2 LITERATURE REVIEW

This Chapter presents a brief review of the scientific literature regarding the usage of stochastic Bayesian methods to estimate the evapotranspiration metric. Although the final product of this thesis is a recommender system, a review of evapotranspiration methods is more pertinent due to this metric's usage in other fields related to climatology (such as drought prediction.)

Section 2.1 provides the framework for this review, including keywords, scientific indexing base and bibliometric data to ensure reproducibility.

Section 2.2 contains a general analysis regarding evapotranspiration's role within the papers' overall scientific contribution. This first filter disjoints the articles according to their focus to present a novel or improved approach to determine evapotranspiration or employ this metric in a broader context.

Section 2.3 discusses the stochastic toolset used in the papers focusing on evapotranspiration methods. The main goal is to provide an overview regarding the popularity of the frequentist versus the Bayesian statistics approach employed throughout the literature.

Lastly, Section 2.4 concludes the Chapter with a brief discussion about the scientific gap identified based on this short review. Its purpose is to highlight the scarcity of Bayesian-focused methods to estimate evapotranspiration.

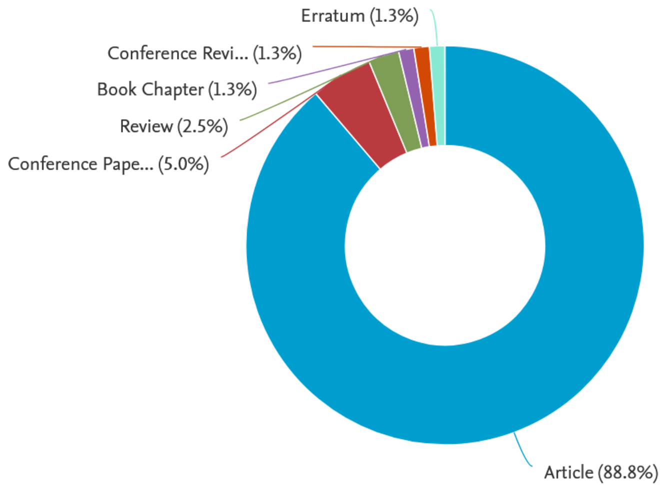
2.1 BIBLIOMETRIC OVERVIEW

The interdisciplinary nature of this research requires a literature review on an indexing base with high multidisciplinarity regarding its sources. Elsevier's Scopus platform presents such characteristics by indexing over 45000 sources across several disciplines. In addition, the Scopus indexing base is often the source for scientific studies based on systematic, bibliometric and bibliographical analysis.

To explore the employment of tools based on Bayesian statistics in modelling evapotranspiration, the author employed the following keyword combination as the main query: `evapotranspiration AND (Bayesian OR stochastic OR probabilistic) AND (computational OR algorithm OR "artificial intelligence")`. This query across the `Article title, Abstract, Keywords` search field results in 131 occurrences, while filtering papers from 2019 onwards and available in English reduces the final number to 80 scientific articles.

The first relevant bibliometric data from these articles is their source type, which allows a discussion regarding the subject's maturation within the scientific community. Figure 2 presents the source type's distribution for the 80 articles provided by the Scopus platform. The high concentration of scientific papers classified as "article" suggests a high degree of maturity, as these works' contributions are usually past the novel and fast-evolving dynamics of "conference papers" but not so consolidated as those published as "book chapters".

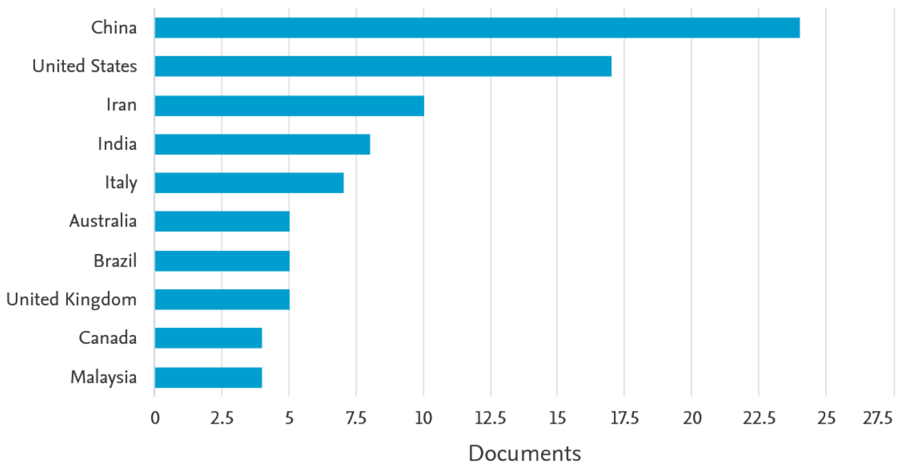
Figure 2 – Distribution by document type.



Source: Extracted from Scopus.

Another relevant bibliometric discussion revolves around the countries listed as the papers' sources. Figure 3 presents the countries' distribution for the 80 articles, with China and the United States as most often represented. Interestingly, according to Figure 3, Iran is the third most represented country, contributing ten articles. This publication volume within the water management context may come from Iran's hydrological characteristics and its scarcity of fresh water.

Figure 3 – Distribution by country/territory.



Source: Extracted from Scopus.

Based on the results in Figure 3, India's contribution as the fourth most represented country may reflect its researchers' interest in maintaining India's food security and water availability. These topics represent a crucial subject within India's context due to its high populational density and dependency on neighbouring countries to access most of its freshwater resources.

These two bibliometric data point to the scientific interest in the evapotranspiration metric aided by a stochastic toolset in recent years. Regardless of the time window since this metric's first rigorous modelling to the present, the discussions and contributions based on evapotranspiration are perennial due to its close relation to food security and water management subjects in an ever-changing climate scenario.

2.2 EVAPOTRANSPIRATION AS SECONDARY COMPONENT

Of the 80 articles, 48 employ evapotranspiration as a secondary metric to assess a correlated event or scenario. The most common uses for this metric are in ecological assessment, crop yield prediction and drought modelling, each category with nine, eight and eight articles, respectively.

Evapotranspiration represents a relevant factor in the water cycle and, subsequently, in ecosystem balance. Some works employ evapotranspiration as a proxy measure to evaluate climate change impact from different standpoints. For instance, Silva *et al.* (2023) estimates climate change's effect on reference evapotranspiration in the Brazilian semi-arid region, and Baltensperger e Joly (2019) employs this metric to analyse changes in migratory patterns in Caribous populations in the Arctic.

Evapotranspiration is also a direct proxy to evaluate changes in large bodies of water, as described by Radmanesh, Esmaeili-Gisavandani e Lotfirad (2022) and Wang *et al.* (2023a). Both Ghaseminejad e Uddameri (2020) and Durighetto, Bertassello e Botter (2022) employ evapotranspiration as a parameter to simulate hydrological dynamics, and Quilty *et al.* (2023) presents a prediction framework based on such dynamics.

Lastly, from a financial point of view, Byers *et al.* (2020) discusses climate change's impact on electricity prices while Tina e Nicolosi (2023) approaches issues regarding water availability in hydroelectric electricity generation.

As evapotranspiration is an indirect measure of crops' consumptive water usage, some works utilise this metric to predict or maximise crop yield. Recent applications include studies based on barley (Zarei; Mahmoudi; Shabani, 2021), maise (Chidzalo; Ngare; Mung'Atu, 2022), rice (Majumdar *et al.*, 2023) and tomatoes (Bazrafshan *et al.*, 2022), which are some of the most important crops to food security. These studies encompass both greenhouse crops (Kocian *et al.*, 2020a; Kocian *et al.*, 2020b) and open-field ones (Chidzalo; Ngare; Mung'Atu, 2022; Rossit *et al.*, 2021).

From a drought perspective, many works use the Standardised Evapotranspiration Precipitation Index (SPEI) as an indicator to evaluate or estimate drought events. Studies in this context relate to studying the evolution of past drought events, such as those in Ullah *et al.* (2023) and Barzkar, Najafzadeh e Homaei (2022), and to prediction/forecast of these, as in Dayal, Deo e Apan (2019) and Kolluru e Kolluru (2021). There are also papers related to drought monitoring (Jehanzaib *et al.*, 2023; Gavahi *et al.*, 2020; Chen *et al.*, 2023), which may support the public sector in its policy-making decisions.

2.3 EVAPOTRANSPIRATION ESTIMATE AS THE MAIN FOCUS

The remaining 32 articles focus on evapotranspiration estimates as their contribution. Despite the numbers, most of these articles fall into two categories: ensemble studies or hybridisation improvements on already established computational methods.

In a broad sense, articles based on ensembles evaluate how to leverage the output of different artificial intelligence methods and their combination to improve evapotranspiration estimates. As the most cited paper in this category, Nourani, Elkiran e Abdullahi (2020) studies evapotranspiration ensemble-based predictions using Adaptive Neuro-Fuzzy Inference System (ANFIS), Artificial Neural Network (ANN), Multiple Linear Regression (MLR), and Support Vector Machine (SVM), all classical and well-established computational methods.

One of the methods most commonly employed on papers based on ensemble methods is the ANN and its variants (Talebi; Samadianfard; Kamran, 2023; Nourani; Elkiran; Abdullahi, 2020; Yin *et al.*, 2021; Walls *et al.*, 2020; Alibabaei; Gaspar; Lima, 2021; Lucas *et al.*, 2020; Ogunrinde *et al.*, 2020). This method is well-established in the literature due to its robust theoretical foundations and their variations (such as convolutional networks, deep ANN and ANFIS) and their applications in several fields with remarkable results.

The papers in the second category, comprised of hybridisation techniques, commonly rely on tree-based methods as their fundamental method. This approach employs some heuristics to improve the original method's convergence, including Gradient Boosting variations such as Natural Gradient Boosting and eXtreme Gradient Boosting (Basagaoglu; Chakraborty; Winterle, 2021; Wang *et al.*, 2023b; Zhao *et al.*, 2024).

Another popular hybridisation technique relies on Bayesian optimisation (Zhao *et al.*, 2024), where the parameters for the model's next iteration are those maximising the parameters' posterior distribution conditioned to the data. Wang *et al.* (2023b) presents an approach based on Bayesian optimisation alongside a Gradient Boosting tree-based model.

There is a small number of papers proposing novel approaches for evapotranspiration estimates. For instance, Smallman e Williams (2019) describes a method based on a terrestrial ecosystem model and several environmental and phytomorphological components to estimate photosynthesis and evapotranspiration. The medium-scale model provides a reasonable estimate within the models' constraints and required computational time.

Another novel approach is the fully Bayesian method proposed by Ribeiro *et al.* (2023), where the authors employ Bayesian statistics for evapotranspiration estimates. The Bayesian approach allows for a seamless incorporation of uncertainties and non-linearity in climatological time series.

2.4 CONCLUSION

Despite evapotranspiration's contribution to the water cycle and food production, few recent articles model this phenomenon from a mostly probabilistic point of view (Basagaoglu; Chakraborty; Winterle, 2021). This lack of research is especially relevant due to the intrinsic stochastic nature of climatological elements, alongside the well-established role of probabilistic methods in other fields with similar characteristics, such as such as medicines (Elmer *et al.*, 2022) and power planning (Shedbalkar; More, 2022).

From the 80 total articles, only three provide models derived from statistics: Chidzalo, Ngare e Mung'Atu (2022) presents a model to estimate maise yield based on copulas (a frequentist technique), and Khashei-Siuki *et al.* (2021) uses a similar toolset for evapotranspiration estimates. Ribeiro *et al.* (2023) is the only paper that employs a toolset based on Bayesian statistics.

The discussions and analysis in this Chapter provide an initial argument for the lack of evapotranspiration estimates – and adjacent applications – based on Bayesian statistics.

3 THEORY

This Chapter contains the theoretical foundation required to understand the terminology and discussions employed throughout this document.

Sections 3.1 and 3.2 define the evapotranspiration phenomenon and classical analytical models, alongside the definition and models for complementary factors relating to evapotranspiration dynamics.

Sections 3.3, 3.4 and 3.5 present the foundation for the Bayesian models employed in this thesis. Section 3.3 describes the Bayesian regression technique; Section 3.4 addresses Bayesian Network theoretical foundations; and Section 3.5 builds upon the Bayesian Network framework to define Influence Diagrams.

Section 3.6 describes the original Physarum Learner, a bio-inspired algorithm developed to provide a Bayesian Network structure from discrete data.

To standardise the mathematical notation, elements in **boldface** represent a set of items, while *italicised* elements represent singular items.

3.1 REFERENCE EVAPOTRANSPIRATION

Estimating the amount of water evaporated and transpired by the culture and evaporated from the soil is vital to determine the irrigation volume required for optimal production. However, it is difficult—or in some cases, practically impossible—to measure directly due to various climatological factors.

In light of the significance of evapotranspiration, the literature offers diverse analytical methods (Hargreaves; Samani, 1985; Linacre, 1977; Turc, 1961; Benavides; Lopez, 1970) to estimate it from climatological variables based on various environmental conditions (Ghiat; Mackey; Al-Ansari, 2021). These traditional methods utilise regression analysis to model and extrapolate the evapotranspiration metric conditioned on the availability of specific climatological variables, such as the Hargreaves–Samani equation, which relies solely on mean temperature and extraterrestrial solar radiation (Hargreaves; Samani, 1985). Another instance is the Linacre method, calibrated according to the Australian climate (Linacre, 1977).

The Food and Agriculture Organization (Allen *et al.*, 1998) has designated the Penman–Monteith equation as the gold standard for estimating standardised evapotranspiration for a grass crop (vegetation height = 0.12 m, albedo = 23%) on soil with a fixed surface resistance of 70 $S \cdot m^{-1}$ without water restrictions. The original Penman–Monteith equation is presented here as Equation 1, which estimates the ET_0 reference evapotranspiration in millimetres per day.

$$ET_0 = \frac{0.408\Delta(R_n - G) + \gamma\left(\frac{900}{T_{mean} + 273}\right)U_2(e_s - e_a)}{\Delta + \gamma(1 + 0.34U_2)} \tag{1}$$

The Penman–Monteith equation necessitates several climatological measurements, including net solar radiation (R_n), mean daily temperature (T_{mean}), on-site wind speed at 2 m (U_2), and actual (e_a) and saturation (e_s) vapour pressure. The slope of the vapour pressure curve (Δ), the soil heat flux density (G), and the psychrometric constant (γ) are additional parameters.

Despite its high correlation with the observed value for the evapotranspiration metric, the primary limitation of this equation is its dependence on several climatological factors. Weather stations that can automatically provide the required data for the Penman–Monteith equation are unattainable to most rural producers, resulting in this method being unfeasible to estimate (Cruz; Marques, 2023).

One analytical approach employed throughout the literature to estimate daily evapotranspiration based on fewer climatological variables is the Hargreaves–Samani method (Hargreaves; Samani, 1985), presented here as Equation 2. This method requires the mean, maximum, and minimum daily temperatures—respectively T_{mean}, T_{max}, and T_{min}—as well as an estimate for the extraterrestrial incident solar radiation (R_a). Another common analytical approach is the Benavides and Lopez method (Benavides; Lopez, 1970), presented as Equation 3, whose equation only requires data on the mean daily temperature (T_{mean}) and mean daily relative humidity (RH_{mean}).

$$ET_{0Hg} = R_a \cdot 0.408 \cdot 0.0023\,(T_{mean} + 17.8)\,(T_{max} - T_{min})^{0.5} \tag{2}$$

$$ET_{0BL} = 1.21 \cdot 10\left(\frac{7.42T_{mean}}{234.7 + T_{mean}}\right)(1 - 0.01RH_{mean}) + 0.21T_{mean} - 2.30 \tag{3}$$

The variety of phytomorphology properties for crops used for plantations implies a high variance for the water amount evaporated and transpired by the plants. For this reason, and to provide a generic framework to estimate evapotranspiration under scenarios of different cultures and development stages, the analytical equations—including the Penman–Monteith approach—provide a standardised reference evapotranspiration ($ET_0 \triangleq ET$). To estimate the evapotranspiration value for a given culture and development stage, one can use the culture coefficient (k_c) to obtain the crop evapotranspiration (ET_c) according to Equation (4).

$$ET_c = k_c \cdot ET_0 \tag{4}$$

Throughout this text, the word "evapotranspiration" refers to the reference evapotranspiration $ET_0 \triangleq ET$ unless stated otherwise.

3.2 CLIMATOLOGICAL WATER BALANCE AND WATER REQUIREMENTS

The water balance for a given region obeys the mass conservation principle summarised by Equation (5), which describes the equilibrium of water inflows and outflows from a control volume (Silva *et al.*, 2021).

$$P = ET + Q + \Delta S \tag{5}$$

In Equation 5, the water inflow from precipitation (P) balances the outflows from evapotranspiration (ET) and surface and subsurface runoff (Q), promoting a variation in soil moisture (ΔS).

When considering a plantation scenario of groundwater flow equilibrium, a linear or exponential function describes the surface runoff as a function of precipitation (Raghunath, 2006). As such, Equation (6) describes the soil moisture dynamics as a function of evapotranspiration and precipitation.

$$P = ET + f(P) + \Delta S \tag{6}$$

Assessing soil water content plays a fundamental role in establishing the water available to crops, which impacts plantation production. The water amount required for optimal production depends on the culture, its development stage, and efforts to minimise water waste.

A simple approach to estimating the water content in a given region relies on climatological data for precipitation and evapotranspiration, the so-called climatological water balance, to estimate the available water deficit (awd), the percentage of water stored in soil (st), and the net water balance (Mutti *et al.*, 2022). Algorithm 1 presents the heuristics to assess the awd and st metrics solely from the climatological factors.

Algorithm 1 Algorithm to estimate the accumulated water deficit $curr\_awd$ and water stored $curr\_st$ based on evapotranspiration ET, precipitation P, and the previous values of these estimates.

Require: ET, P, $prev\_awd$, $prev\_st$
if $ET > P$ **then**
  $curr\_awd \leftarrow prev\_awd + (P - ET)$
  $curr\_st \leftarrow 100 \cdot e^{curr\_awd/100}$
else
  $curr\_st \leftarrow prev\_st + (P - ET)$
  if $curr\_st > 100$ **then**
    $curr\_st \leftarrow 100$
  end if
  $curr\_awd \leftarrow 100 \cdot e^{curr\_st/100}$
end if
  return $curr\_awd$, $curr\_st$

Given all these nuances, several methods estimate soil water content based on readings of low-cost sampling equipment, such as tensiometers and capacitors, to provide decision-support for the irrigation responsible (Saha; Sekharan; Manna, 2020). One of these methods is the van Genuchten equation, which estimates soil water content for a specific soil (θ, in percentage) based on soil moisture content estimated by suction pressure (ψ, in kPa) and experimental coefficients (Genuchten, 1980). Equation (7) provides the generic expression for the van Genuchten equation.

$$vG(\psi) = \theta_r + \frac{\theta_s - \theta_r}{(1 + (\alpha|\psi|)^n)^m} \tag{7}$$

where θ_s is the saturated soil water content or field capacity, and θ_r is the residual soil water content or wilting point under the physical limitations of the tensiometers. The parameters

α [cm^{-1}], n, and $m = 1 - 1/n$ are constants obtained experimentally based on the soil's characteristic water retention curve.

3.3 BAYESIAN REGRESSION

Regression analysis is a tool used to estimate the influence of a set of input variables over an output (response) variable (McElreath, 2020). Let y be the target variable and $X = (X_1, X_2, ..., X_n)$ be the set of components (or covariates) that may contribute to the values of the outcome. Equation (8) presents a multiple linear regression of y as a linear combination of elements X_i plus a constant value as the series' DC component. The main task of this regression is to estimate $A = \{a_0, a_1, ..., a_n\}$ as the set of regression parameters from the data set.

$$y \approx a_0 + \sum_{i=1}^{n} a_i X_i + \varepsilon \tag{8}$$

A random variable usually assumed with a Gaussian $\mathcal{N}(0, \sigma^2)$ distribution models the regression error term ε (non-observed) in Equation (8). Under the frequentist approach, the estimators of the regression parameters $a_0, a_1, ..., a_n$ are usually obtained using the minimum squares method, i.e., the set of coefficients that minimises the square sum of the errors (Draper; Smith, 1981). The variance parameter σ^2 should also be estimated from the data.

Under a Bayesian approach (Bayesian regression model), we assume the same regression model from Equation (8), while considering Gaussian errors, with the regression parameters β_i representing random quantities with specified probability density functions—also denoted as prior distributions—with hyperparameters φ_i. These hyperparameters may be univariate or multivariate, whose values are usually known, as represented in Equation (9). If the hyperparameters are unknown, which comes from a hierarchical Bayesian approach, each hyperparameter also requires—over itself—a prior distribution.

$$y \approx \beta_0(\varphi_0) + \sum_{i=1}^{n} (\beta_i(\varphi_i) \cdot X_i) + \varepsilon \tag{9}$$

Bayes' theorem provides the means to obtain the posterior distributions for the regression parameters β_i (Gelman; Hill; Vehtari, 2020). Let D represent the data set, $\pi(B)$ represent the joint prior distribution for the vector of regression parameters, $B = (\beta_0, \beta_1, ..., \beta_n)$, and $L(B \mid D)$ represent the likelihood function for B based on the data D. Under the assumption of Gaussian distribution for the error ε in Equation (9), the joint posterior distribution for B, denoted as $\pi(B \mid D)$, is given by the combination of the prior distribution $\pi(B)$ with the likelihood function $L(B \mid D)$, as described in Equation (10). The integral in the denominator represents multiple integrals with dimension *dim(B).*

$$\pi(B \mid D) = \frac{\pi(B)L(B \mid D)}{\int_B \pi(B)L(B \mid D)dB} \tag{10}$$

The integral in Equation (10)'s denominator represents multiple integrals with dimension *dim(B)*, which presents the main drawback of the Bayesian regression analysis. An analytical approach to estimate the term $\int_B \pi(B)L(B \mid D)dB$ usually becomes unfeasible.

Popular approaches extensively used in applied Bayesian analysis are Markov Chain Monte Carlo (MCMC) method, such as Gibbs sampling or Metropolis–Hastings algorithms, to simulate samples for the joint posterior distribution of interest (Gelman; Hill; Vehtari, 2020). The MCMC method generates values from each conditional posterior distribution $\pi(\beta_i \mid \beta_{(i)}, D)$, where $\beta_{(i)}$ denotes the vector of all parameters, except β_i, and is available in existing Bayesian software.

From the simulated samples of the joint posterior distribution of interest, we obtain the posterior summaries of interest, such as Monte Carlo point Bayesian estimates for the parameters of the model (usually the posterior means) or 95% credible intervals for the parameters of interest.

The authors acknowledge the intricacies and hardships of the MCMC methods, such as chain convergence, the sampler chosen, and initial values for the parameters in the iterative method. These characteristics, albeit relevant, lie outside the scope of the present work and were omitted. Please refer to reference (Doucet; Wang, 2005) for more details on MCMC methods.

3.4 BAYESIAN NETWORKS

Bayesian Networks (B.N.s) are a class of Probabilistic Graphical Models. The role of a B.N. is to establish stochastic associations between random variables of a data set observed from a system and present these relationships in an easy-to-understand format. A B.N. may represent the knowledge from a body of specialists on the system, be built exclusively from system data, or both (Pearl, 2000).

The first element that comprises a B.N. is a Directed Acyclic Graph (D.A.G.) $G = (\mathbf{V}, \mathbf{E})$. The vertices set $\mathbf{V}$ contains one vertex for each random variable from the observed system; for a system with n discrete variables, $\mathbf{V} = \{v_i, i = 1, 2, \ldots, n\}$ (Pearl, 1988). In this context, a random variable describes a variable whose values follow a probability distribution.

Given two vertices v_i and v_j, $i \neq j$, a directed edge from v_i to v_j represents an association between these variables. The element $e_{ij} \in \mathbf{E}$ defines this edge, and the set $\mathbf{Pa}(v_j) \subset \mathbf{V}$ contains all vertices that have an edge which ends in v_j, referred as the set of parents to v_j.

From a system standpoint, this interaction between variables corresponds to a relationship where a value from v_i may associate with a state from v_j, or vice-versa. These edges do not necessarily represent causal relationships, just observational associations acquired from the data set (Pearl; Mackenzie, 2018).

Each association relationship established by the elements in $\mathbf{E}$ formally corresponds to a conditional probability function over the values from a variable and those from its parents. Let $\mathbf{S}_j$ represent the finite and discrete set of possible states for each vertex $v_j \in \mathbf{V}$, and $\mathbf{Pa}(v_j) \neq \varnothing$. Then, this representation describes that each random variable from the system follows a probability distribution function of its states and conditioned to the states of its parents, presented formally as eq. 11.

$$v_j \sim P\left(\mathbf{S}_j | \mathbf{Pa}(v_j)\right) \tag{11}$$

Each random variable is assumed to be discrete and sampled from a bounded system, which makes enumerable the set of states for every vertex in G. This characterisation and the relationship provided in eq. 11 allow for a tabular representation of the probability for all combinations of values from a variable conditioned to every permutation of its parents' values, defined as a Conditional Probability Table (C.P.T.). Each vertex $v_i \in \mathbf{V}$ has a C.P.T. attached to it, denoted as θ_i, and the set of all C.P.T.s $\mathbf{\Theta} = \{\theta_i, i = 1, 2, \ldots, n\}$ is the second element required to define a Bayesian Network (Pearl, 1988).

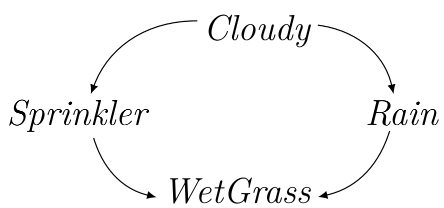
As an example of a Bayesian Network, Figure 4 contains the graphical structure and C.P.T.s for the *Sprinkler* data set (Murphy, 1998). This example consists of a system with four binary random variables (*Cloudy*, *Sprinkler*, *Rain*, and *WetGrass*), and the association between these variables, such as the assumption that a cloudy sky influences the probability of rain (edge *Cloudy* $\rightarrow$ *Rain*).

The B.N. in Figure 4 also presents the set of all C.P.T.s, which describe the probability of observing an outcome for a variable given any combination of states from its parents. For instance, both the outcomes for the events of the sprinkler being on or off and the occurrence of rain exert influence over the probability distribution for the grass being wet or dry.

Figure 4 – Complete representation of the Bayesian Network for the toy problem *Sprinkler*, presented in Murphy (1998). This network consists of four binary variables, the association relationships between those, and explicit C.P.T.s for each node.

Cloudy = F	Cloudy = T
0.5	0.5

Cloudy	Sprinkler = F	Sprinkler = T
F	0.5	0.5
T	0.9	0.1



Cloudy	Rain = F	Rain = T
F	0.8	0.2
T	0.2	0.8

Sprinkler	Rain	WetGrass = F	WetGrass = T
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

Source: Adapted from Murphy (1998).

A B.N. model also provides the ability to perform stochastic inferences over its structural elements based on the outcome with the *maximum a posteriori* probability. An inference query evaluates the most probable values for a set of variables provided the observed states for some or all other variables on the network, respectively referred to as incomplete or complete evidence (Koller; Friedman, 2009).

Let v_i be a target variable whose possible states are $\mathbf{S}_i$, and $\mathbf{\Xi} \subseteq \mathbf{V} \setminus \{v_i\}$ represent the set of observed values. Let ξ be an instance for the observed values. A *maximum a posteriori* query for v_i given evidence ξ returns the state $s_M \in \mathbf{S}_i$ that maximises the probability distribution over v_i conditioned to the observed occurrences in ξ. Equation 12 provides the mathematical representation for this operation.

$$s_M = \operatorname*{argmax}_{s \in \mathbf{S}_i} P(v_i = s | \mathbf{\Xi} = \xi) \tag{12}$$

From the *Sprinkler* B.N. in Figure 4, one may perform a query as "What is the most probable state of the lawn, observed *Sprinkler* $= T$, *Cloudy* $= F$, and *Rain* $= F$?". Given this complete evidence over the system, the query identifies the state for the target variable $WetGrass = s_M \in \{T, F\}$ which corresponds to the maximum probability according to eq. 13. Under this scenario, the most probable outcome is $s_M = T$ with a probability of 90%, but there is also a 10% chance for dry grass.

$$s_M = \operatorname*{argmax}_{s \in \{T,F\}} P(WetGrass = s | Sprinkler = T, Cloudy = F, Rain = F) \tag{13}$$

A Bayesian Network $B = (G, \mathbf{\Theta})$ describes the joint probability distribution from the model over a data set (Pearl, 2000). The Bayes' Rule defines this distribution from the elements in B as presented in eq. 14.

$$P(v_1, v_2, \ldots, v_n) = \prod_{v_i \in \mathbf{V}} P\left(\mathbf{S}_i | \mathbf{Pa}(v_i)\right) \tag{14}$$

Absent from the data, any D.A.G. with an appropriate number of vertices may represent the associations between a system's variables. The task of finding a suitable network B whose joint probability function adjusts over a data set D devolves into a graph optimisation problem (Broom; Do; Subramanian, 2012).

Let $B^* = (G^*, \mathbf{\Theta}^*)$ describe the structure which provides the best joint probability distribution over the data set. Equation 15 formally describes this optimisation problem.

$$B^* = \operatorname*{argmax}_{B} P(B|D) \tag{15}$$

According to Bayes' rule, and considering a uniform distribution for $P(D)$, eq. 15 expands into eq. 16.

$$B^* \propto \operatorname*{argmax}_{B} P(D|B)P(B) = \operatorname*{argmax}_{G,\mathbf{\Theta}} P(D|G, \mathbf{\Theta})P(\mathbf{\Theta}|G)P(G) \tag{16}$$

Equation 17 provides a marginalisation of the right-hand side of eq. 16 by the distribution over $\mathbf{\Theta}$.

$$\operatorname*{argmax}_{G} \int_{\mathbf{\Theta}} P(D|G, \mathbf{\Theta})P(\mathbf{\Theta}|G)d\mathbf{\Theta} = \operatorname*{argmax}_{G} P(D|G) \tag{17}$$

After multiplying both sides of eq. 17 by a factor of $P(G)$ which is independent from $\mathbf{\Theta}$, follows eq. 18.

$$\underset{G}{\operatorname{argmax}} P(G) \int_{\mathbf{\Theta}} P(D|G, \mathbf{\Theta}) P(\mathbf{\Theta}|G) d\mathbf{\Theta} = \underset{G}{\operatorname{argmax}} P(D|G) P(G) \tag{18}$$

Bayes' theorem provides an estimative for the distribution of $P(G|D)$ from the right-hand side of eq. 18, presented in eq. 19.

$$\underset{G}{\operatorname{argmax}} P(D|G) P(G) \propto \underset{G}{\operatorname{argmax}} P(G|D) \tag{19}$$

Finally, when considering a uniform distribution over $P(G)$, eqs. 16 and 19 combine into eq. 20.

$$B^* \propto \underset{B}{\operatorname{argmax}} P(D|B) P(B) \propto \underset{G}{\operatorname{argmax}} P(D|G) \tag{20}$$

The right-hand side of eq. 20 suggests that the optimal Bayesian Network B^* is the one whose graph G best fits the data set D. The literature presents several fitness functions that evaluate the adherence of a graph over a data set from the Bayesian perspective, such as the *Minimum Description Length* (MDL) (Suzuki, 1993), the *Bayesian Information Criterion* (BIC) (Schwarz, 1978), and the *Bayesian Dirichlet equivalence with uniform prior metric* (BDeu) (Buntine, 1991).

The issue of graph optimisation belongs to the NP-hard class, and the space containing all graph possibilities is super-exponential over the number of vertices. Researchers rely on heuristic methods that provide a good candidate graph within a reasonable time with no optimality guarantee (Gross *et al.*, 2019).

3.5 INFLUENCE DIAGRAMS

An Influence Diagram (I.D.) is another member of the probabilistic models based on graphs and is an extension of the Bayesian Network. This modelling also presents probabilistic associations between variables based on D.A.G.s and conditional probability functions (Koller; Friedman, 2009).

The increment over the B.N. modelling consists of two new kinds of nodes on the I.D. structure, which provides a complimentary set of functions to the base model. The first new type is a decision node, each representing an action that an external agent exerts over the system, preferably implemented over a quantitative scale.

The second type represents a utility function. Each function provides a utility value for every possible scenario involving a subset of random variables and actions of the external agent.

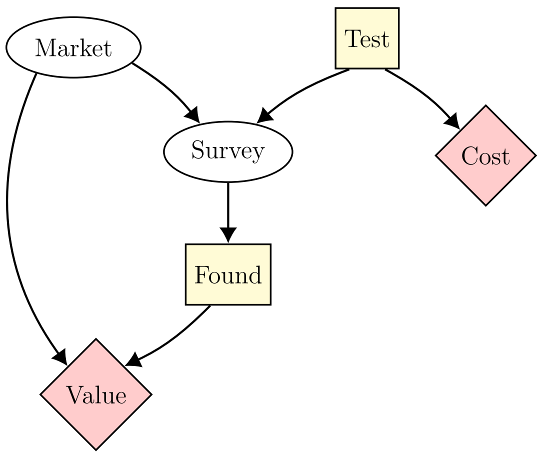
Let $\chi = \{\mathbf{V}, \mathbf{E}, \mathbf{\Theta}, \mathbf{D}, \mathbf{U}\}$ represent an I.D. based on the notation from Section 3.4. $\mathbf{V}$ is the set of observed random variables, i.e. stochastic factors in the system; $\mathbf{E}$ is the set of all edges establishing relations between variables of the system, independent of the nature of those; $\mathbf{\Theta}$ is the set of C.P.T.s, with one element for each in $\mathbf{V}$.

The set $\mathbf{D}$ encompasses the set of all actions from external agents, with elements $\mathbf{d} \in \mathbf{D}$. The elements $\delta \in \mathbf{d}$ represent the options for a single decision d, and finally $\boldsymbol{\sigma}$ is the set of decisions taken for a *realisation* over the model, i.e. one action taken for each element on a subset of decisions.

Lastly, $\mathbf{U}$ is the set of nodes representing utility functions. Each node $u \in \mathbf{U}$ is associated with a deterministic function $f_u\left(\mathbf{Pa}(u)\right)$ defined over the parents of u according to $\mathbf{E}$, regardless of their nature.

Figure 5 contains a visual representation for an Influence Diagram. According to the convention, an elliptic node represents a random variable, a rectangle node stands for a decision and a diamond for utility; added colours simply as a visual aid.

Figure 5 – Influence Diagram for the *Entrepreneur* scenario. This diagram contains *Market* and *Survey* as random variables, *Test* and *Found* as decisions to be taken, and *Value* and *Cost* as utilities associated with the decisions.



Source: Adapted from Koller e Friedman (2009).

The modelling in Figure 5 corresponds to a problem where an entrepreneur has to decide about founding a company (decision node *Found*). To help assess the viability of this venture, this person can hire a survey company to perform a test (decision node *Test*), which incurs costs for the project (utility node *Cost*).

The market as a random variable determines the survey's results. If the entrepreneur didn't choose to pay for the test, the model considers a uniform distribution for the analysis. This survey's result is the factor determining the founding of the company.

The final value for this undertaking is a function of the market openness for the product at the time of launch and the decision to found the company or save the initial financing.

As a model derived from the B.N., Influence Diagrams also allow for inference tasks of maximum a posteriori. Besides that, this extended model also performs estimative of optimal strategies for agents based on the system's expected utility (Marinescu; Lee; Dechter, 2021).

The expected utility EU for an I.D. $\mathcal{X}$ over any realisation of actions $\boldsymbol{\sigma}$ is given by eq. 21.

$$EU\left(\mathcal{X}[\boldsymbol{\sigma}]\right) = \sum_{u \in \mathbf{U}} f_u\left(\mathbf{Pa}(u) | \boldsymbol{\sigma}\right) \tag{21}$$

Finally, an optimal realisation for this system is the combination $\boldsymbol{\sigma}^*$ which results in the maximum expected utility, as described in eq. 22.

$$MEU(\mathcal{X}) = EU\left(\mathcal{X}[\boldsymbol{\sigma}^*]\right) = \underset{\boldsymbol{\sigma}}{\operatorname{argmax}} \sum_{u \in \mathbf{U}} f_u\left(\mathbf{Pa}(u)|\boldsymbol{\sigma}\right) \tag{22}$$

3.6 ORIGINAL *PHYSARUM LEARNER*

Physarum polycephalum ("multi-headed slime mould") is a member of the Myxomycetes class whose ability to find the shortest path on a maze first reported by Nakagaki, Yamada e Toth (2000) sparked interest from the scientific community (Smith-Ferguson; Beekman, 2019).

Physarum Solver is a mathematical model that simulates the reported *P. polycephalum* behaviour and finds the shortest path between two nodes on an undirected graph. Given a start and finish node – referred to as source and sink – this meta-heuristic employs a system of equations based on Poiseuille flow and mass conservation to iterate over the graph and to find the shortest path after convergence (Tero; Kobayashi; Nakagaki, 2006).

The original *Physarum Learner* is a meta-heuristic to learn a Bayesian Network from data. It employs at its core the elements from the *Physarum Solver* algorithm to build an auxiliary structure based on an undirected graph (Schön *et al.*, 2014).

Given a data set D with n variables, the first step of the Physarum Learner is to build two graphs: one fully-connected undirected graph and one directed graph with edges at random directions, both with n vertices each.

The algorithm initialises the edges of the undirected graph, henceforth the *PhyMaze*, with a random uniform parameter value. This structure will serve as an external memory for the Bayesian Network creation.

After this initial set-up of graph structures, the algorithm creates a list of every pair of distinct nodes in random order. This step ensures that the subsequent procedures evaluate each edge at least once to be a part of the final Bayesian Network.

Next, the algorithm removes a pair of nodes and evaluates the *PhyMaze* with the *Physarum Solver* engine for the shortest path between them. After reaching convergence to a course, the algorithm creates a list of all the edges on this path and the direct edge between the source and sink for this iteration.

In possession of the shortest path between the two nodes for each edge on it, the algorithm attempts to insert it into the Bayesian Network while not creating a cycle.

If the insertion of one such connection in any proper direction results in a score improvement for the Bayesian Network over the data, the algorithm provides positive feedback to this edge's counterpart on the *PhyMaze*. This feedback increases the likelihood of good connections repeating throughout iterations; on the other hand, if the insertion of a link would decrease the score of the Bayesian Network, the algorithm provides negative feedback to this edge's counterpart on *PhyMaze*.

After evaluating all the edges on the shortest path between a pair of nodes, the algorithm picks another combination of nodes and repeats the process; the method iterates over the entire list of $\frac{n(n-1)}{2}$ combinations. Once this list is exhausted, the final Bayesian Network is considered the best according to the algorithm.

4 MATERIAL AND METHOD

This Chapter presents detailed descriptions regarding the benchmark (synthetic) data sets and climatological data sets employed throughout this thesis, alongside the description of the data collectors.

This Chapter also contains a geomorphological analysis of the soil in a particular plantation. Although these soil characteristics are relevant to subsequent results and discussions in this document, the original tests were conducted independently by a competent, independent third party, and these methods' descriptions are beyond the scope of this thesis.

Furthermore, this Chapter details the modifications to the original Physarum Learner algorithm, the mathematical models for the Bayesian regression, and the proposed Influence Diagram for irrigation accompanied by five different scenarios and their respective utility functions.

4.1 DATA COLLECTORS

INMET employs the commercial Vaisala HydroMat™ model MAWS301 as a standard climatological data collector for automatic stations, responsible for data acquisition and communication with the governmental systems (Moura *et al.*, 2019). The climatological factors monitored hourly by INMET are as follows:

- Temperature [°C], available at maximum, minimum, and instant values;
- Relative humidity [%], available at maximum, minimum, and instant values;
- Dew point temperature [°C], available at maximum, minimum, and instant values;
- Atmospheric pressure [hPa], available at maximum, minimum, and instant values;
- Wind speed and gust [$\mathrm{m} \cdot \mathrm{s}^{-1}$], measured at 10 m height;
- Wind direction [degrees];
- Incident solar radiation [$\mathrm{kJ} \cdot \mathrm{m}^{-2}$];
- Precipitation [mm].

INMET provides climatological data hourly within an open-access policy from their automatic collectors after acquisition, communication, and pre-processing. The final data set is available at https://tempo.inmet.gov.br/TabelaEstacoes/xxxx (accessed on 01 July 2023, site in Brazilian Portuguese), where xxxx is the required station's code. The station code for the automatic collector in Patrocínio/MG is $A523$.

4.2 TIME SERIES FOR CLIMATOLOGICAL DATA

The data employed throughout this paper originate predominantly from the Brazilian National Institute of Meteorology (INMET—Instituto Brasileiro de Meteorologia).

The data sets consists of time series of seven climatological variables, i.e., temperature, atmospheric pressure, relative humidity, dew point temperature, wind speed, incident solar radiation, and precipitation.

The particular monitored location is the city of Patrocínio in the state of Minas Gerais (18°59′48″ S, 46°59′09″ W), at 978.11 m above sea level, and a warm temperate with a hot summer climate profile (Cwa), according to the Köppen classification. Figure 6 presents the geographical location of the state of Minas Gerais (MG) within the Brazilian map and the Patrocínio municipality within the former, alongside four other municipalities whose climatological time series that the authors employed as validation sets. Patrocínio is Brazil's largest coffee producer and the fourth-largest producer of bovine milk, among other goods. The initial data set consists of 5473 daily samples for each climatological variable from Patrocínio/MG. The data span from 22 August 2008 to 31 January 2022, with a 97% rate of valid elements.

Figure 7a–g present graphical representations of the time series, with temperature, atmospheric pressure, relative humidity, and dew point temperatures sub-divided into maximum, mean, and minimum values for each day. Figure 7h presents the estimated evapotranspiration according to the Penman–Monteith equation over the available climatological data.

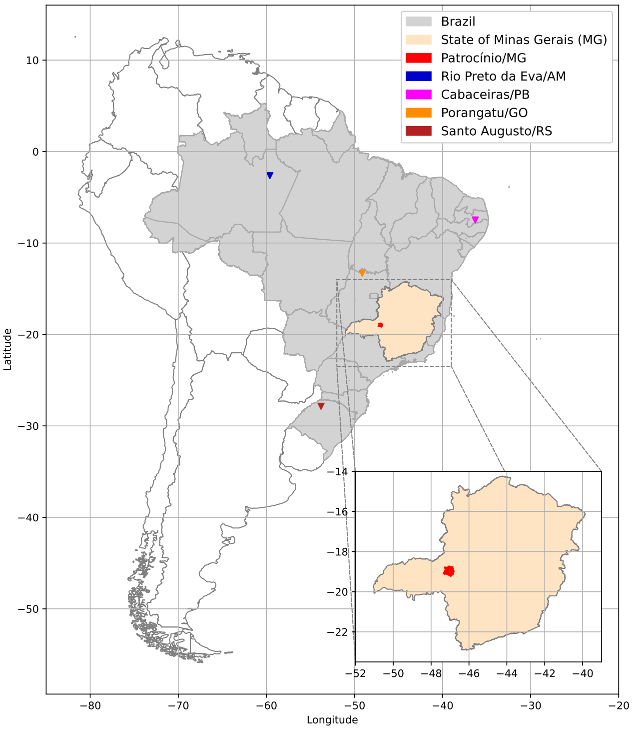
An additional data set with the same climatological variables from Patrocínio/MG acts as a test set to evaluate the proposed method. This extra set spans from 1 February to 31 December 2022, with a total of 334 samples. Furthermore, four supplementary data sets, each containing 334 samples from the same period previously described, from four different cities across Brazil, serve to evaluate the model when considering distinct regions and climate profiles. Table 1 presents the names of cities, their geographic coordinates, the climate profiles according to the Köppen classification, and their distances from Patrocínio/MG. The only criterion for choosing these cities is to provide good coverage across Brazil's territory.

Table 1 – Localisation and profile climate of four different cities within Brazil, alongside their distances to Patrocínio/MG.

City's Name	**Coordinates**	**Climate Profile**	**Distance to Patrocínio/MG (km)**
Rio Preto da Eva/AM	2°41′56″ S 59°42′00″ W	Tropical rain forest (Af)	2278
Cabaceiras/PB	7°29′20″ S 36°17′13″ W	Hot semi-arid (BSh)	1723
Porangatu/GO	13°26′27″ S 49°08′56″ W	Tropical wet (Aw)	659
Santo Augusto/RS	27°51′03″ S 53°46′37″ W	Humid subtropical (Cfa)	1202

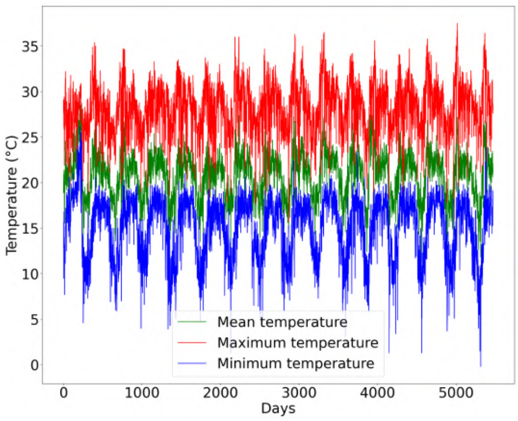
Source: Developed by the author.

Figure 6 – The geographical location of the Minas Gerais (MG) state is highlighted in yellow within Brazil's map. The red detail within the blue area highlights the Patrocínio/MG municipality's geographical location. The other four markers represent cities with different climate profiles from Patrocínio/MG's, whose climatological time series were employed by the authors to validate and deepen the discussion around the proposed model.

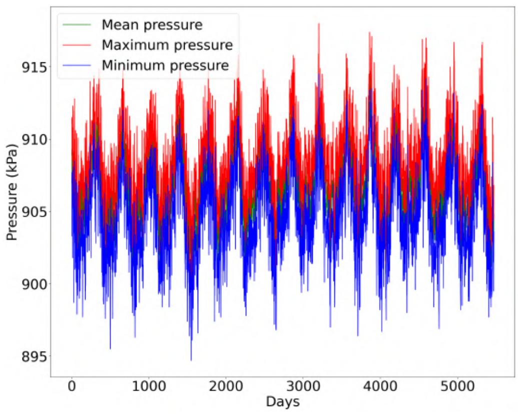


Source: Developed by the author.

Figure 7 – Visualisation of the complete time series for each of the fifteen climatological variables observed for the experiments in this paper, totalling 5473 samples; (**a**–**d**) follow the same colour pattern, where the red line represents the maximum value for that variable, green for the mean observed value, and blue for the minimum value; (**e**–**h**) represent the daily observations for each variable.

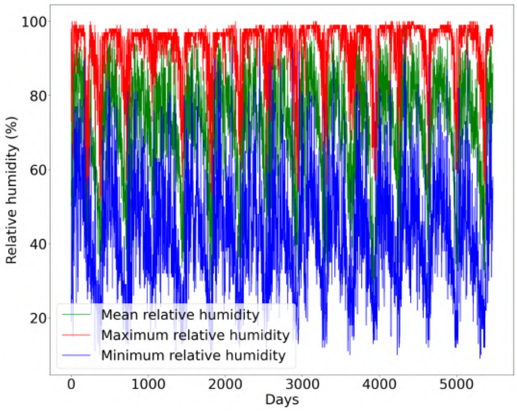


(a) Temperature.

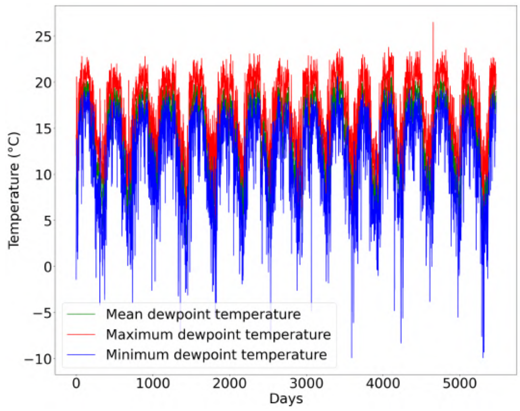


(b) Atmospheric pressure.

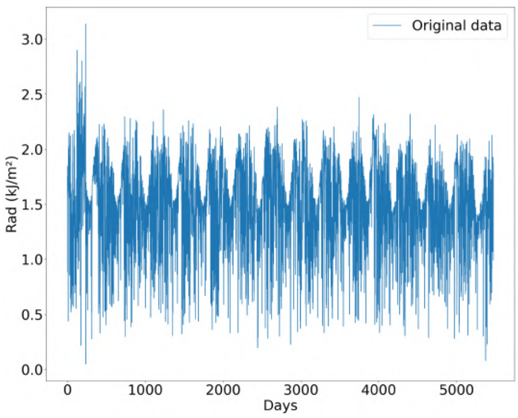
Figure 7 – *Cont.*



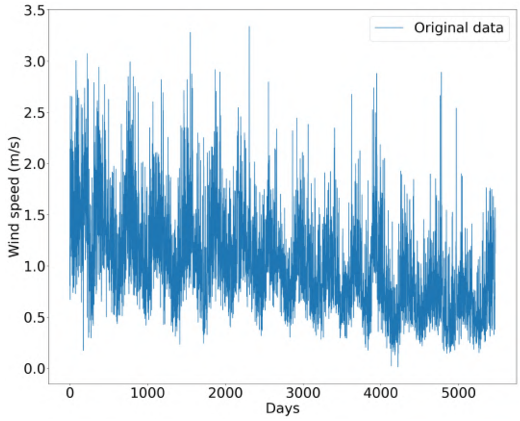
(c) Relative humidity.



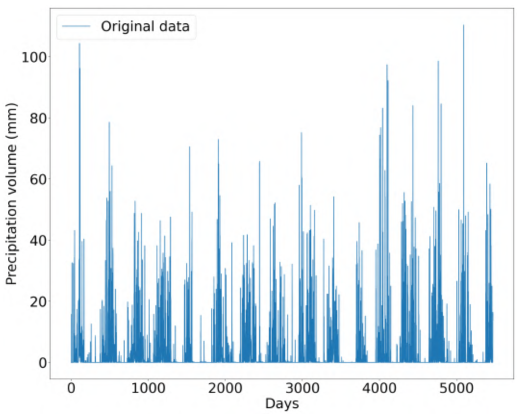
(d) Dew point temperature.



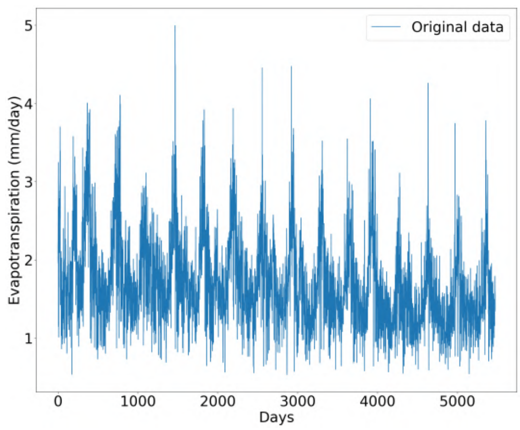
(e) Incident solar radiation.



(f) Wind speed.



(g) Precipitation.



(h) Estimated evapotranspiration.

Source: Developed by the author.

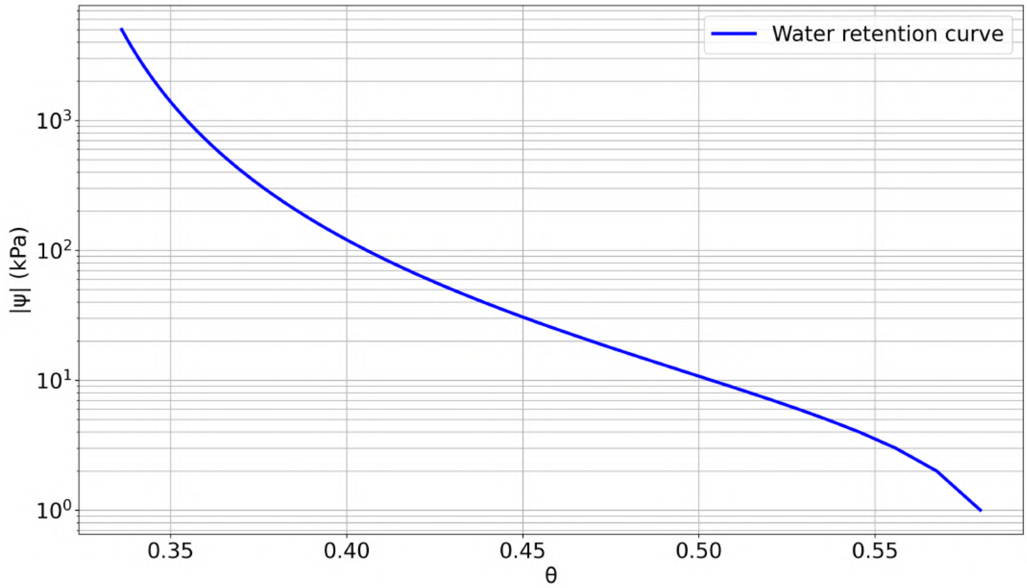
4.3 SOIL CHARACTERISATION

The soil characteristics around the Patrocínio/MG municipality correspond to those of the Ferralsol reference soil group, according to the International Union of Soil Sciences classification, with a high concentration of hematite (Santos *et al.*, 2018; Mantel; Dondeyne; Deckers, 2023).

The physical water parameters for this soil (field capacity (FC) and wilting point (WP)) alongside the water retention curve were obtained via the pressure plate method applied to three different soil samples, with one from each layer, i.e., from 0 cm to 20 cm, from 20 cm to 40 cm, and 40 cm to 60 cm.

The experimental results suggest uniform physical water parameters throughout the three layers. As such, Figure 8 presents the resulting water retention curve for the three strata.

Figure 8 – Semi-log representation for the water retention curve obtained from three soil samples from different depths near the Patrocínio/MG municipality. The uniform distribution of soil characteristics throughout the layers allows for the representation of a single water retention curve.



Source: Developed by the author.

Lastly, Table 2 presents the van Genuchten equation parameters based on the observed soil retention curve in Figure 8. The authors include the parameters for each of the three layers as a matter of completeness despite their uniformity.

Table 2 – Parameters for the van Genuchten equation obtained via the water retention curve and pressure plate method.

Layer	$\boldsymbol{\theta_s}$ **($cm^3 \cdot cm^{-3}$)**	$\boldsymbol{\theta_r}$ **($cm^3 \cdot cm^{-3}$)**	$\boldsymbol{\alpha}$ **(cm^{-3})**	$\boldsymbol{n}$	$\boldsymbol{m}$
0–20 cm	0.59	0.31	0.25	1.33	0.25
20–40 cm	0.59	0.31	0.25	1.33	0.25
40–60 cm	0.59	0.31	0.25	1.33	0.25

Source: Developed by the author.

4.4 BAYESIAN NETWORK BENCHMARK DATA SETS

Three benchmark data sets comprise the basis for evaluating this proposed *Improved Physarum Learner* implementation. These are the well-known *Asia*, *Sachs*, and *Alarm* data sets.

The work in (Lauritzen; Spiegelhalter, 1988) proposes the *Asia* network as a tool to illustrate how the medical expert system on the original paper works. The original structure is reproduced here as Figure 12a.

The structure consists of eight nodes and eight edges. Each node codifies a binary variable, which can only assume the values of *True* or *False*. This structure is one of the classical benchmarks due to its small number of nodes, and the fact that it is a weakly connected graph, i.e. there is a path between every pair of nodes after degenerating the edges' directions.

The *Sachs* network originates from (Sachs *et al.*, 2005), where the authors study possible causal relationships in a protein-signalling environment. The structure obtained from the original article is reproduced here as Figure 14a.

This structure contains eleven nodes, each representing a component in human immune system cells. The possible values for each node are the concentration of that component which adds to the computational effort required on the BDeu score function. This structure is relevant based on its relatively small number of vertices. Also, this graph is unconnected, which is a valuable characteristic due to the method using a fully-connected undirected graph as an auxiliary structure.

Finally, the *Alarm* network encodes a medical monitoring system built from the authors' expertise and medical textbook literature (Beinlich *et al.*, 1989).

The graph contains thirty-seven nodes, each representing one out of three categories of variable: a diagnostic, a measurement, or an intermediate variable. The values for the nodes are discrete, each with two or three levels. 12a represents the original structure.

The *Alarm* is a relevant network as a benchmark due to the expert knowledge that encodes it. Each of the "diagnostic" variables is independent of the others and is supposed to have no parents on the graphical structure, given the assumption that an identified condition characterised by a diagnostic is the cause of the variables associated with it. As the method deals primarily with a data set, this *a priori* knowledge is concealed from the structure and used afterwards to analyse the obtained network.

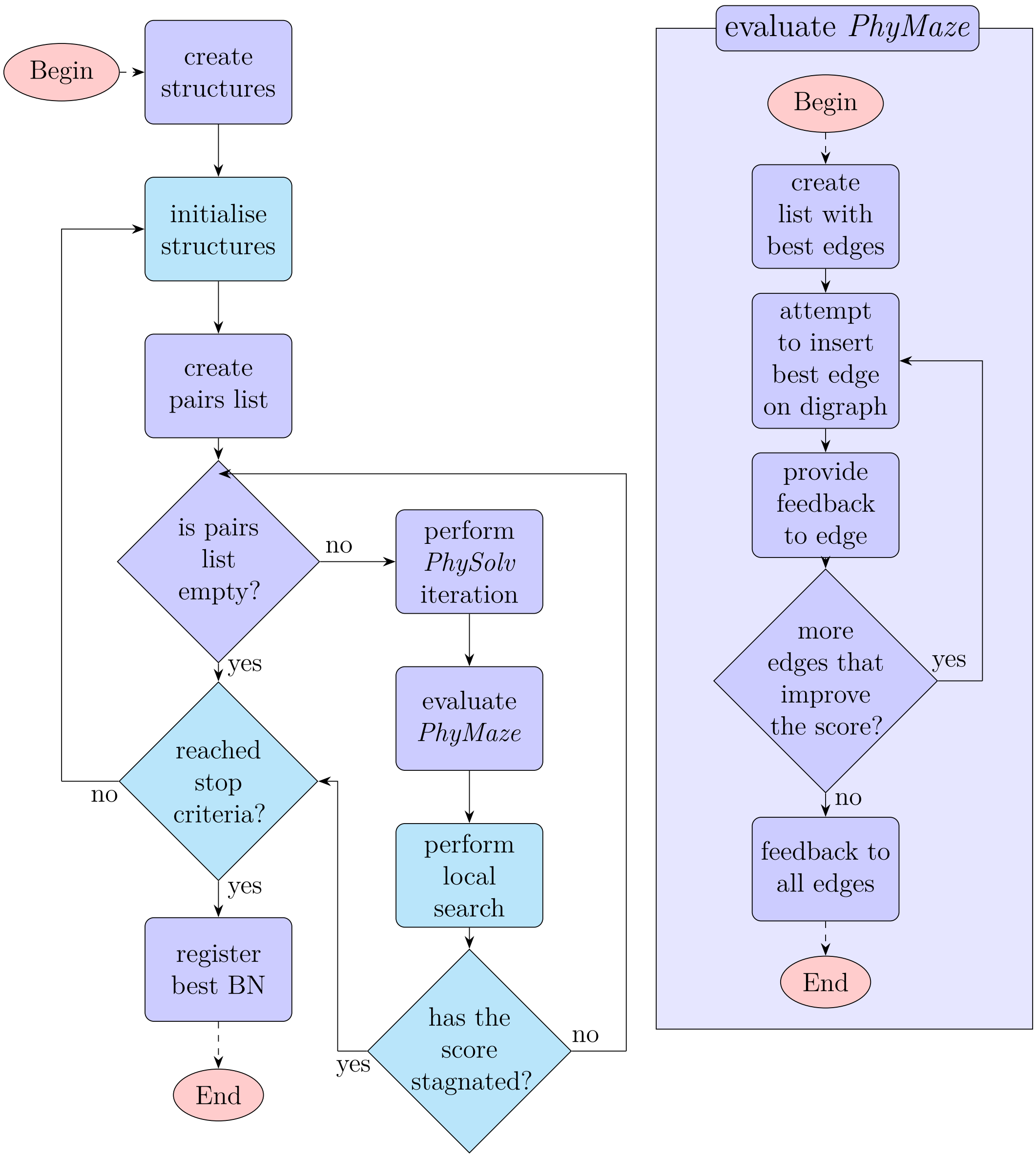
4.5 IMPROVED *PHYSARUM LEARNER*

The results presented in this Section are a complementary version of the work presented at the XXIV Brazilian Congress on Automatics (*Congresso Brasileiro de Automática* - C.B.A.). The complete paper is provided here as Annex A.

The first development for this project is a bio-inspired algorithm to tackle the problem of finding a Bayesian Network super-structure from data. An improved implementation of the *Physarum Learner* algorithm was proposed and implemented. The algorithm was applied to benchmark data sets to evaluate the method's accuracy.

Figure 9 contains a flowchart with the major components of the *Physarum Learner* algorithm (in cyan) and the elements added or improved on the *Improved Physarum Learner* implementation (in light blue).

Figure 9 – Flowchart depicting the main components of the original *Physarum Learner* as described in Schön *et al.* (2014), coloured in cyan. The elements in blue represent components added or modified from the original algorithm, resulting in a revised method called *Improved Physarum Learner*.



Source: Extracted from Ribeiro *et al.* (2022b).

The first improvement involves how the algorithm creates its initial structure (“initialise structures” node). The original method creates a directed graph with edges in random directions. This approach may result in cyclical graphs that violate the non-cyclical requirements for a Bayesian Network and cause issues with the BDEu scoring function.

The proposed version of the algorithm creates an initial Bayesian Network without edges. This approach ensures the initial structures have no cycles. Also, from a theoretical standpoint, isolated nodes on the graph mean the observed variables are statistically independent; and is a reasonable first step for every Bayesian Network score-and-search procedure.

Another improvement over the original algorithm is a local search after adding each edge. This step is convenient where each vertex added may impact the parenthood relations of several nodes. In this context, a "local search" is defined as operations which attempt to change the direction of edges, checking for a score improvement and ensuring a non-cyclical structure.

Two other aspects modified from the original method are related to stagnant structure and score. The presented method sets a threshold for the minimal improvement in score across iterations to check for structural stagnation on local optima. After achieving a scoring plateau, the algorithm breaks the current iteration and starts from a new possible network. This approach ensures that the algorithm won't waste time trying to build a structure based on one with a stagnant score.

4.6 BAYESIAN REGRESSION PARAMETRISATION

The statistical models employed in this paper address three factors to assess the quality of the Bayesian regression for the evapotranspiration metric: (1) Which mathematical transformation applied to the target variable yields regression results with the lowest error metric? (2) The influence of auto-regressive components on the quality of the regression. (3) An evaluation concerning the impact of the sample size on the regression learning process and its results.

The experiments evaluate five distinct operations over the evapotranspiration metric: linear—$f(x) = x$, Box-Cox transformation, square root, natural logarithm, and exponential. Equation (1) presents the definition of the one-parameter Box-Cox transformation (Box; Cox, 1964) with an estimation of the value of the parameter λ from the data set, which leads to better normality of the transformed series or is assumed to be known in each application. The number of auto-regressive elements is, at most, three. Lastly, experiments employed the last 1000 or 2500 samples or the entire set with 5473 entries. A complete permutation of these components results in a total of sixty experiments.

$$f(x,\lambda) = \begin{cases} \frac{(x\lambda - 1)}{\lambda}, & \lambda \neq 0 \\ log(x), & \lambda = 0 \end{cases} \tag{1}$$

Equations (2)–(8) establish the experiments' framework based on the posterior distribution in Equation (10). Let $N \in \{1000, 2500, 5473\}$ represent the number of samples for each experiment and X_i represent each of the fifteen climatological variables employed, according to Section 4.2. Equation (2) establishes the target series y composed of the evapotranspiration series after applying one of the five equations presented, while Equation (3) provides a prior probability distribution assumed for this series based on a Gaussian density function.

Equations (4a) and (4b) describe the parameterisation of the hyperparameter μ_i as a linear combination of the climatological elements. Equation (4a) applies when no auto-regressive window is present, while Equation (4b) denotes the experiments with $w \in \{1, 2, 3\}$ of such components.

Lastly, Equations (5)–(8) provide initial, approximately non-informative prior distributions for the parameter τ and regression parameters β_i. Equation (8) only applies to the experiments with at least one auto-regressive component.

The authors employed MultiBUGS 2.0 software (Goudie *et al.*, 2020) (available at https://www.multibugs.org/, accessed on 15 August 2023). Each experiment contained five chains, 1000 samples as the burn-in, and one in every ten values for the next 5000 obtained for the estimates.

$$y_i = f(ET_{0_i}) \qquad i = \{1, 2, ..., N\} \tag{2}$$

$$y_i \sim \mathcal{N}(\mu_i, \tau) \qquad i = \{1, 2, ..., N\} \tag{3}$$

$$\mu_i = \beta_0 + \sum_{j=1}^{15} \beta_j X_{ji} + \varepsilon \qquad i = \{1, 2, ..., N\} \tag{4a}$$

$$\mu_i = \beta_0 + \sum_{j=1}^{15} \beta_j X_{ji} + \sum_{k=1}^{w} \beta_k y_{(i-k)} + \varepsilon \qquad i = \{1, 2, ..., N\}, \quad w \in \{1, 2, 3\} \tag{4b}$$

$$\tau \sim \Gamma(1, 1) \tag{5}$$

$$\beta_0 \sim \mathcal{N}(4, 0.01) \tag{6}$$

$$\beta_j \sim \mathcal{N}(0, 1) \qquad j = \{1, 2, ..., 15\} \tag{7}$$

$$\beta_k \sim \mathcal{N}(0, 1) \qquad k = \{1, ..., w\} \tag{8}$$

4.7 INFLUENCE DIAGRAM FOR IRRIGATION

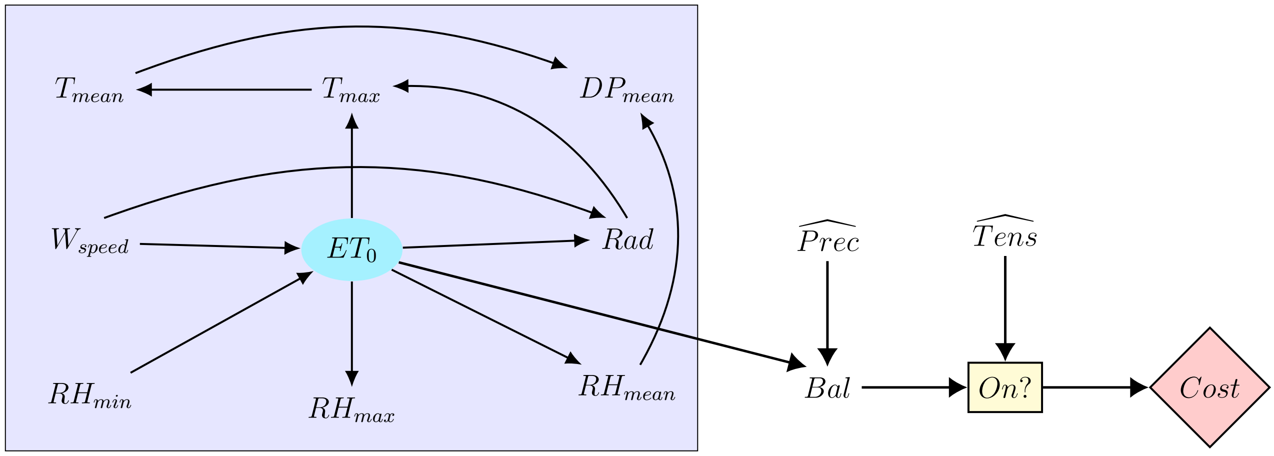
This project proposes an Influence Diagram for a decision support system for irrigation based on hydro-meteorological variables, soil parameters, and crop characteristics.

Figure 10 contains the graphical representation of the proposed model. The elements highlighted in blue correspond to the Bayesian Network for estimating ET_0 based on climatological data. An estimand for the water balance takes the estimated ET_0 and rainfall to evaluate whether the soil water content is above or below a given threshold.

The decision to activate the irrigation system considers a binary input for water balance and an estimate for the matric potential. The proposed utility is a function that penalises the decision for turning on when the soil water content is near the saturated level or the water balance estimate requires no extra irrigation and favours the other way around.

The climatological water balance random variable (Bal) represents a binary variable whose possible states are “surplus” or “deficit” – unless stated otherwise – based on the estimands

Figure 10 – Proposed Influence Diagram to provide decision support for irrigation. Elements in ellipses represent random variables, while rectangle is for decisions and diamond for utility function. Elements highlighted in blue form the Bayesian Network to estimate ET_0 from climatological data from Ribeiro *et al.* (2023).



Source: Adapted from (Ribeiro *et al.*, 2023).

for evapotranspiration and precipitation. The tensiometer estimates ($\widehat{Tens}$) are a normal distribution whose parameters come from an MCMC estimation process over *in-situ* readings from this instrument.

Finding a relevant utility function to describe the preferred outcome of each stakeholder is a non-trivial task due to the high level of subjectivity involved in its decisions. In this thesis context on agriculture, the decision to irrigate or not revolves around several additional factors, such as the produce, their growing stage, market conditions and possible water deprivation to induce faster production.

The five scenarios below represent generic situations where the irrigation decision is based solely on the climatological water balance and tensiometer readings, alongside their proposed utility function.

The first and most basic scenario models the decision based solely on the soil-water content above or below the field capacity threshold. Equation 9 presents a utility function f_1 as the mathematical model for this scenario, which relies on an estimate for the readings of a tensiometer $\hat{\psi}$ and the binary, stochastic decision regarding the irrigation system's activation On. The value ψ^* is the soil matric pressure at its field capacity, representing the optimal threshold for water-soil content.

$$f_1(\widehat{\psi}) = On \cdot \left(\frac{\theta(\psi^*)}{\theta(\widehat{\psi})} - 1 \right), \qquad On \in \{-1, 1\} \tag{9}$$

This first simple model achieves the maximum utility when the system irrigates under the soil lacking water content or when there is no irrigation under soil presenting more water-soil content than the minimum.

The second model employs solely tensiometer readings on a specific depth and a linear decay of water content based on the culture roots' depth. This scenario allows for a decision regarding cultures with a radicular system of any (reasonable) depth and the position of the available tensiometers. As such, let h represent the radicular depth of a given culture, $h1$ and $h2$ be the depths of two tensiometers with $h1 < h < h2$; eq. 10 presents the matric potential for any given depth based on the depth and matric potential estimates from two tensiometers. Eq. 11 describes a utility function f_2 based on the estimated value of matric potential at an arbitrary depth h.

$$\theta(\widehat{\psi_h}) = \theta(\widehat{\psi_{h2}}) + \left(\theta(\widehat{\psi_{h1}}) - \theta(\widehat{\psi_{h2}})\right) \cdot \frac{h - h_1}{h_2 - h_1} \tag{10}$$

$$f_2(\widehat{\psi_h}) = On \cdot \left(\frac{\theta(\psi^*)}{\theta(\widehat{\psi_h})} - 1\right), \qquad On \in \{-1, 1\} \tag{11}$$

The third scenario stratifies the climatological water balance Bal into four classes according to the severity of estimated water-soil content deficit or surplus - classes “high deficit”, “high surplus”, “deficit” and “surplus” - with a discrete probability distribution over these states. In this scenario, the decision to turn on the irrigation system becomes directly conditioned to Bal’s probability mass function. Figure 11 contains a graphical representation of this conditional relation and the proposed value for each state.

Figure 11 – Conditional probability tables and association proposed for the fourth scenario. The probability for Bal’s states reflect the most probable condition observed in Patrocínio/MG during the data collection period.

Bal	$p(Bal)$
“High deficit”	0.25
“Deficit”	0.50
“Surplus”	0.20
“High surplus”	0.05

→

Bal	$On? = -1$	$On? = 1$
“High deficit”	0.00	1.00
“Deficit”	0.20	0.80
“Surplus”	0.80	0.20
“High surplus”	1.00	0.00

Source: Developed by the author.

The fourth scenario stipulates a utility function which combines the water-soil content estimates with the stratified water balance, with the latter divided into four classes like in the third scenario. Eq. 12 presents the utility function f_4 for this scenario, considering a numerical value for each possible state in Bal where positive values indicate water surplus and negative represent water deficit.

$$f_4(\widehat{\psi}, \widehat{Bal}) = On?\left((-\widehat{Bal})^3 + 20\left(\theta(\psi^*) - 1.1\theta(\widehat{\psi})\right)\right) \tag{12}$$
$$On? \in \{-1, 1\}; \widehat{Bal} \in \{-3, -1, 1, 3\}$$

Lastly, the fifth scenario proposes a model considering a crop resilience factor to consider cases where the producer does want to induce water stress to force production. Eq. 13 encompasses this characteristic with an exponential factor regarding the number of days the crop is under water stress n_d and the maximum number of days before reaching its wilting point n_{max}.

$$f_5(\widehat{\psi}, \widehat{Bal}) = e^{(n_d - n_{max})} - \left(\frac{\theta(\psi^*)}{\theta(\widehat{\psi})}\right) \widehat{Bal} \tag{13}$$

5 RESULTS AND DISCUSSION

This Chapter contains the results and discussions obtained during this thesis' time frame.

Section 5.1 contains the obtained results and relevant discussions from the developed bio-inspired method applied to synthetic benchmark data sets and show its adequacy from both structural and score standpoints.

Section 5.2 contains a Bayesian Network for inference of the evapotranspiration metric, considering climatological variables from an open-source and organic data set provided by the Meteorological National Institute (Instituto Nacional de Meteorologia - INMET).

Section 5.3 provides the results and discussions regarding the sixty experiments in Section 4.6 and the ability of an approach based on Bayesian regression to model non-stationary and non-linear phenomena.

Section 5.4 expands the numerical analysis from Section 5.3 to include a comparison with real-world data. This evaluation is substantial to assess the modelling quality when dealing with non-synthetic data.

Section 5.5 presents the Bayesian Network that allows for evapotranspiration estimates considering scenarios with complete or incomplete evidence, alongside the results from Section 5.3.

Lastly, Section 5.6 provides the results and discussions regarding the results from the Influence Diagram model and scenarios presented in Section 4.7, with estimates from tensiometers and climatological water balance as inputs for the decisions.

Aside from the results in Section 5.6, the results in this Chapter culminated in publications in two international conferences and one journal. Annexes A, B and C contain the original version of these publications.

5.1 BIO-INSPIRED METHOD TO BUILD BAYESIAN NETWORK FROM DATA

The results presented in this Section are a complementary version of the work presented at the XXIV Brazilian Congress on Automatics (*Congresso Brasileiro de Automática - CBA*). The complete paper is provided here as Annex A.

The visual representations in this Section of the Bayesian Networks obtained by the Improved Physarum Learner employ a graph super-structure, i.e. a graph with the omitted direction of edges. This depiction is a reasonable approach given the fundamental role of a Bayesian Network as a display of *association* links instead of *causal* ones (Cox, 2018).

The colour scheme for the edges on the resulting Bayesian Network for this Section is as follows: green represents correct edges in the right direction; blue for correct ones but wrong direction; red for incorrect; and magenta for missing edges.

Figure 12b contains the best super-structure for the Bayesian Network found by the *Improved Physarum Learner* after the score stagnation. From the structure presented in Figure 12b, all the red edges except $Tuberculosis \rightarrow Bronchitis$ come from the concept of *d-separation* – given three random variables with an association chain $A \rightarrow B \rightarrow C$ it stands to reason that A may exert influence over C when fixed a value for B (Koller; Friedman, 2009).

Although the red edge *Tuberculosis* $\rightarrow$ *Bronchitis* can not be justifiable according to d-separation, the specialised medical literature presents an association of both respiratory diseases (Martin; Cochran; Katsura, 1968).

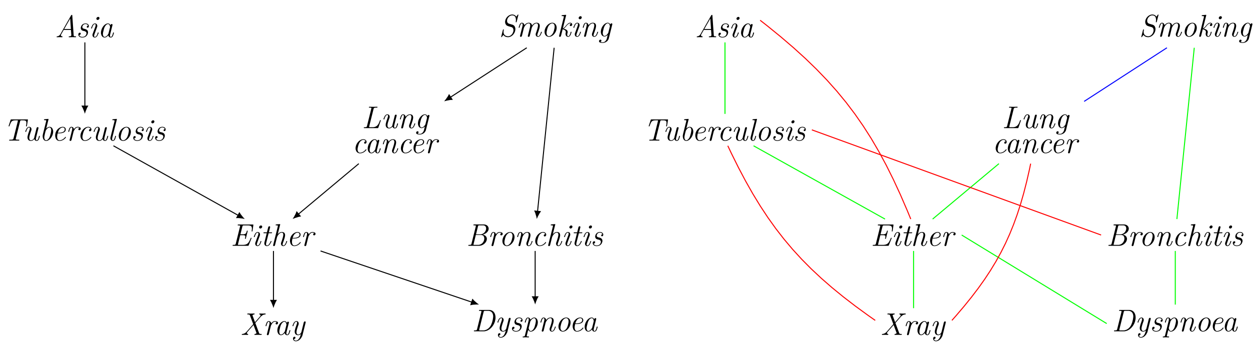
Figure 13 contains a semi-log representation of the score evolution across iterations from the *Improved Physarum Learner* on the *Asia* data set as the blue line. For comparison, the red baseline is the score for the original network structure over the same data set.

The results in Figure 13 illustrate the fast convergence of the *Improved Physarum Learner* to achieve a structure with a good score compared to the original network. This version of the algorithm reaches score stagnation after three iterations and may stop after a couple of iterations. The original method would execute all twenty-eight iterations.

A result presented in the detail of Figure 13 is that the obtained structure provides a score marginally better than the one obtained from the original network over the same data set. This result is a theoretically possible outcome, which occurs from the fact that the score function is sensible to both structure and data set evaluated.

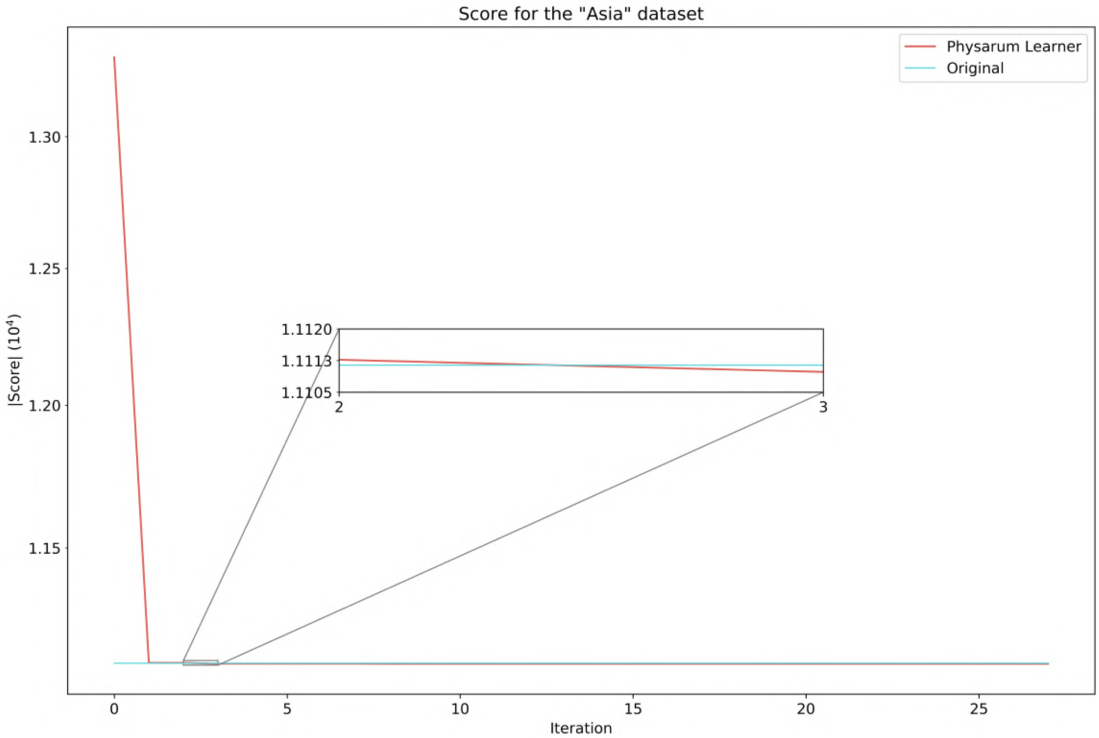
Figure 12 – Representations of the Bayesian Networks for the *Asia* data set, which contains eight binary variables. Figure 12a presents the original structure proposed in Lauritzen e Spiegelhalter (1988). Figure 12b presents a super-network structure found by the *Improved Physarum Learner* method, where green and blue edges represent correct edges, but in the correct or incorrect direction respectively, and red means one that does not exist compared to the original structure.

(a) Original structure. (b) Obtained structure.



Source: Extracted from Ribeiro *et al.* (2022b).

Figure 13 – Semi-log representation of score convergence of the *Improved Physarum Learner* for the *Asia* data set in blue; the red line is a baseline of the score obtained from the original structure for the same data set. The detail shows the iteration after which the score for the obtained structure was marginally better than the one from the original network.



Source: Extracted from Ribeiro *et al.* (2022b).

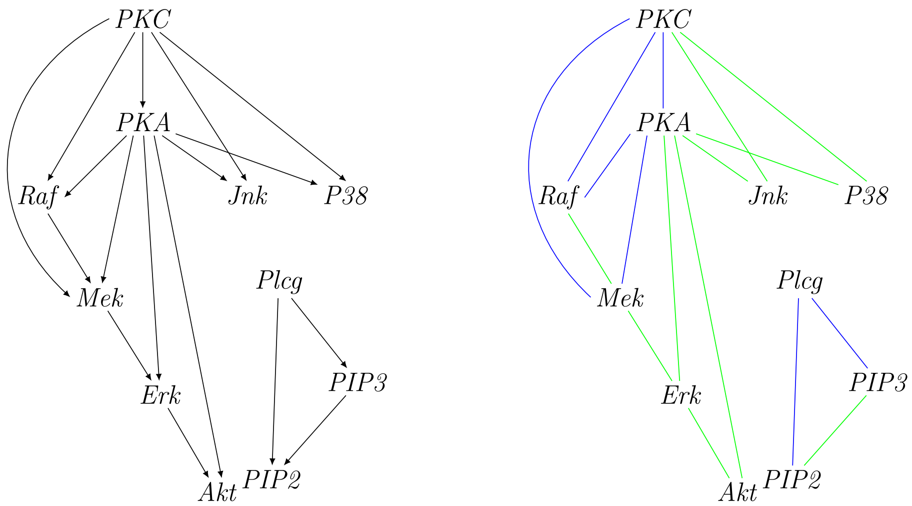
Figure 14b presents the graphical structure obtained by the *Improved Physarum Learner* for the *Sachs* data set. The algorithm created an unconnected graph with the nodes $Plcg$, $PIP2$, and $PIP3$ segregated from the other vertices. This result shows that the algorithm can create connected or unconnected graphs, even with the auxiliary structure as a fully connected graph.

Figure 15 presents the convergence results obtained for the *Sachs* network. This result illustrates the score stagnation issue and the mechanism to find a new structure after determining this condition. The method encountered a score stagnation after the ninth iteration and left the local optima for a network with a better score after iteration thirty-five.

The detail in the Figure 15 shows that the score for the obtained network is the same as for the traditional structure, despite the identified network containing both green and blue edges. This result points to one of the fundamental characteristics of the Bayesian Networks learnt exclusively from data: the requirement of external validation for the model from a specialised agent, which can evaluate the proper associations between variables in real-world scenarios.

Figure 14 – Representations of the Bayesian Networks for the *Sachs* data set, which contains eleven variables. Figure 14a presents the original structure proposed in Sachs *et al.* (2005). Figure 14b presents a super-network structure found by the *Improved Physarum Learner* method, where green and blue edges represent correct edges, but in the correct or incorrect direction respectively.

(a) Original structure. (b) Obtained structure.



Source: Extracted from Ribeiro *et al.* (2022b).

The scalability of a method to build Bayesian Networks from data is an essential factor, given that this is a problem of super-exponential class. The *Alarm* data set is a medium-sized benchmark used to evaluate the scalability of the *Improved Physarum Learner* method. The original network is present in Figure 17.

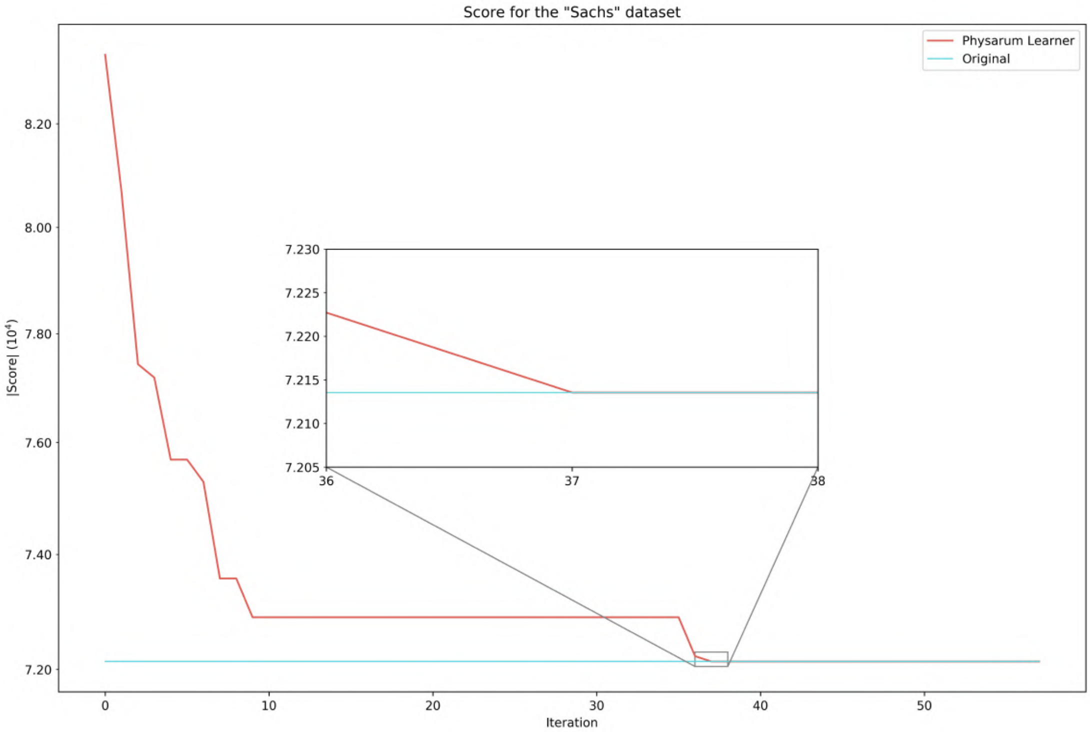
Figure 18 contains the graphical structure obtained by the *Improved Physarum Learner* on the *Alarm* data set. The properties of d-separation justify some of the eighteen red edges, such as $Intubation \rightarrow Vent.Tube$ or $LVFailure \rightarrow PCWP$.

Another subset of wrong edges reinforces the necessity of model validation from an expert in the field. For instance, as the authors that first proposed this network in Beinlich *et al.* (1989) described, the seven variables of the *Diagnostics* class, such as *Pulm.Embolus*, *Anaphylaxis*, and *LVFailure* should not interact with each other. The red edges $Pulm.Embolus \rightarrow Anaphylaxis$ and $Anaphylaxis \rightarrow LVFailure$ are evidence of issues that may occur with methods that rely solely on data.

To evaluate the structural accuracy of the *Improved Physarum Learner* for this medium-sized data set, Figure 19 contains all the edges from Figure 18 that are also present in the original structure and the only four edges the method did not identify from the data set, those represented in magenta.

Conditional independence tests performed to evaluate the association of these pairs of variables from the data set indicate independence. As the data does not suggest dependency between the components is reasonable that a data-based method could not identify those four missing edges.

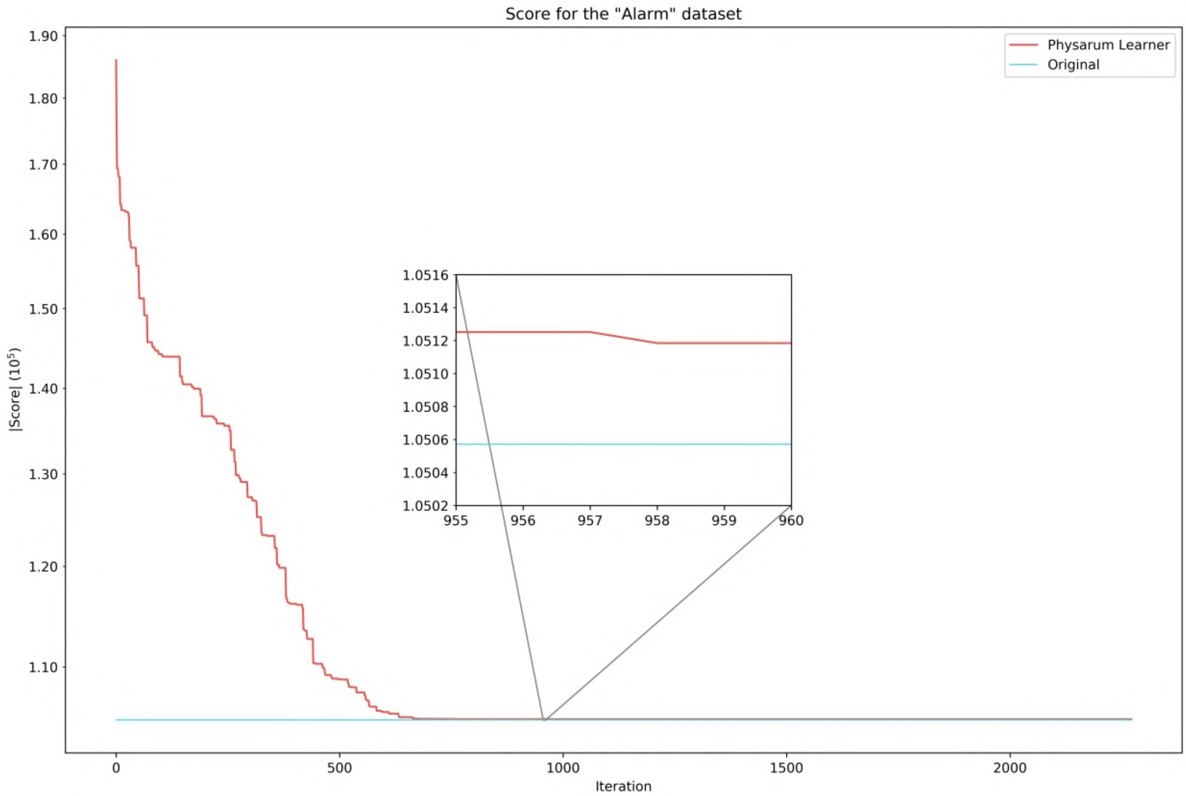
Figure 15 – Semi-log representation of score convergence of the *Improved Physarum Learner* for the *Sachs* data set in blue; the red line is a baseline of the score obtained from the original structure for the same data set. The detail shows the iteration after which the score for the obtained structure exactly the same as the one from the original network, despite the obtained structure presenting some edges in different directions.



Source: Extracted from Ribeiro *et al.* (2022b).

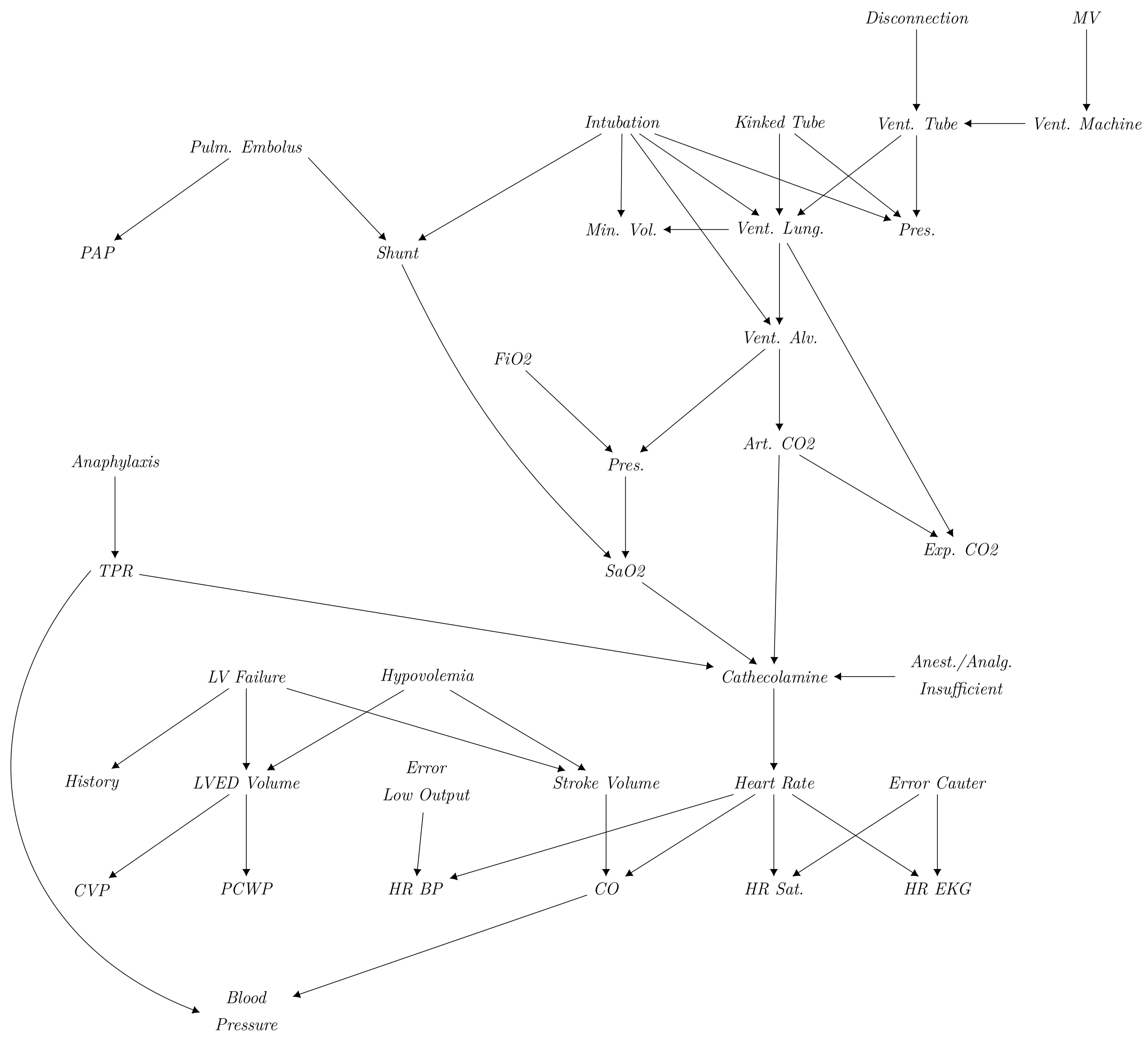
Lastly, Figure 16 contains the score convergence for the structure obtained by the *Improved Physarum Learner* on the *Alarm* data set. As this network contains the highest number of edges of the benchmark data sets, the *Improved Physarum Learner* takes more iterations to reach local optima. The fact that the data suggest independence across the pairs connected by magenta edges on Figure 19 may be the factor that stops the algorithm from achieving better scores.

Figure 16 – Semi-log representation of score convergence of the *Improved Physarum Learner* for the *Alarm* data set in blue; the red line is a baseline of the score obtained from the original structure for the same data set. The detail shows the iteration after which the method did not find a network that provided any score improvement within a threshold.



Source: Developed by the author.

Figure 17 – Original Bayesian Network for the *Alarm* medical data set. This network contains thirty-seven nodes each with two or three possible values and from one out of three categories: a diagnostic, a measurement, or an intermediate variable.

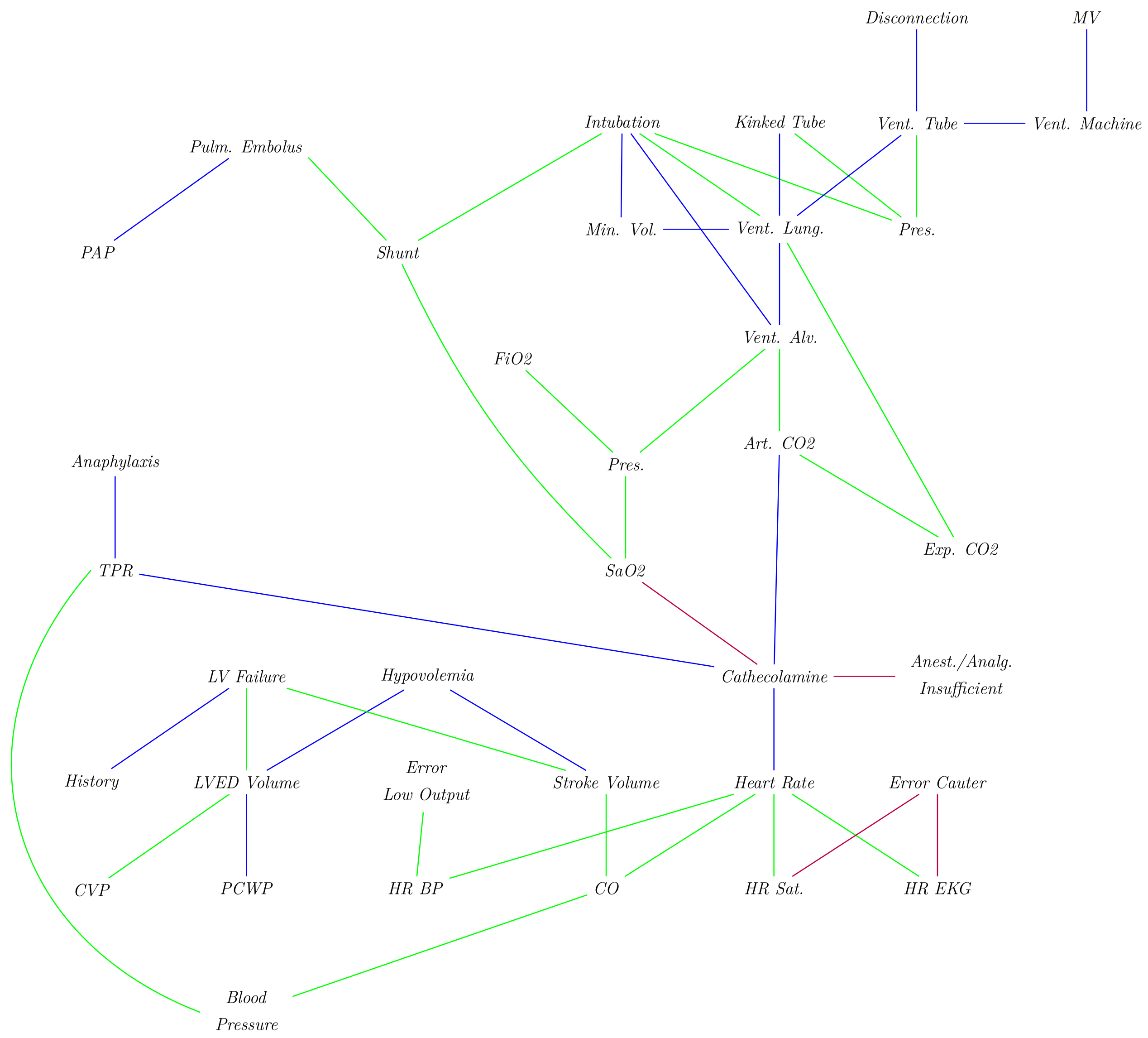


Source: Adapted from Beinlich *et al.* (1989).

Figure 18 – Bayesian Network obtained by the *Improved Physarum Learner* for the *Alarm* data set. Although the structure presents several wrong edges, the properties from d-separation and the analysis of a medical expert may remove several red edges.

Source: Developed by the author.

Figure 19 – Alternative representation for the Bayesian Network presented in Figure 18. This graphical structure has all green and blue edges from the obtained graph. In magenta, the only four edges present on the original network that the *Improved Physarum Learner* did not identify.



Source: Developed by the author.

5.2 STOCHASTIC INFERENCE OF EVAPOTRANSPIRATION FROM CLIMATOLOGICAL DATA

The results presented in this Section are a complementary version of the work presented at the 2022 International Joint Conference on Neural Networks (IJCNN). The complete paper is provided here as Annex B.

Given the difficulty of directly evaluating the evapotranspiration metric and the multiple analytical methods to estimate it from climatological data, this project establishes the Penman-Monteith (FAO56) equation as the gold standard.

The climatological variables provided by the INMET and required to estimate evapotranspiration according to the FAO56 method are: maximum, minimum, and mean temperature, wind speed (at a 2m height), net solar radiation, and maximum and minimum relative humidity.

A Bayesian Network to analyse the stochastic associations among the climatological variables and the required target metric contains eight nodes – one for each observed element. To provide a set of relationships extracted from data, this model employs a set created with real-world climatological data from several cities in the MATOPIBA region and the evapotranspiration metric from the FAO56 method.

Figure 20 contains the graph for the Bayesian Network obtained from the Improved Physarum Learner optimisation method using the data set from Luís Eduardo Magalhães/BA (5474 samples). An out-of-the-box optimisation algorithm achieved the same structure. Although some of the edges in Figure 20 and their directions may seem inappropriate considering the evapotranspiration dynamics, a Bayesian Network is suited to extract association relations from the data, which is more nuanced than causal relations.

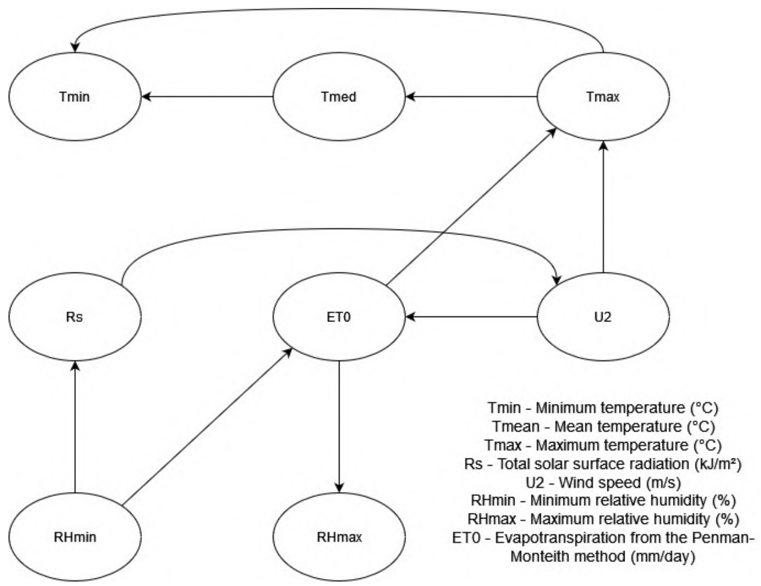
This stochastic modelling provides inferences over a target variable under scenarios where not all the other variables' states are known. That is a powerful characteristic for the evapotranspiration subject, given that the gold standard can estimate evapotranspiration only if all the required climatological variables are known.

The model in Figure 20 stipulates that wind speed and minimum relative humidity variables probabilistically condition the values for evapotranspiration. Also, the model specifies that the target variable conditions the value for maximum temperature and maximum relative humidity.

The algorithm then performs a 5-fold cross-validation on the data set to evaluate the accuracy and precision of the stochastic model in inferring the evapotranspiration from climatological data. This approach divides the data set into five sets with the same amount of samples, and then each four-fifths of the set is used to train the model, while the remaining fifth is employed to test the accuracy.

Table 3 contains the accuracy of evapotranspiration estimates from the model presented in Figure 20 compared to the expected value from the gold standard, trained and tested with data from five cities in the MATOPIBA region and one in Minas Gerais. The first row holds the accuracy of inferences under the scenario of complete information, i. e. all seven climatological variables presented as evidence. These results show that the model achieves an accuracy of over 80% in all cities, even with data from a location with a different climate profile than the one in Luís Eduardo Magalhães.

Figure 20 – Bayesian network obtained for the reference evapotranspiration (ET_0), considering the variables for the Penman-Monteith method applied on the data set from Luís Eduardo Magalhães. The climatological data represented by each node is described on the bottom right of the figure.



Source: Extracted from Ribeiro *et al.* (2022a).

The second and third rows from Table 3 contain the accuracy values for evapotranspiration estimates after removing each of the target's parents from the evidence to query. The drop from over 80% to closer to 70% highlights the importance of wind speed and minimum relative humidity for an accurate evapotranspiration estimate according to the proposed model. This result is appropriate given that the model lacks one of the required values to infer the most probable state for the evapotranspiration metric.

Table 3 – Inference accuracy (%) for the ET_0 variable from complete and missing data across several cities from the MATOPIBA region and Patrocínio/MG. These results suggest the adequacy of the proposed graphical model in providing estimates for ET_0, even under scenarios of incomplete climatological evidence.

Removed variable(s)	**Luís Eduardo Magalhães**	**Correntina**	**Imperatriz**	**Formoso do Araguaia**	**Alvorada do Gurgueia**	**Patrocinio**
–	82.5	86.3	85.5	83.1	80.2	84.6
$U2$	65.7	84.8	79.5	78.6	70.6	77.5
$RHmin$	79.2	80.4	80.8	79.1	73.8	84.1
$Tmax$	83.0	86.5	85.5	83.1	79.8	85.1
$RHmax$	80.7	84.2	87.2	78.0	76.3	83.1
All parents	64.3	79.3	74.2	75.6	66.7	79.4
All children	80.9	84.3	87.1	77.8	77.5	83.3
Both parents and children	53.0	67.8	73.7	59.9	54.8	71.3

Source: Adapted from Ribeiro *et al.* (2022a).

The fourth and fifth rows contain the accuracy of estimates for the target variable after removing its children's values from the evidence to query. The negligible loss of accuracy for the inference is reasonable considering the graphical model in Figure 1. Although the children's values may indicate the most probable state of the evapotranspiration metric, the parents of the target variable are the ones who rule over its value.

The subsequent two rows present the results of removing the whole set of parents or children respectively from the evidence to query. The removal of both parents reinforces the discussion on the role of these variables in the estimative's accuracy. Likewise, removing only the children from the evidence marginally reduces the model's accuracy, in line with the observed from the fourth and fifth rows.

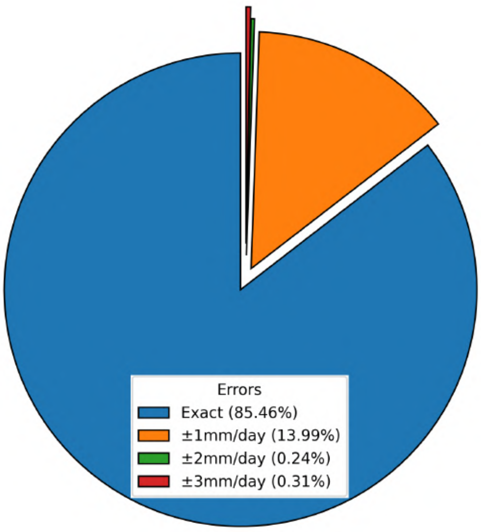
Lastly, the final row contains the accuracy under the scenario of evidence lacking information on any variables of those four adjacent to evapotranspiration, which results in the most degraded accuracy performance. This outcome provides an insight into the importance of the children's value in the inference task, given the discrepancies observed between the sixth and eighth rows.

To evaluate the scattering of the inferred evapotranspiration values under the presented scenarios, Figure 21 contains a set of pie charts for the difference between the estimated value against the expected one obtained with the gold standard. These pie charts are the inference results over the data set from Patrocínio/MG.

The results in Figures 21a to 21h suggest that the evapotranspiration value inferred from the stochastic modelling, when incorrect, is seldom outside the ±1mm/day range, even under the scenario of the least number of available climatological variables.

Figure 21 – Accuracy results for estimates of evapotranspiration according to the proposed Bayesian Network modelling. These results are from the city of Patrocínio/MG. Each pie chart from Figure 21a to Figure 21h represent the inference error after the removal of a subset of climatological variables from the available data.

(a) No variables removed.



(b) Removing both parents' and children's data.

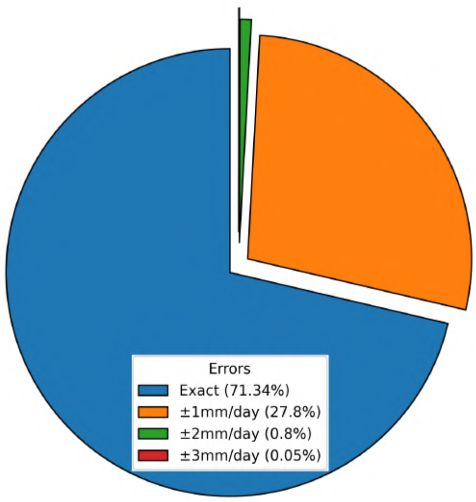
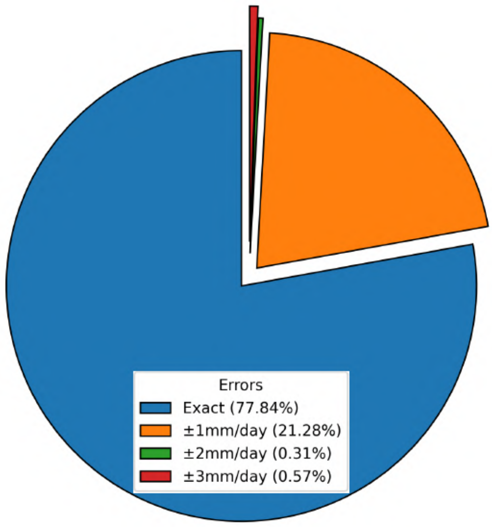
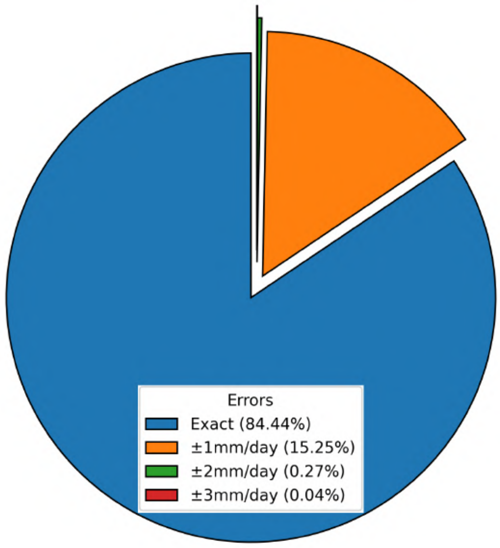


Figure 21 – *Cont.*

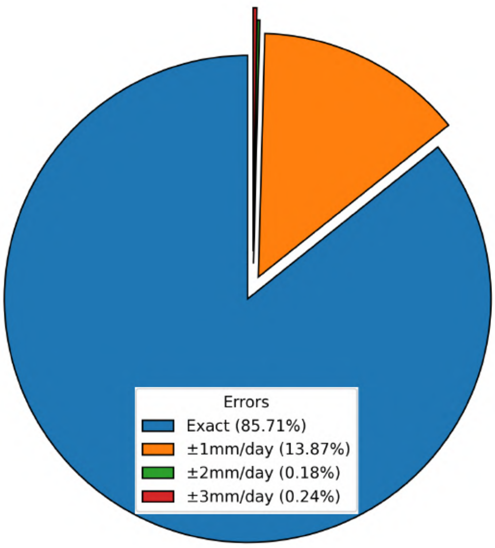
(c) Removing wind speed.



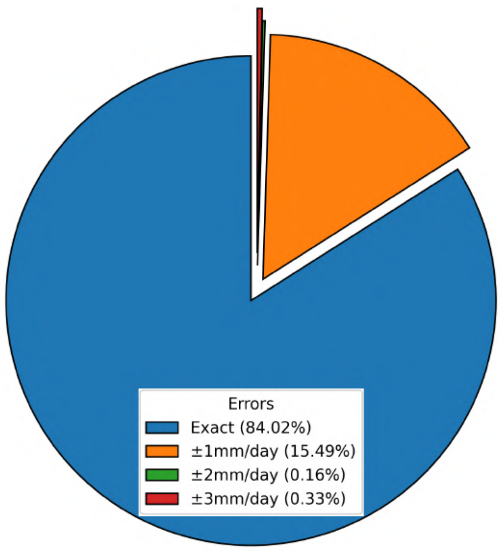
(d) Removing minimum relative humidity.



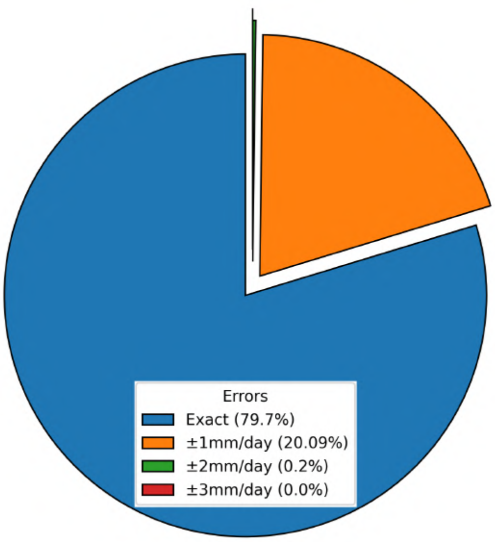
(e) Removing maximum temperature.



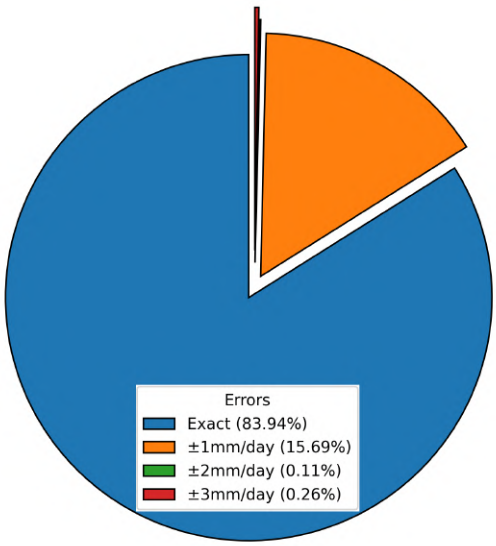
(f) Removing maximum relative humidity.



(g) Removing both parents' data.



(h) Removing both children's data.



5.3 CLIMATOLOGICAL VARIABLES SELECTION

This section presents the results and relevant discussions about the experiments in Section 4.6. Appendix A presents the complete tables for the numeric results of all the experiments, and presents the graphical comparisons of the evapotranspiration metric, considering the inference task of each climatological variable.

The root mean square error (RMSE) metric for each of the sixty experiments against the expected time series for the Penman–Monteith equation over the original data set ranges from $0.086 \text{ mm} \cdot \text{day}^{-1}$ to $3.502 \text{ mm} \cdot \text{day}^{-1}$. These results range from the experiment with logarithmic transformation, one auto-regressive element, and 1000 samples, to the test with the exponential function, no auto-regressive terms, and 5473 samples, respectively. Table 12 presents the RMSE results for the sixty experiments proposed in Section 4.6, alongside their mean absolute error (MAE) metrics; the closer the value of each error metric to zero, the better adherence between the two series.

The total time required to reach convergence for each experiment on MultiBUGS depends mainly on the amount of data provided to the software. Experiments with 1000 samples took 12.2 min, 2500 samples required 36.3 min, and the complete dataset needed 67.3 min, on average. These times were achieved on a desktop with an Intel® Core™ i5-6600 CPU at 3.30 GHz, 8 GB of RAM, and operating a Windows® 10 Pro 64-bit.

The first finding is that the RMSE metric exhibits minor fluctuations within the same transformation but higher fluctuations across functions. For instance, the experiments within the logarithm function present RMSE from $0.086 \text{ mm} \cdot \text{day}^{-1}$ to $0.100 \text{ mm} \cdot \text{day}^{-1}$, while experiments within the Box-Cox transform—with each λ parameter automatically obtained by the transformation algorithm for each series—achieve RMSEs from $0.097 \text{ mm} \cdot \text{day}^{-1}$ to $0.107 \text{ mm} \cdot \text{day}^{-1}$. These results suggest that the mathematical function choice is a relevant first step to minimize the estimation error based on this approach.

Within the same mathematical transformation, when comparing the error metric according to the number of samples of each experiment, the error metric consistently presents the lowest value for experiments with 2500 samples. Although this improvement is marginal compared to the errors from those with 1000 samples, this result suggests the existence of an optimal data volume for the inference task on these variables, and a large volume of data may impair the inference task by introducing more noise into the model.

Finally, when observing the error performance based on the amount of past information provided to the model, there is a marginal improvement across all sixty experiments for each past value added. This result is consistent with the observed meteorological dynamics, as the amount of water evaporated and transpired from the previous day plays a role in the maintenance of relative humidity, dew point temperature, and surface temperature for the current day, which are factors that directly impact the evapotranspiration phenomenon.

The exponential transformation is the one that consistently presents inference results with the highest error metric, according to Table 12. The results and discussions from this point onward do not consider the experiments with the exponential function, given their errors within one or two orders of magnitude greater than those obtained by the other four transformations.

The output of the MultiBUGS software includes the interval at 95% credibility for each beta regression parameter and the most probable value within this interval. As such, a climatological element is relevant if the credible interval for its beta coefficient does not encompass zero, as any value for the climatological variable would represent no impact in the regression.

Table 13 contains a complete overview of the occurrence of each climatological variable deemed relevant by the regression model, grouped by transformation. Across all forty-eight experiments—excluding the twelve from the exponential function—precipitation and all three measurements for atmospheric pressure do not contribute to the regression model. The variables for maximum dew point temperature and minimum temperature and dew point temperature contributed to a few experiments.

Eight variables consistently supported the regression analysis in more than half of the twelve experiments for each function. Wind speed and net solar radiation are relevant across all forty-eight experiments; mean temperature and dew point temperature, maximum temperature and relative humidity, and minimum relative humidity are significant for several experiments. Mean relative humidity is the last of the relevant climatological variables, although this component did not contribute to any regression with the square root function.

The high relevance observed from the wind speed and net solar radiation throughout all experiments is consistent with the observed meteorological phenomena. Both components directly contribute to the regional climate, represented by temperature, relative humidity, and dew point temperature. The irrelevance of the precipitation is also within expectation, given the assumption that elements within a much wider region contribute to this component with their dynamics not considered here. The atmospheric pressure components directly impact the dew point temperature, and the absence of this variable from the set of relevant variables is likely due to a discrepancy in the magnitude of the pressure measurements compared to the other elements

While Table 12 shows that the natural logarithm is the function that provides a regression with the lowest errors, Table 13 presents this function as the most assertive to which variables are most relevant to the stochastic regression for evapotranspiration. Further, the error improvements within this transformation are negligible after providing more data and auto-regressive components, given the trade-off of obtaining these. Based on these considerations, the authors establish that the stochastic model with a natural logarithm transformation and no auto-regressive elements, trained with 1000 samples, shows the best adherence to the evapotranspiration metric.

Table 4 presents the numerical values achieved for the beta coefficients from the experiment utilising natural logarithm transformation on the evapotranspiration data, with 1000 samples for the parameter learning and no auto-regressive component. The coefficients and their respective climatological variables—whose 95% credible intervals do not contain zero and, as such, characterise a relevant variable in this context—are highlighted in blue. The independent coefficient β_0 for this experiment is $\beta_0 \approx 0.6373$.

Based on the maximum a posteriori (MAP) values of each beta parameter and the relationships detailed in Equations (2) and Equation (4a), Equation (1) presents the approximation for the hierarchical Bayesian regression for the evapotranspiration metric; this is based on the experiment identified as the best among the sixty conducted. Although this representation for Equation (1) is similar to linear regression, each beta coefficient corresponds to a density distribution function, and its best point estimator value lies within the non-symmetrical 95% credible interval.

$$
\begin{aligned}
ln(ET_0) \approx 0.6373 - 0.024 \cdot T_{mean} + 0.022 \cdot T_{max} \\
- 0.006 \cdot RH_{mean} - 0.005 \cdot RH_{max} - 0.010 \cdot RH_{min} \\
+ 0.028 \cdot DP_{mean} + 0.498 \cdot W_{speed} + 0.113 \cdot Rad
\end{aligned} \tag{1}
$$

Table 4 – Credible interval at 95% for the beta regression coefficients obtained from the experiment with natural logarithm transformation, using a data partition of 1000 samples, and excluding the auto-regressive window feature. The relevant beta coefficients (i.e., those that do not include zero on the interval or whose value is close to zero) are highlighted in blue.

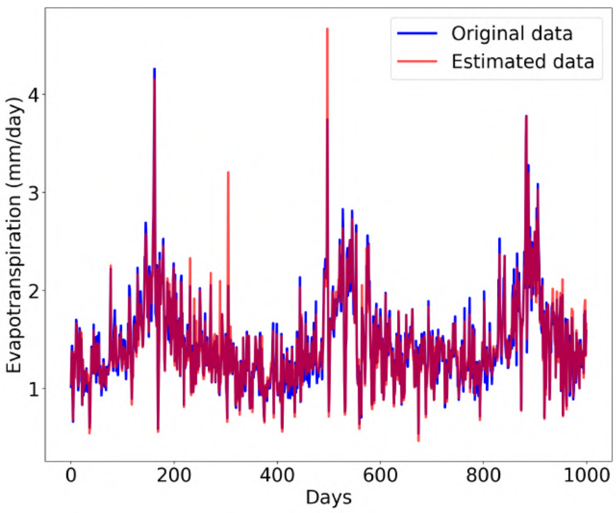
Variable	**Coefficient**	**2.5%**	**97.5%**	**MAP Value**
Tmean	β_1	−0.035	−0.010	−0.024
Tmax	β_2	0.014	0.028	0.022
Tmin	β_3	−0.008	0.003	0
RHmean	β_4	−0.009	−0.004	−0.006
RHmax	β_5	−0.005	−0.004	−0.005
RHmin	β_6	−0.011	−0.008	−0.010
DPmean	β_7	0.014	0.039	0.028
DPmax	β_8	−0.005	0.006	0
DPmin	β_9	−0.008	0.004	0
Pmean	β_{10}	−0.001	0.002	0
Pmax	β_{11}	−0.003	0.001	0
Pmin	β_{12}	−0.003	0.001	0
Wspeed	β_{13}	0.484	0.513	0.498
Rad	β_{14}	0.093	0.133	0.113
Prec.	β_{15}	−0.001	0.000	0

Source: Extracted from Ribeiro *et al.* (2023).

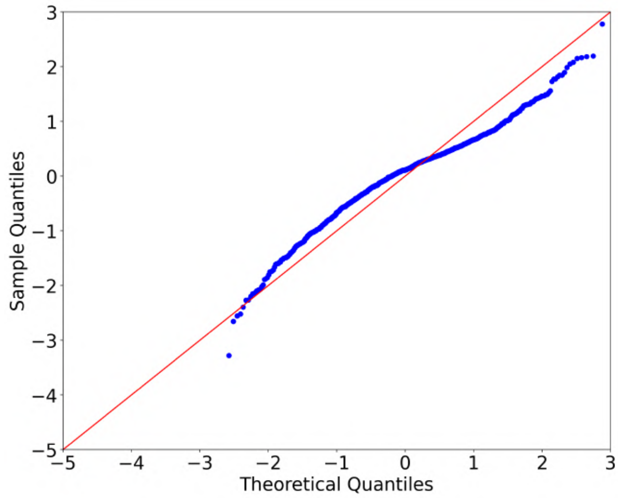
Figure 22a contains a graphical comparison of the evapotranspiration time series based on the Penman–Monteith equation (blue continuous line) against the estimated time series from Equation (1) (red line) throughout 1000 samples. This result suggests that the proposed method and subsequent equation properly conciliated the series' non-stationarity.

It is relevant to evaluate if the error component follows a Gaussian distribution for the estimated time series against the expected one, as the proposed Bayesian regression in Equation (1) is similar to the model in Equation (4a). To this end, Figure 22b presents a quantile–quantile plot for the distribution of the residual errors. As the blue dots lie close to the red 45-degree line, the distribution for the error is close to a Gaussian distribution, as required by the theoretical model.

Figure 22 – Graphical representation of the results obtained from the experiment with natural logarithm transformation, using a data partition of 1000 samples, and excluding the auto-regressive window feature; (**a**) presents the original time series as the blue line and the estimated time series as the red line; (**b**) presents the quantile–quantile plot of the residuals for the estimated time series.



(a) Comparison.

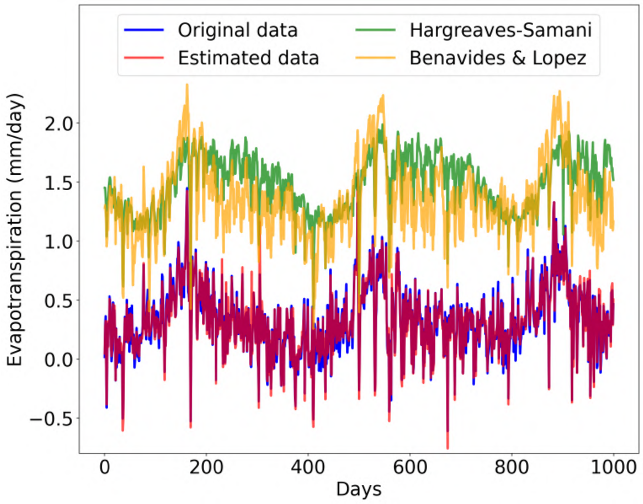


(b) Quantile–quantile plot.

Source: Extracted from Ribeiro *et al.* (2023).

One relevant comparison is to ascertain the proposed method's adherence to the expected evapotranspiration according to the gold-standard FAO56 method against the evapotranspiration estimated by other classical and deterministic approaches. As such, Figure 23 presents a graphical comparison between the natural logarithm of the evapotranspiration from the FAO56 method (blue line), the Benavides and Lopez method (orange line), and the Hargreaves–Samani method (green line); all three are based on the same data set, alongside the estimate from the proposed method (red line). This result confirms the ability of the proposed approach to closely model the complex behaviour of the FAO56 estimate, with an adherence better than the ones presented by other classical equations.

Figure 23 – Graphical comparison of the results obtained from the experiment with natural logarithm transformation, using a data partition of 1000 samples, and excluding the auto-regressive window feature (as the red line). The blue line represents the natural logarithm of the reference evapotranspiration according to the FAO-56 equation; the orange and green lines represent the natural logarithm for the reference evapotranspiration according to the Benavides and Lopez and the Hargreaves–Samani equations, respectively. The latter two methods represent some of the analytical approximations commonly employed under the unavailability of the climatological components required by the FAO-56 equation.



Source: Extracted from Ribeiro *et al.* (2023).

5.4 STOCHASTIC CLIMATOLOGICAL WATER BALANCE COMPARISON WITH *IN SITU* MEASUREMENTS

A proxy measurement for the water balance dynamics is necessary, mainly due to the characteristics of the evapotranspiration metric. Given the measurement hardship, this paper employs an estimate for the available soil water content on a coffee plantation in the proximity of Patrocínio/MG based on tensiometer readings and the van Genuchten equation.

Although the authors cannot divulge the precise location of this plantation due to confidentiality concerns, Equation (2) presents the estimated soil water content θ as a function of the matric potential ψ (kPa), according to the original Equation (7). For this particular soil, $\theta_s \approx 59\%$ and $\theta_r \approx 31\%$, $\alpha \approx 0.25$ cm^{-1}, $n \approx 1.33$ and $m \approx 0.25$, based on the experiments described in Section 4.3. Tensiometer readings were provided by a PalmaFlex digital tensiometer, model SL-PF-TNSE.

$$\theta = vG\,(\psi) \approx 0.31 + \frac{0.28}{\left(1 + (0.25|\psi|)^{1.33}\right)^{0.25}} \tag{2}$$

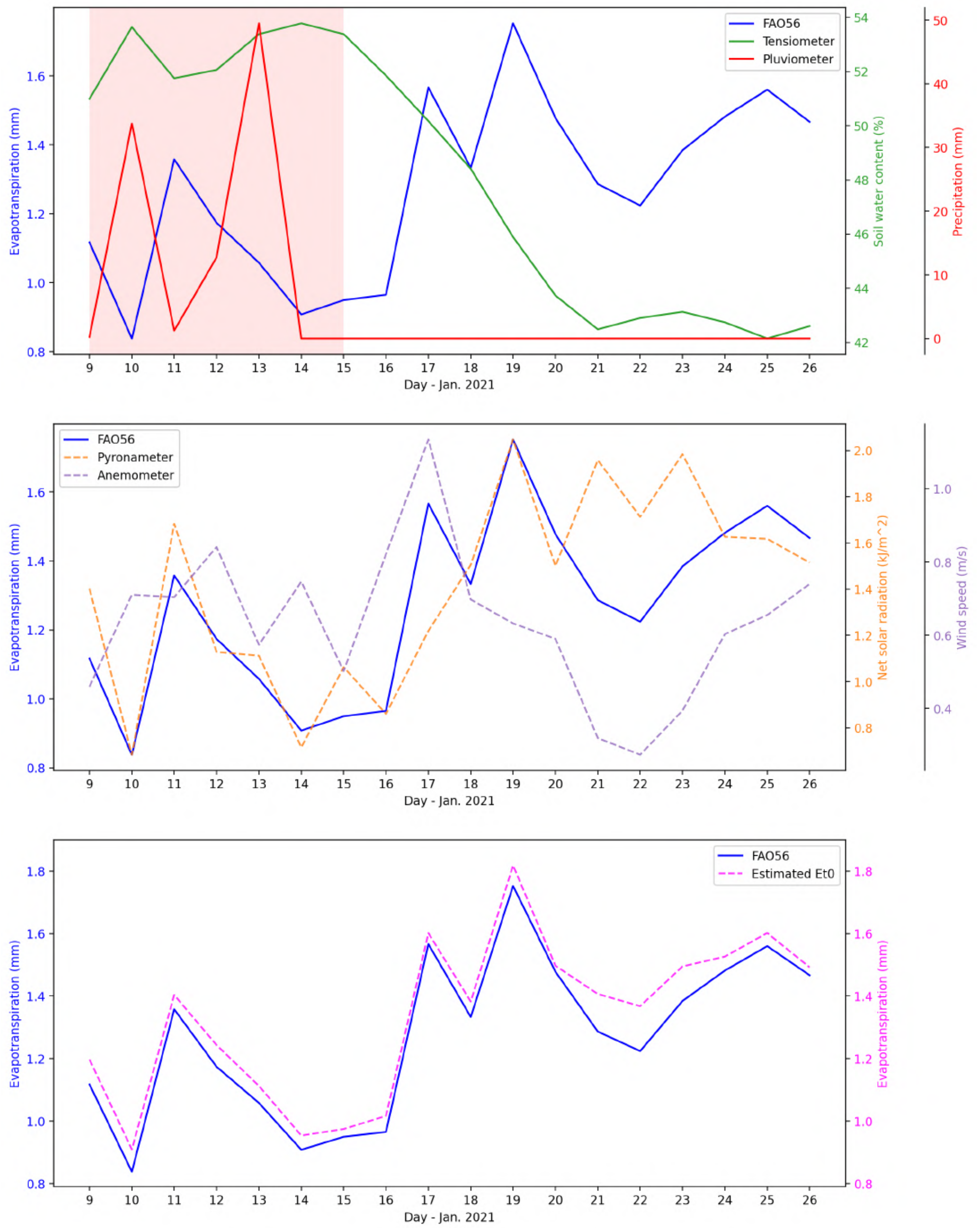
Figure 24 presents the relationships between rainfall, estimated soil water content, and estimated evapotranspiration on the top panel, based on in situ measurements from a tensiometer monitoring the substrate from the surface to a depth of 20 cm, and from a pluviometer, with both sampling data from the 9th to the 25th of January 2021. In the highlighted period (from the 9th to the 15th), the observed rainfall is enough to keep the soil water content close to its maximum capacity ($\approx 59\%$). From the 16th onward, when there is no observed rainfall, the curves for soil water content and evapotranspiration have a symmetrical relationship: the higher the estimated evapotranspiration, the faster the reduction in soil water content. This characteristic against the in situ measurements reinforces the Penman–Monteith method as adequate to model the consumptive use of water due to evaporation and transpiration.

The middle panel of Figure 24 presents the influence on the estimated evapotranspiration by the two most relevant climatological variables, according to the MAP values in Equation (1). This panel exhibits the correlations for evapotranspiration trends with wind speed ($\beta_{13} \approx 0.113$) and net incident solar radiation ($\beta_{14} \approx 0.498$). These relationships of the target variable with the two most relevant climatological factors, according to the Bayesian regression model, combined with the adherence observed in the first panel against the in situ measurements, suggest the adequacy of the proposed method to represent the hydrological phenomenon, even under scenarios where the traditional method cannot be employed.

Lastly, the third panel compares the analytical time series for evapotranspiration according to the Penman–Monteith metric and the results from Equation (1) over the same data set. These three panels suggest that this method models the expected analytical time series and the underlying real-world dynamics.

These results support the adequacy of the proposed method in tracking the analytical FAO-56 equation and functioning as a framework to model dynamic and complex phenomena based on non-linear, non-stationary time series and their underlying factors.

Figure 24 – Graphical representation of the climatological water balance compared to in situ samplings from the 9th to the 26th of January 2021. The top panel contains a time series for the two principal factors of the climatological water balance—rainfall and evapotranspiration—compared to the dynamics of the soil water percentage. The area highlighted in red represents the days when the water intake to the plantation due to rainfall was sufficient to counteract the amount of water evaporated and transpired. The middle panel presents the interaction between the estimated evapotranspiration, wind speed, and incident net solar radiation, the latter two being the most relevant factors for evapotranspiration estimation based on the results presented in Section 5.3. Lastly, the bottom panel presents the estimated evapotranspiration based on the FAO-56 method compared to the values obtained by Equation (1), applied to the same data set during the established timeframe.

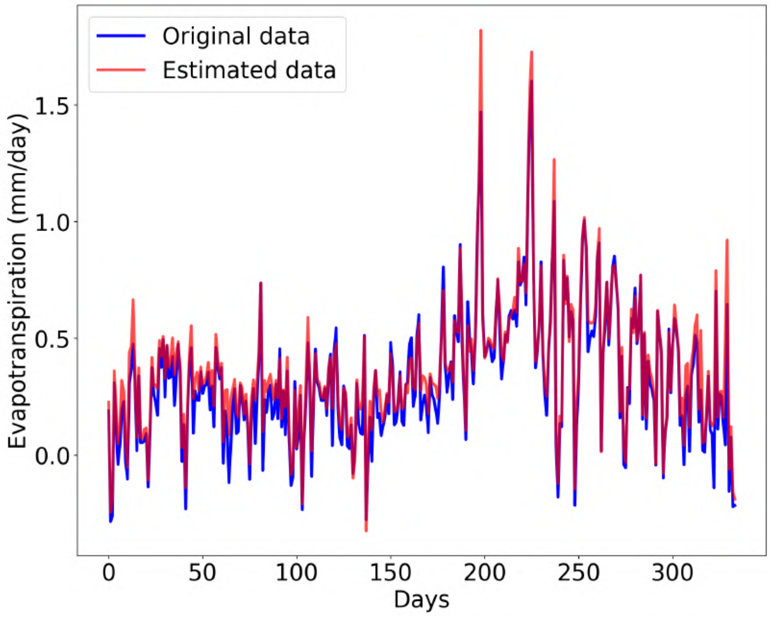


Source: Extracted from Ribeiro *et al.* (2023).

5.5 STOCHASTIC EVAPOTRANSPIRATION INFERENCE WITH COMPLETE OR INCOMPLETE DATA

To evaluate the model's adequacy in estimating evapotranspiration over a data set with new information, Figure 25 presents a graphical comparison of the natural logarithm of the evapotranspiration calculated with the Penman–Monteith equation against the time series obtained from Equation (1). This new data set, comprising 334 samples not used in the learning process, contains complete information on all the climatological variables presented in Section 4.2.

Figure 25 – Graphical comparison of the reference evapotranspiration from the Penman–Monteith equation over the complimentary data set against that obtained via the regression presented in Equation (1). The RMSE metric between the natural logarithm of both series is 0.087 $\mathrm{mm} \cdot \mathrm{day}^{-1}$, consistent with the results in Table 12.



Source: Extracted from Ribeiro *et al.* (2023).

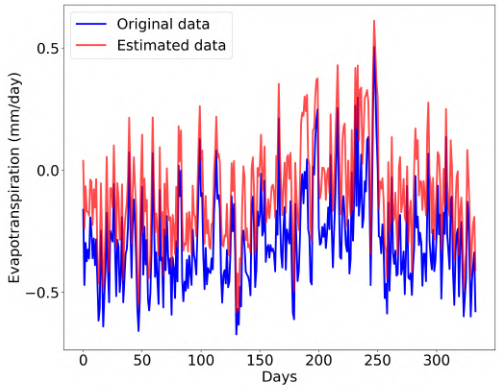
Another way to check the model's performance is to compare the expected time series for the Penman–Monteith equation against the result from Equation (1) when dealing with climatological data from regions with different climate profiles. Figure 26a–d provide a graphic comparison similar to that in Figure 25, but this time it is applied to the cities described in Table 1, with RMSE metrics and distances from Patrocínio presented in Table 5. As expected, the RMSE error for each region increases as the distance from the original city increases, reinforcing the notion that Penman–Monteith equation relies on local climatological factors for an accurate estimate.

Table 5 – Error metrics for the natural logarithm of the reference evapotranspiration derived from the Penman–Monteith equation versus the values from Equation (1), based on the climatological data from different cities in Brazil with different climate profiles compared to Patrocínio/MG.

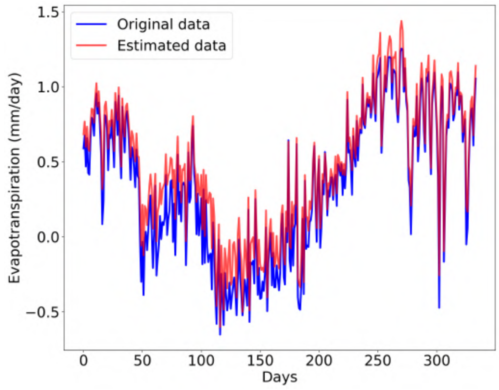
City	**RMSE ($\mathbf{mm \cdot day^{-1}}$)**	**Distance to Patrocínio/MG (km)**
Rio Preto da Eva/AM	0.202	2278
Cabaceiras/PB	0.140	1723
Porangatu/GO	0.102	659
Santo Augusto/RS	0.185	1202

Source: Extracted from Ribeiro *et al.* (2023).

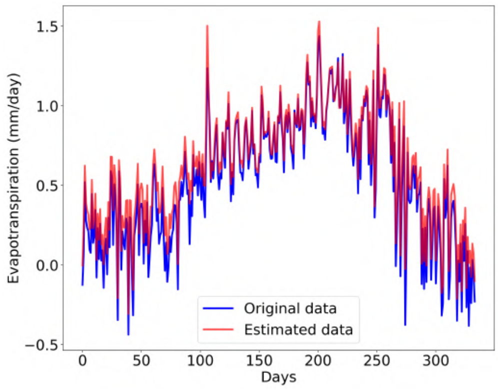
Figure 26 – Graphical comparison of the reference evapotranspiration from the Penman–Monteith equation over the complimentary data set for each different city presented in Table 1, against the value obtained via the regression presented in Equation (1) over that city's climatological time series; (**a**) presents the comparison for Rio Preto da Eva/AM, which achieved RMSE $= 0.202$ mm $\cdot$ day^{-1}; (**b**) presents the comparison for Cabaceiras/PB, which achieved RMSE $= 0.140$ mm $\cdot$ day^{-1}; (**c**) presents the comparison for Porangatu/GO, which achieved RMSE $= 0.102$ mm $\cdot$ day^{-1}; lastly, (**d**) presents the comparison for Santo Augusto/RS, which achieved RMSE $= 0.185$ mm $\cdot$ day^{-1}.



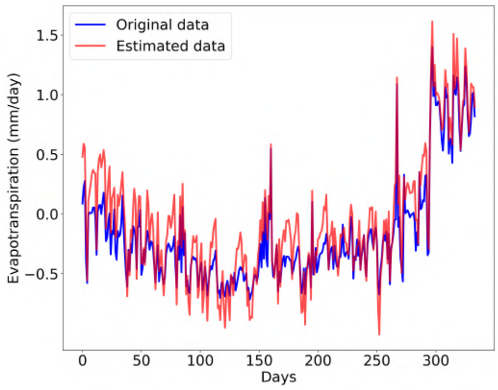
(a) Rio Preto da Eva/AM.



(b) Cabaceiras/PB.



(c) Porangatu/GO.



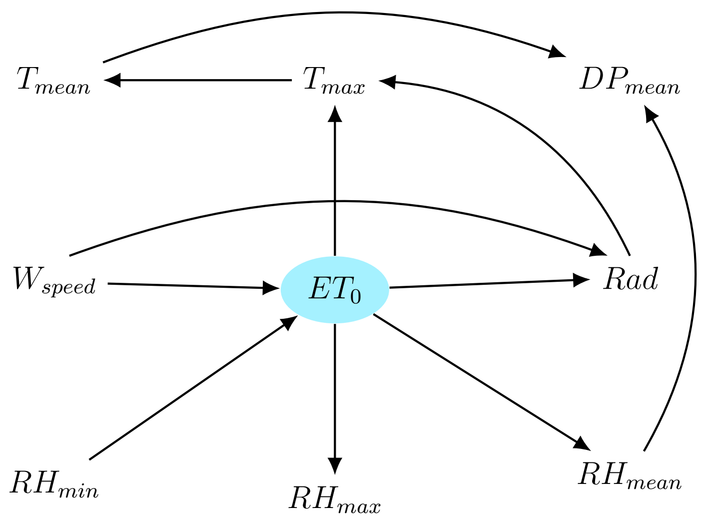
(d) Santo Augusto/RS.

Source: Extracted from Ribeiro *et al.* (2023).

One shortcoming of every analytical approach to estimating evapotranspiration is their prerequisite of having all the required climatological variables available. Under the Bayesian framework, the BN provides a probabilistic approach to supply the most probable value for missing data based on complete or incomplete evidence from the system.

Figure 27 presents a graphical representation of the BN's DAG, which provides the best posterior likelihood estimates based on the original dat aset containing 1000 samples. The nodes represent only the random variables relevant to the Bayesian regression from Table 4 instead of the complete set of fifteen variables. This approach reduces the graph search space for the learning algorithm and employs only the relevant climatological variables, which decreases the number of required components and sensors. The evapotranspiration node is highlighted in blue; the omission of the conditional probability tables from the BN is intentional due to the high dimensionality of each one.

Figure 27 – Bayesian network for estimating the natural logarithm of reference evapotranspiration from the climatological variables most frequently present as relevant, according to the set of experiments with natural logarithm transformation.



Source: Extracted from Ribeiro *et al.* (2023).

To build a BN that presents the best likelihood to the climatological and evapotranspiration data, the authors employed both the standard $K2$ greedy algorithm (Cooper; Herskovits, 1992) and the proposed bio-inspired score-and-search algorithm (Ribeiro *et al.*, 2022b).

The relevant nodes and the amount of required data were consonant with the results of the set of sixty experiments described in Section 4.3, resulting in eight climatological variables, alongside evapotranspiration as the target component and a data set with 1000 samples.

The parameters for both methods were the default ones, with the Bayesian Dirichlet equivalence with the uniform prior metric (BDeu) as a scoring function and k-fold (k = 5) cross-validation. Both methods resulted in the same graphical structure.

After learning the conditional probability tables based on the maximum likelihood of the model over the training data set, the BN framework represented in Figure 27 allows for queries about the most probable state of a climatological variable, providing the Bayesian regression analysis with a new level of flexibility to deal with missing data. For instance, if a meteorological data collector experiences a failure in its anemometer, an estimate of the most probable value for wind speed may allow for the use of the underlying methods that require this variable, albeit within an error margin. Figure 28 presents the natural logarithm of the expected evapotranspiration time series against the values obtained from Equation (1) after replacing each climatological variable from the original data set with values inferred from the BN.

As expected, wind speed estimates based on the BN provide a time series with the lowest adherence to the baseline time series. This result occurs because the wind speed is one of the independent variables in the BN, resulting in the inference task for this variable becoming a query for the most probable value that would induce the observed values for the other variables in the system. This type of inverted query increases the error margin for the estimated value and, alongside the relatively high MAP value for wind speed's beta coefficient, culminates in the amplification of the inference error on the final estimate for evapotranspiration when wind speed data are lacking.

Table 6 presents the RMSE metric for the evapotranspiration estimates according to Equation (1), considering the removal and subsequent inference of each one of the climatological variables from the BN in Figure 27 and the expected evapotranspiration time series. The variables presenting the highest error values are wind speed and minimum relative humidity, which is reasonable within the BN structure as both components are independent elements of the BN. The lowest error values are from the maximum relative humidity component, likely due to its position on the graphical structure, where only one random variable influences it, and it does not influence any others, and due to the lowest absolute MAP value of the beta parameter for maximum relative humidity in Equation (1).

The results in Table 6 suggest that the BN approach provides efficient estimates for the required variable under scenarios lacking a single component and allows for the maintenance and usage of methods that may require these missing factors.

Table 6 – Error metric for the natural logarithm of the reference evapotranspiration metric from the Penman–Monteith equation against the value from Equation (1), after the removal of each relevant climatological variable and the inference of its value from the BN in Figure 27.

Removed and Inferred Variable	**Evapotranspiration RMSE ($\mathbf{mm \cdot day^{-1}}$)**
T_{mean}	0.087
T_{max}	0.092
RH_{mean}	0.087
RH_{max}	0.074
RH_{min}	0.156
DP_{mean}	0.085
W_{speed}	0.286
Rad	0.096

Source: Extracted from Ribeiro *et al.* (2023).

5.6 INFLUENCE DIAGRAM FOR IRRIGATION

The outcome from an MCMC estimation process over water-soil content from the monitored location suggests a normal distribution with parameters $\mu \approx 0.498$ and $\sigma \approx 0.064$ as the best curve to estimate tensiometer readings. Based on these results and the optimal matric potential set as 10 kPa, the most probable value (MPV) for the matric potential is $\theta(\psi)_{MPV} = 49.8\%$ and the optimal value is $\theta(\psi^*) \approx 50.4\%$, which suggest that this soil's location is running on water deficit.

The climatological water balance corroborates this data, as the Bernoulli distribution over the outcomes “deficit” and “surplus” results in an 87% deficit against a 13% surplus probability. As such, the most probable value for the climatological water balance is “deficit”.

With the graphical model and scenarios discussed in Section 4.7, samples from a distribution $\mathcal{N}(0.498, 0.004)$ to model matric potential indirectly measured by tensiometers and samples from a distribution $Ber(0.870)$ act as each scenario’s input. A set random number generator seed allows for experiment reproducibility and result comparison between different scenarios.

The first scenario allows for model validation regarding the irrigation decision based on matric potential estimate according to eq. 9. This model ideally suggests irrigation if, and only if, the matric potential is below the optimal threshold of 50.4%, as lower matric potential corresponds to lower water-soil content readily available. Table 7 presents the recommendation for a small set of samples from the matric potential distribution. The results are consistent with the expected values, suggesting an initial model conforming.

Table 7 – Irrigation suggestions based on the first scenario proposed, solely a function of the matric potential estimates. The results are consistent with the expected outcome as the lower the matric potential, less water is readily available for the agricultural produce.

Matric potential	Turn on?
51.8%	No
43.1%	Yes
54.6%	No
55.8%	No
37.3%	Yes

Source: Developed by the author.

The second scenario provides irrigation recommendations based on eq. 10 and eq. 11 for agricultural produce with arbitrary radicular depth, which may not coincide with the tensiometer depth. Let this system monitor a culture with radicular depth $h = 25$ cm and tensiometers at depths $h_1 = 20$ cm and $h_2 = 40$ cm. The matric potential estimator at 40 cm is $\mathcal{N}(0.452, 0.002)$ based on a new MCMC parameter estimation process.

Table 8 contains the results from the same matric potential samples regarding the tensiometer at 20 cm with the addition of samples from the estimator for the one at 40 cm. As the soil moisture in this place gets lower for the deeper strata, some recommendations to not irrigate should be to turn on the irrigation system. This change in recommendation is present in the first and fourth rows in Table 8 compared to those entries in Table 7, which is the first indicative of adequacy in compounding elements in this system.

The third scenario provides a complete stochastic modelling for the irrigation recommendation based on conditional probability tables, as in Figure 11. The results presented in Table 9 show that this approach is also suited for instances requiring the final recommendation and a degree of certainty for the suggested outcome. As pictured in Table 9, the recommendation becomes

Table 8 – Irrigation suggestions based on the second scenario, which also takes into consideration the estimates for a second tensiometer and a produce with radicular system at a depth of 25 cm.

Matric potential (20 cm)	Matric potential (25 cm)	Matric potential (40 cm)	Turn on?
51.8%	49.0%	48.1%	Yes
43.1%	45.7%	46.6%	Yes
54.6%	55.7%	56.1%	No
55.8%	44.8%	41.1%	Yes
37.3%	38.3%	38.7%	Yes

Source: Developed by the author.

more confident the farthest the tensiometer estimate is from the optimal matric potential, one way or another.

Table 9 – Irrigation suggestions based on the stochastic relationship in Figure 11. Under this scenario the model is shown to be more or less assertive in suggesting irrigation based on the difference between the estimated matric potential and the optimal value for this variable.

Matric potential	Turn on?
51.8%	No (85%)
43.1%	Yes (55%)
54.6%	No (55%)
55.8%	No (85%)
37.3%	Yes (85%)

Source: Developed by the author.

The fourth scenario combines the information from the stratified water balance with the tensiometer estimates, according to eq. 12. Under this scenario, Table 10 shows that the decision prioritises the irrigation under "high deficit" ($Bal = -3$) and no irrigation under "high surplus" ($Bal = 3$), independently from the tensiometer outcome. Once the climatological water balance becomes closer to a neutral state, the tensiometer reading performs the decisive role regarding the recommendation.

The results from the fourth scenario illustrate the potential of this stochastic model to adjust to any utility function defined by the system's agent.

The fifth and final scenario adds a deterministic contextual factor, providing flexibility regarding the system's recommendation. This toolset is relevant if the agent wants to submit the plantation to a water stress regime for days. Although not the most complex scenario, it provides another layer of external control over the recommendation system. Table 11 contains such recommendations, where the outcome is always to not irrigate independently from the tensiometer or the climatological water balance estimates until the n^{th} day, after which the system suggests irrigation.

Table 10 – Results from the fourth scenario considering both the tensiometer estimates alongside the stratified water balance, according to eq. 12. These results suggest a preference ordering for the random variables, where the tensiometer estimate is considered only if the climatological water balance does not present an extreme value.

	Turn on?		
Matric potential	$Bal = -3$	$Bal = 3$	Else
51.8%	Yes	No	No
43.1%	Yes	No	Yes
54.6%	Yes	No	No
55.8%	Yes	No	No
37.3%	Yes	No	Yes

Source: Developed by the author.

Table 11 – Crop resilience.

Soil-water content	Turn on?
51.8%	No
43.1%	No
54.6%	No
55.8%	No
37.3%	No
...	...
n^{th} day	Yes

Source: Developed by the author.

6 CONCLUSION

Most of the results from this thesis lead to three distinct international publications: two conference papers and one journal article. Annexes A, B and C contain these papers' integral versions.

This thesis' first contribution – a bio-inspired algorithm to obtain a Bayesian Network exclusively from data – achieved good performance dealing with synthetic data for small and medium benchmark data sets. As the results show, this method obtained a structure with the same BDeu score as the original *Sachs* network, a graph with a marginally better score than the *Asia* network's original and a sound structure for the *Alarm* network within a reasonable timeframe.

When evaluating the efficacy of the Bayesian Network in modelling climatological real-world phenomena, this thesis shows that this approach can model and allows for probabilistic inference of these phenomena considering complete or incomplete evidence and based on a model extracted from the same city or another locality. The results also point to non-uniform relevancy among the climatological variables for the inference task, which may allow for a reduction in the graph search space.

This thesis concludes that the Bayesian regression toolset presents a suitable approach based on Bayesian statistics to model the numerical dependency on the evapotranspiration target variable on its most relevant climatological factors based on non-synthetic data. The non-linearity and non-stationarity from the real-world time series highlight the prominence of this finding. Furthermore, the obtained regression equation better fits the theoretical gold standard compared to well-established analytical methods, and the regression equation provides a good evapotranspiration estimate performance when applied to different geographical localities and within different climates.

Complimentary to the evapotranspiration's numerical modelling, this thesis also establishes the Bayesian regression model's adequacy to reproduce the real-world phenomenon indirectly inspected via tensiometers. The results further confirm the importance of the net incident solar radiation and wind speed to the evapotranspiration dynamics.

Another conclusion from this thesis revolves around the Bayesian framework providing stochastic inference based on complete or incomplete evidence. As the results suggest, this approach brings an adaptability layer to estimate evapotranspiration, even absenting mandatory climatological data for the analytical methods. The possible lack of data for any required climatological variable is a central issue from the gold standard and its related classical approximations, turning them unfeasible.

Lastly, this thesis shows a reasonable Influence Diagram to guide the irrigation decision based on the relevant stochastic variables and utility functions that try to represent policies from an external agent. The algorithmic suggestions are consistent with the expected based on each scenario. Adversely, eliciting utility functions for this problem is challenging due to these functions' subjective nature and the myriad of components which may impact the final decision. These elements may include the agricultural cultures' characteristics and the socio-economic environment surrounding the producer.

REFERENCES

ABDELMONEIM, A. A. *et al.* Internet of Things (IoT) for soil moisture tensiometer automation. **Micromachines**, Switzerland, v. 14, n. 2, p. 263, 2023. ISSN 2072-666X.

ALIBABAEI, K.; GASPAR, P. D.; LIMA, T. M. Modeling soil water content and reference evapotranspiration from climate data using deep learning method. **Applied Sciences**, Switzerland, 2021.

ALLEN, R. *et al.* **Crop evapotranspiration**: guidelines for computing crop water requirements. [S.l.]: Food and Agriculture Organization of the United Nations, 1998. (FAO irrigation and drainage paper 56). ISBN 9789251042199.

AMIRASHAYERI, A. *et al.* Evapotranspiration estimation using hybrid and intelligent methods. **Soft Computing**, Germany, 2023. ISSN 1432-7643, 1433-7479.

ANAYAH, F. M.; KALUARACHCHI, J. J. Estimating global distribution of evapotranspiration and water balance using complementary methods. **Atmosphere-Ocean**, Downsview, v. 57, n. 4, p. 279–294, 2019.

BALTENSPERGER, A.; JOLY, K. Using seasonal landscape models to predict space use and migratory patterns of an arctic ungulate. **Movement Ecology**, Cham, v. 7, n. 18, 2019.

BARZKAR, A.; NAJAFZADEH, M.; HOMAEI, F. Evaluation of drought events in various climatic conditions using data-driven models and a reliability-based probabilistic model. **Natural Hazards**, Dordrecht, v. 110, p. 1931–1952, 2022.

BASAGAOGLU, H.; CHAKRABORTY, D.; WINTERLE, J. Reliable evapotranspiration predictions with a probabilistic machine learning framework. **Water**, Switzerland, 2021.

BAZRAFSHAN, O. *et al.* Predicting crop yields using a new robust bayesian averaging model based on multiple hybrid ANFIS and MLP models: Predicting crop yields using a new robust bayesian averaging model. **Ain Shams Engineering Journal**, Egypt, 2022.

BEINLICH, I. A. *et al.* The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. In: HUNTER, J.; COOKSON, J.; WYATT, J. (Ed.). **AIME 89**. Berlin: Heidelberg, 1989. p. 247–256. ISBN 978-3-642-93437-7.

BENAVIDES, J. G.; LOPEZ, D. Formula para el calculo de la evapotranspiracion potencial adaptada al tropico (15º N - 15º S). **Agronomia Tropical**, Venezuela, v. 20, n. 5, p. 335–345, 1970.

BIABI, H.; MEHDIZADEH, S. A.; SALMI, M. S. Design and implementation of a smart system for water management of lilium flower using image processing. **Computers and Electronics in Agriculture**, Netherlands, v. 160, p. 131–143, 2019. ISSN 0168-1699.

BOX, G. E. P.; COX, D. R. An analysis of transformations. **Journal of the Royal Statistical Society: Series B (Methodological)**, United Kingdom, v. 26, n. 2, p. 211–243, 1964. ISSN 0035-9246.

BROOM, B. M.; DO, K.-A.; SUBRAMANIAN, D. Model averaging strategies for structure learning in bayesian networks with limited data. **BMC Bioinformatics**, United Kingdom, v. 13, n. S13, p. S10, 2012. ISSN 1471-2105.

BUNTINE, W. Theory refinement on bayesian networks. In: D'AMBROSIO, B. D.; SMETS, P.; BONISSONE, P. P. (Ed.). **Uncertainty Proceedings 1991**. San Francisco (CA): Morgan Kaufmann, 1991. p. 52–60. ISBN 978-1-55860-203-8.

BYERS, E. A. *et al.* Drought and climate change impacts on cooling water shortages and electricity prices in Great Britain. **Nature Communications**, United Kingdom, 2020.

CHEN, L. *et al.* Construction of an agricultural drought monitoring model for karst with coupled climate and substratum factors—a case study of Guizhou province, China. **Water**, Switzerland, 2023.

CHIDZALO, P.; NGARE, P. O.; MUNG'ATU, J. K. Trivariate stochastic weather model for predicting maize yield. **Journal of Applied Mathematics**, United Kingdom, 2022.

COOPER, G. F.; HERSKOVITS, E. A bayesian method for the induction of probabilistic networks from data. **Machine Learning**, United States, v. 9, n. 4, p. 309–347, 1992. ISSN 0885-6125, 1573-0565.

COX, L. A. Modernizing the Bradford Hill criteria for assessing causal relationships in observational data. **Critical Reviews in Toxicology**, United States, v. 48, n. 8, p. 682–712, 2018.

CRUZ, T. A. C. D.; MARQUES, P. A. A. Low-cost irrigation management system: improving data confidence through artificial intelligence. **Engenharia Agrícola**, Brazil, v. 43, n. spe, p. e20210164, 2023. ISSN 1809-4430, 0100-6916.

DAI, J. *et al.* An improved evolutionary approach-based hybrid algorithm for bayesian network structure learning in dynamic constrained search space. **Neural Computing and Applications**, United Kingdom, Aug 2018.

DAYAL, K. S.; DEO, R. C.; APAN, A. A. Book chapter. **Development of copula-statistical drought prediction model using the standardized precipitation-evapotranspiration index**. [S.l.: s.n.], 2019.

DOBRIYAL, P. *et al.* A review of the methods available for estimating soil moisture and its implications for water resource management. **Journal of Hydrology**, Netherlands, v. 458–459, p. 110–117, 2012. ISSN 0022-1694.

DOUCET, A.; WANG, X. Monte Carlo methods for signal processing: a review in the statistical signal processing context. **IEEE Signal Processing Magazine**, United States, v. 22, n. 6, p. 152–170, 2005. ISSN 1053-5888.

DRAPER, N. R.; SMITH, H. **Applied regression analysis**. 2d ed. ed. New York: Wiley, 1981. (Wiley series in probability and mathematical statistics). ISBN 9780471029953.

DURIGHETTO, N.; BERTASSELLO, L. E.; BOTTER, G. Eco-hydrological modelling of channel network dynamics - part 1: Stochastic simulation of active stream expansion and retraction. **Royal Society Open Science**, United Kingdom, 2022.

ELMER, J. *et al.* Bayesian outcome prediction after resuscitation from cardiac arrest. **Neurology**, United States, v. 99, n. 11, p. e1113–e1121, 2022. ISSN 0028-3878, 1526-632X.

FAO. **FAOSTAT: Crops and livestock products**. 2021. Available at: https://www.fao.org/faostat/en/#data/QCL. Access in: 31 Oct. 2023.

FONTENELLE, T. H.; FUCKNER, M. A.; SOARES, S. R. A. Produção animal e usos consuntivos da água no brasil. In: _____. **Produção animal e recursos hídricos**: Uso da água nas dimensões quantitativa e qualitativa e cenários regulatórios e de consumo. 1. ed. Brasil: Embrapa, 2021. p. 323. ISBN 978-65-87380-37-7.

GAVAHI, K. *et al.* Multivariate assimilation of remotely sensed soil moisture and evapotranspiration for drought monitoring. **Journal of Hydrometeorology**, United States, 2020.

GELMAN, A.; HILL, J.; VEHTARI, A. **Regression and other stories**. United Kingdom: Cambridge University Press, 2020.

GENUCHTEN, M. V. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. **Soil Science Society of America Journal**, United States, v. 44, 1980.

GHASEMINEJAD, A.; UDDAMERI, V. Physics-inspired integrated space-time artificial neural networks for regional groundwater flow modeling. **Hydrology and Earth System Sciences**, Germany, 2020.

GHIAT, I.; MACKEY, H. R.; AL-ANSARI, T. A review of evapotranspiration measurement models, techniques and methods for open and closed agricultural field applications. **Water**, Switzerland, v. 13, n. 18, 2021. ISSN 2073-4441.

GOUDIE, R. J. B. *et al.* MultiBUGS: A parallel implementation of the BUGS modeling framework for faster bayesian inference. **Journal of Statistical Software**, United States, v. 95, n. 7, 2020. ISSN 1548-7660.

GRANATA, F. Evapotranspiration evaluation models based on machine learning algorithms—a comparative study. **Agricultural Water Management**, Netherlands, v. 217, p. 303–315, 2019. ISSN 0378-3774.

GROSS, T. J. *et al.* An analytical threshold for combining bayesian networks. **Knowledge-Based Systems**, Netherlands, 2019. ISSN 0950-7051.

HANSEN, E. A. An integrated approach to solving influence diagrams and finite-horizon partially observable decision processes. **Artificial Intelligence**, Netherlands, v. 294, p. 103431, 2021. ISSN 00043702.

HARGREAVES, G. H.; SAMANI, Z. A. Reference crop evapotranspiration from temperature. **Applied engineering in agriculture**, United States, v. 1, n. 2, p. 96–99, 1985.

HUGHES, A. *et al.* A probabilistic risk-based decision framework for structural health monitoring. **Mechanical Systems and Signal Processing**, United Kingdom, v. 150, p. 107339, 2021. ISSN 0888-3270.

HUI, E. *et al.* Bayesian networks as a novel tool to enhance interpretability and predictive power of ecological models. **Ecological Informatics**, Netherlands, v. 68, p. 101539, 2022. ISSN 1574-9541.

IANISHI, P. *et al.* Probability on graphical structure: A knowledge-based agricultural case. **Annals of Data Science**, Germany, p. 1–19, 2020.

ILIĆ, M. *et al.* Irrigation water fitness assessment based on bayesian network and FAO guidelines. **Irrigation and Drainage**, United Kingdom, v. 71, n. 3, p. 665–675, 2022. ISSN 1531-0353, 1531-0361.

JAYASHREE, T. *et al.* Prediction of reference crop evapotranspiration: Empirical and machine learning approaches. In: _____. **Handbook of Hydroinformatics**. [S.l.]: Elsevier, 2023. p. 253–268. ISBN 9780128219614.

JEHANZAIB, M. *et al.* Modeling hydrological non-stationarity to analyze environmental impacts on drought propagation. **Atmospheric Research**, Netherlands, 2023.

KHASHEI-SIUKI, A. *et al.* Simulation of potential evapotranspiration values based on vine copula. **Meteorological Applications**, United Kingdom, 2021.

KOCIAN, A. *et al.* Bayesian sigmoid-type time series forecasting with missing data for greenhouse crops. **Sensors**, Switzerland, 2020.

KOCIAN, A. *et al.* Dynamic bayesian network for crop growth prediction in greenhouses. **Computers and Electronics in Agriculture**, Netherlands, 2020.

KOLLER, D.; FRIEDMAN, N. **Probabilistic graphical models**: Principles and techniques. [S.l.]: MIT Press, 2009. ISBN 978-0-262-01319-2.

KOLLURU, V.; KOLLURU, S. Development and evaluation of pre and post integration techniques for enhancing drought predictions over India. **International Journal of Climatology**, United Kingdom, 2021.

LAURITZEN, S. L.; SPIEGELHALTER, D. J. Local computations with probabilities on graphical structures and their application to expert systems. **Journal of the Royal Statistical Society. Series B (Methodological)**, United Kingdom, v. 50, n. 2, p. 157–224, 1988. ISSN 0035-9246.

LINACRE, E. T. A simple formula for estimating evaporation rates in various climates, using temperature data alone. **Agricultural Meteorology**, Netherlands, v. 18, n. 6, p. 409–424, 1977. ISSN 0002-1571.

LU, J. *et al.* A comparison of six potential evapotranspiration methods for regional use in the southeastern United States. **JAWRA Journal of the American Water Resources Association**, United States, v. 41, n. 3, p. 621–633, 2005.

LUCAS, P. d. O. e. *et al.* Reference evapotranspiration time series forecasting with ensemble of convolutional neural networks. **Computers and Electronics in Agriculture**, Netherlands, v. 177, p. 105700, 2020. ISSN 0168-1699.

MAJUMDAR, P. *et al.* Demand prediction of rice growth stage-wise irrigation water requirement and fertilizer using bayesian genetic algorithm and random forest for yield enhancement. **Paddy and Water Environment**, Germany, 2023.

MANTEL, S.; DONDEYNE, S.; DECKERS, S. World reference base for soil resources (WRB). In: _____. **Encyclopedia of Soils in the Environment**. [S.l.]: Elsevier, 2023. p. 206–217. ISBN 9780323951333.

MARINESCU, R.; LEE, J.; DECHTER, R. A new bounding scheme for influence diagrams. **Proceedings of the AAAI Conference on Artificial Intelligence**, United States, v. 35, n. 13, p. 12158–12165, 2021. ISSN 2374-3468.

MARTIN, C. J.; COCHRAN, T. H.; KATSURA, S. Tuberculosis, emphysema, and bronchitis. **American Review of Respiratory Disease**, United States, v. 97, n. 6P1, p. 1089–1094, 1968.

MCELREATH, R. **Statistical rethinking**: A bayesian course with examples in R and Stan. [S.l.]: Chapman and Hall/CRC, 2020.

MITCHELL, E. G. *et al.* Bayesian network analysis reveals resilience of the jellyfish aurelia aurita to an irish sea regime shift. **Scientific Reports**, United Kingdom, v. 11, n. 1, p. 3707, 2021. ISSN 2045-2322.

MOURA, A. D. *et al.* **Nota técnica no. 001/2011/SEGER/LAIME/CSC/INMET**. Instituto Nacional de Meteorologia, 2019. Available at: http://www.cemtec.ms.gov.br/wp-content/uploads/2019/02/Nota_Tecnica-Rede_estacoes_INMET.pdf. Access in: 07 Jan. 2024.

MURPHY, K. **A brief introduction to graphical models and bayesian networks**. 1998. Available at: https://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html. Access in: 22 Sep. 2023.

MUTTI, P. R. *et al.* Long-term meteorological drought characterization in the São Francisco watershed, Brazil: A climatic water balance approach. **International Journal of Climatology**, United Kingdom, v. 42, n. 16, p. 8162–8183, 2022. ISSN 0899-8418, 1097-0088.

NAKAGAKI, T.; YAMADA, H.; TOTH, A. Maze-solving by an amoeboid organism. **Nature**, United Kingdom, v. 407, n. 6803, p. 470–470, 2000.

NATORI, K. *et al.* Constraint-based learning bayesian networks using bayes factor. In: SUZUKI, J.; UENO, M. (Ed.). **Advanced Methodologies for Bayesian Networks**. Cham: Springer International Publishing, 2015. p. 15–31.

NOURANI, V.; ELKIRAN, G.; ABDULLAHI, J. Multi-step ahead modeling of reference evapotranspiration using a multi-model approach. **Journal of Hydrology**, Netherlands, 2020.

OGUNRINDE, A. T. *et al.* Application of artificial neural network for forecasting standardized precipitation and evapotranspiration index: A case study of Nigeria. **Engineering Reports**, United States, 2020.

PEARL, J. **Probabilistic reasoning in intelligent systems**: Networks of plausible inference. [S.l.]: Elsevier, 1988. ISBN 9780080514895.

PEARL, J. **Causality**: models, reasoning, and inference. United States: Cambridge University Press, 2000. ISBN 0521773628.

PEARL, J.; MACKENZIE, D. **The book of why**: the new science of cause and effect. First edition. [S.l.]: Basic Books, 2018. ISBN 9780465097609.

QUILTY, J. *et al.* Bayesian extreme learning machines for hydrological prediction uncertainty. **Journal of Hydrology**, Netherlands, 2023.

RADMANESH, F.; ESMAEILI-GISAVANDANI, H.; LOTFIRAD, M. Climate change impacts on the shrinkage of Lake Urmia. **Journal of Water and Climate Change**, United Kingdom, 2022.

RAGHUNATH, H. M. **Hydrology**: principles, analysis and design. Rev. 2nd ed. New Delhi: New Age International (P) Ltd., Publishers, 2006. ISBN 9788122418255.

RIBEIRO, V. P. *et al.* Bayesian network for hydrological model: an inference approach. In: **2022 International Joint Conference on Neural Networks (IJCNN)**. Italy: IEEE, 2022. p. 1–7. ISBN 9781728186719.

RIBEIRO, V. P. *et al.* An improved bayesian network super-structure evaluation using physarum polycephalum bio-inspiration. In: **Proceedings of the XXIV Brazilian Congress on Automation**. Brazil: [s.n.], 2022.

RIBEIRO, V. P. *et al.* A stochastic bayesian artificial intelligence framework to assess climatological water balance under missing variables for evapotranspiration estimates. **Agronomy**, Switzerland, v. 13, n. 12, 2023. ISSN 2073-4395.

ROSSIT, D. *et al.* Stochastic forestry harvest planning under soil compaction conditions. **Journal of Environmental Management**, Netherlands, 2021.

SACHS, K. *et al.* Causal protein-signaling networks derived from multiparameter single-cell data. **Science**, United States, v. 308, n. 5721, p. 523–529, 2005. ISSN 0036-8075.

SAHA, A.; SEKHARAN, S.; MANNA, U. Evaluation of capacitance sensor for suction measurement in silty clay loam. **Geotechnical and Geological Engineering**, Netherlands, v. 38, n. 4, p. 4319–4331, 2020. ISSN 0960-3182, 1573-1529.

SANTOS, H. G. d. *et al.* **Brazilian Soil Classification System.** Brasília, DF: Embrapa, 2018., 2018. ISBN 9788570358219. Available at: http://www.infoteca.cnptia.embrapa.br/handle/doc/1094001. Access in: 21 Oct. 2023.

SCHÖN, T. *et al.* Physarum learner: A bio-inspired way of learning structure from data. **Expert Systems with Applications**, United Kingdom, v. 41, n. 11, p. 5353–5370, 2014.

SCHWARZ, G. Estimating the dimension of a model. **The Annals of Statistics**, United States, v. 6, n. 2, p. 461–464, 1978.

SHEDBALKAR, K. H.; MORE, D. S. Bayesian regression for solar power forecasting. In: **2022 2nd International Conference on Artificial Intelligence and Signal Processing (AISP)**. Vijayawada, India: IEEE, 2022. p. 1–4. ISBN 9781665442909.

SILVA, L. A. P. *et al.* Mapeamento dos efeitos das mudanças climáticas na evapotranspiração de referência em cenários futuros no semiárido brasileiro. **Revista Brasileira de Geografia Fisica**, Brazil, 2023.

SILVA, R. O. *et al.* Response of water balance components to changes in soil use and vegetation cover over three decades in the eastern Amazon. **Frontiers in Water**, Switzerland, v. 3, 2021. ISSN 2624-9375.

SINGH, P.; KAUR, A. A systematic review of artificial intelligence in agriculture. In: _____. **Deep Learning for Sustainable Agriculture**. [S.l.]: Elsevier, 2022. p. 57–80. ISBN 9780323852142.

SMALLMAN, T. L.; WILLIAMS, M. Description and validation of an intermediate complexity model for ecosystem photosynthesis and evapotranspiration: ACM-GPP-ETv1. **Geoscientific Model Development**, Germany, 2019.

SMITH-FERGUSON, J.; BEEKMAN, M. Who needs a brain? slime moulds, behavioural ecology and minimal cognition. **Adaptive Behavior**, United Kingdom, p. 105971231982653, 2019.

SUZUKI, J. A construction of bayesian networks from databases based on an MDL principle. In: _____. **Uncertainty in Artificial Intelligence**. [S.l.]: Elsevier, 1993. p. 266–273. ISBN 9781483214511.

TALEBI, H.; SAMADIANFARD, S.; KAMRAN, K. V. Estimation of daily reference evapotranspiration implementing satellite image data and strategy of ensemble optimization algorithm of stochastic gradient descent with multilayer perceptron. **Environment, Development and Sustainability**, Germany, 2023.

TERO, A.; KOBAYASHI, R.; NAKAGAKI, T. Physarum solver: A biologically inspired method of road-network navigation. **Physica A: Statistical Mechanics and its Applications**, Netherlands, v. 363, n. 1, p. 115–119, 2006.

TINA, G. M.; NICOLOSI, C. F. A statistical assessment of water availability for hydropower generation in the context of adequacy analyses. **Applied Sciences**, Switzerland, 2023.

TURC, L. Estimation of irrigation water requirements, potential evapotranspiration: A simple climatic formula evolved up to date. **Annales Agronomiques**, France, v. 12, n. 1, p. 13–49, 1961.

ULLAH, I. *et al.* Spatiotemporal characteristics of meteorological drought variability and trends (1981–2020) over South Asia and the associated large-scale circulation patterns. **Climate Dynamics**, Germany, 2023.

United Nations Department for Economic and Social Affairs. **Sustainable Development Goals Report 2020**. United Nations, 2020. ISBN 9789211014259. Available at: https://unstats.un.org/sdgs/report/2020/The-Sustainable-Development-Goals-Report-2020.pdf. Access in: 05 Jul. 2023.

WALLS, S. *et al.* Prediction of actual evapotranspiration by artificial neural network models using data from a Bowen ratio energy balance station. **Neural Computing and Applications**, United Kingdom, 2020.

WANG, D. *et al.* Comparison of machine learning models to predict lake area in an arid area. **Remote Sensing**, Switzerland, 2023.

WANG, T. *et al.* Hybrid machine learning approach for evapotranspiration estimation of fruit tree in agricultural cyber-physical systems. **IEEE Transactions on Cybernetics**, United States, 2023.

WU, K. *et al.* Ground-penetrating radar full-wave inversion for soil moisture mapping in trench-hill potato fields for precise irrigation. **Remote Sensing**, Switzerland, v. 14, n. 23, p. 6046, 2022. ISSN 2072-4292.

YIN, L. *et al.* Improving terrestrial evapotranspiration estimation across China during 2000–2018 with machine learning methods. **Journal of Hydrology**, Netherlands, 2021.

ZAREI, A. R.; MAHMOUDI, M. R.; SHABANI, A. Investigating of the climatic parameters effectiveness rate on barley water requirement using the random forest algorithm, bayesian multiple linear regression and cross-correlation function. **Paddy and Water Environment**, Germany, 2021.

ZHAO, L. *et al.* Selecting essential factors for predicting reference crop evapotranspiration through tree-based machine learning and bayesian optimization. **Theoretical and Applied Climatology**, Austria, 2024.

APPENDIX A – COMPLETE TABLE FOR THE ERROR METRICS

Table 12 – The RMSEs and MAEs of all sixty experiments, combining each proposed transformation, original dataset partition, and autoregressive window. From a scoring standpoint, every experiment achieved error metrics within the same order, except for those with the exponential transformation.

Transformation	**Nº of Samples**	**Window**	**RMSE ($mm \cdot day^{-1}$)**	**MAE ($mm \cdot day^{-1}$)**
Linear	1000	0	0.115	0.074
		1	0.112	0.073
		2	0.112	0.073
		3	0.112	0.073
	2500	0	0.118	0.076
		1	0.116	0.075
		2	0.116	0.075
		3	0.116	0.075
	5473	0	0.139	0.090
		1	0.134	0.089
		2	0.134	0.088
		3	0.133	0.088
Box-Cox	1000	0	0.107	0.054
		1	0.105	0.054
		2	0.105	0.054
		3	0.105	0.054
	2500	0	0.097	0.059
		1	0.095	0.058
		2	0.095	0.058
		3	0.095	0.058
	5473	0	0.101	0.070
		1	0.098	0.069
		2	0.097	0.068
		3	0.097	0.068

Table 12 – *Cont.*

Transformation	**Nº of Samples**	**Window**	**RMSE ($\mathbf{mm \cdot day^{-1}}$)**	**MAE ($\mathbf{mm \cdot day^{-1}}$)**
Square root	1000	0	0.094	0.061
		1	0.092	0.061
		2	0.092	0.061
		3	0.092	0.061
	2500	0	0.093	0.062
		1	0.092	0.062
		2	0.092	0.061
		3	0.092	0.061
	5473	0	0.108	0.074
		1	0.105	0.072
		2	0.104	0.072
		3	0.104	0.072
Natural logarithm	1000	0	0.088	0.054
		1	0.086	0.053
		2	0.086	0.054
		3	0.086	0.053
	2500	0	0.088	0.058
		1	0.087	0.057
		2	0.086	0.057
		3	0.086	0.057
	5473	0	0.100	0.070
		1	0.098	0.069
		2	0.097	0.068
		3	0.097	0.068

Table 12 – *Cont.*

Transformation	**Nº of Samples**	**Window**	**RMSE ($\mathbf{mm \cdot day^{-1}}$)**	**MAE ($\mathbf{mm \cdot day^{-1}}$)**
Exponential	1000	0	2.408	1.010
		1	2.296	0.979
		2	2.283	0.988
		3	2.282	0.993
	2500	0	2.190	1.029
		1	2.129	1.007
		2	2.130	1.010
		3	2.128	1.009
	5473	0	3.502	1.609
		1	3.458	1.560
		2	3.430	1.561
		3	3.429	1.560

Source: Developed by the author.

APPENDIX B – COMPLETE TABLE FOR THE RELEVANT VARIABLES

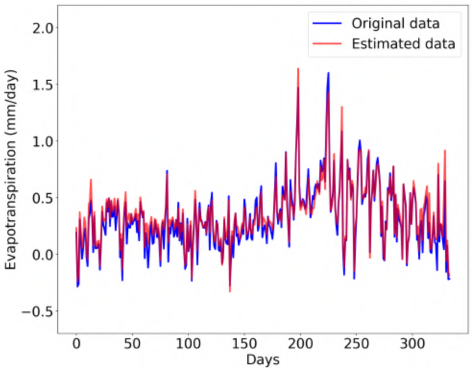
Table 13 – The frequency of each climatological variable found relevant for each set of twelve experiments, according to the transformation applied to the evapotranspiration metric. The variables most frequently present are highlighted in blue.

Variable	Linear	Box-Cox	Square Root	Natural Logarithm
T_{mean}	6	12	8	12
T_{max}	9	12	12	12
T_{min}	3	4	1	4
RH_{mean}	8	10	0	11
RH_{max}	11	8	4	12
RH_{min}	11	12	12	12
DP_{mean}	6	12	8	12
DP_{max}	2	1	0	0
DP_{min}	8	1	1	2
P_{mean}	0	0	0	0
P_{max}	0	0	0	0
P_{min}	0	0	0	0
W_{speed}	12	12	12	12
Rad	12	12	12	12
$Prec.$	0	0	0	0
$ET_0\,[-1]$	9/9	9/9	9/9	9/9
$ET_0\,[-2]$	4/6	4/6	4/6	4/6
$ET_0\,[-3]$	1/3	1/3	1/3	2/3

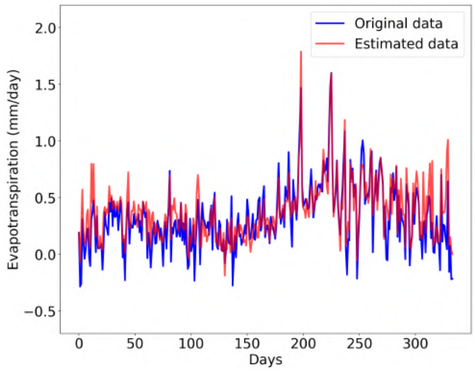
Source: Developed by the author.

APPENDIX C – GRAPHICAL COMPARISON BETWEEN THE EXPECTED TIME SERIES AND INFERRED DATA

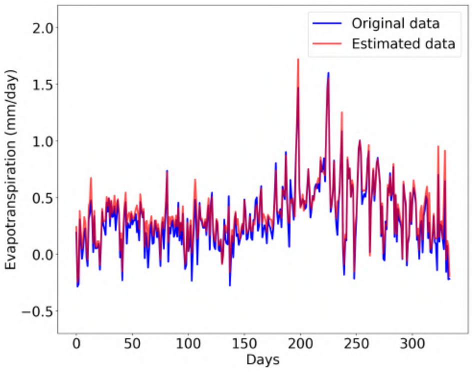
Figure 28 – A graphical comparison between the expected time series and the results obtained after each climatological variable was removed from the dataset; subsequent inference of the components based on the Bayesian Network structure.



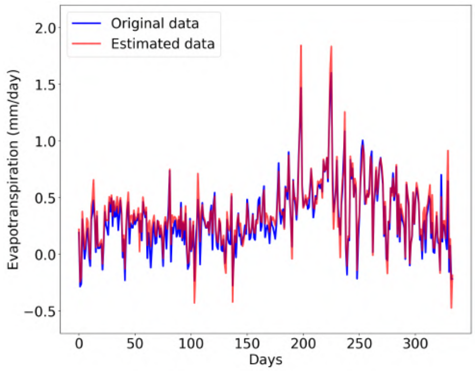
(a) Maximum relative humidity.



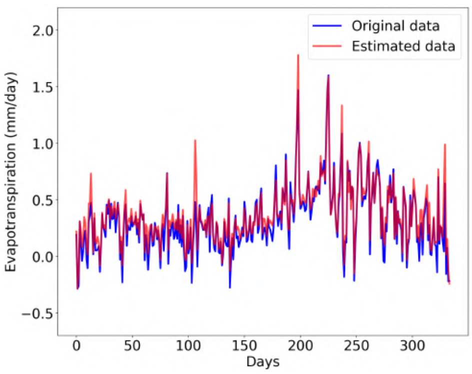
(b) Minimum relative humidity.



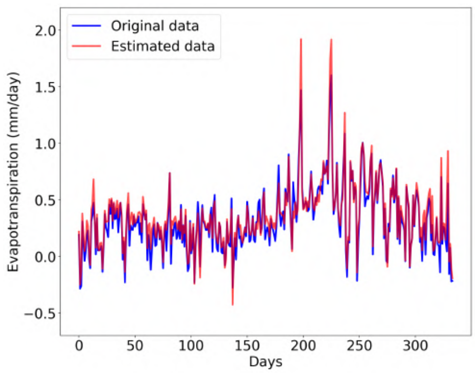
(c) Mean relative humidity.



(d) Mean dew point temperature.

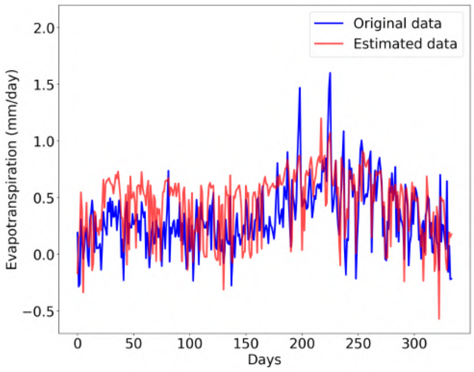


(e) Maximum temperature.

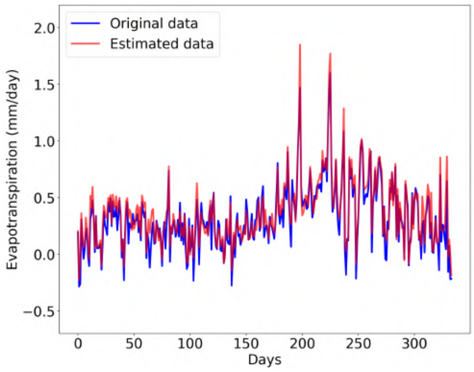


(f) Mean temperature.

Figure 28 – *Cont.*



(g) Wind speed.



(h) Net solar radiation.

Source: Developed by the author.

ANNEX A – *AN IMPROVED BAYESIAN NETWORK SUPER-STRUCTURE EVALUATION USING* PHYSARUM POLYCEPHALUM *BIO-INSPIRATION*

This Annex contains the complete paper presented at the XXIV Brazilian Congress on Automatics (Congresso Brasileiro de Automática - CBA/2022).

This paper provides and discusses an improved implementation for a bio-inspired score-and-search method to build a Bayesian Network super-structure from data.

As discussed in the original paper, one of the main contributions of this work is its fast convergence to a good structure, from structural and score-based standpoints.

Another contribution regards its innate ability to deal with this class of NP-hard super-exponential problems. As the original *Physarum Solver* finds the optimal path between two nodes even on a complete graph, this method can deal with graphs with many nodes and edges.

An Improved Bayesian Network Super-Structure Evaluation using *Physarum polycephalum* Bio-inspiration

Vitor P. Ribeiro [*] **Benvindo R. Pereira Jr.** [**] **Carlos D. Maciel** [**]
José Antônio P. Balestieri [*]

[*] *School of Engineering, São Paulo State University (UNESP), Guaratinguetá, SP, Brazil (e-mail: {vitor.p.ribeiro;jose.perrella}@unesp.br).*
[**] *Department of Electrical Engineering, São Paulo University (USP), São Carlos, SP, Brazil (e-mail: {brpjunior;cdmaciel}@usp.br)*

Abstract: Bayesian Networks are a powerful tool to assess association relationships between variables on a given observed system. Despite their usage on several areas, the task of obtaining such structure exclusively from observational data is a well-documented NP-hard problem. *Physarum Learner* is a method to obtain a Bayesian Network from a data set, based on the foraging behaviour of the *Physarum polycephalum*, a member of the Myxomycetes class. However, the original method overlook crucial characteristics of this combinatorial problem. In the work here presented, we implement three changes to the *Physarum Learner* original algorithm and apply this improved method on three well-known benchmark data sets. Our results indicate a faster convergence point identification, while achieving good structures from a scoring standpoint.

Keywords: Equivalence class; Bayesian networks; Physarum Autonomous Optimisation; Bio-inspired algorithm.

1. INTRODUCTION

Bayesian Networks (BNs) are a type of Probabilistic Graphical Model that can be used to build models from data or expert knowledge, and provide a robust and mathematically coherent framework for the analysis of several kind of problems (Uusitalo, 2007). This method is becoming increasingly popular to model uncertain and complex domains such as medicine (Shen et al., 2022), traffic engineering (Wang et al., 2022), and risk assessment (Ruiz-Tagle et al., 2022). This wide usage is mostly due to its function to point at associations between a large set of variables from the observed system, in an explainable format.

Despite the usefulness of BNs in analysing complex systems, the construction of the structure which best comprehends the associations between a variables of a given system, referred to as an optimal solution, is a well-known NP-hard problem (Chickering et al., 2004). Most of the algorithms proposed to build BNs from data fall into three categories: constrain-based (CB) methods, search-and-score (SS) methods, and hybrid methods.

The constrain-based (CB) methods employ conditional independence (CI) tests to create their structure (Koller and Friedman, 2009), which in one hand scale up better to problems with many variables, but on the other provide a structure with lower accuracy when compared to the next two methods (Natori et al., 2015).

The score-and-search (SS) methods use any suitable scoring function to rate a candidate Bayesian network, and search for a structure that optimises said score through adjustments on the elements of said network (Cooper and Herskovits, 1992). Although the results with this approach are usually more accurate for smaller networks, it becomes unfeasible when dealing with a problem with many variables, due to the super-exponential increase on the number of candidate structures that it would have to score (Gross et al., 2019).

Lastly, hybrid methods such as the one presented in Dai et al. (2018) make use of techniques from the CB approach in order to reduce the search space, then use procedures based on the SS approach to examine this constrained search space and provide a solution. This class of methods allow for a fast reduction of the search space, which is then used as a small subset of structures to be scored and searched upon.

The *Physarum Learner* algorithm employs a hybrid approach to provide a Bayesian Network exclusively from data (Schön et al., 2014), based on a Physarum Autonomic Optimisation (PAO) (Tero et al., 2007) methodology. Although this algorithm provides good structures according to a given score, as provided in the original article, the procedure described there presents an approach that can be improved in terms of score convergence, computational time, and coverage of the evaluated structures.

As described in Chickering (2013), the procedure to build a BN exclusively from fitting the structure to the data according to a given score function generates a family of graphical structures with the same or very similar scores. This family of structures is called an *equivalence class*, where the most complete structure that encompasses all possible graphs from the same equivalence class is referred to as the *super-structure* for the association relationships on that data set.

The original *Physarum Learner* performs an extensive search on the space of possible structures given a starting point, this initial configuration is then used throughout all iterations of the method, which causes the search procedure to be trapped in a local optima, a common issue with heuristics with a single starting point for the solution (Mao et al., 2021).

The exhaustive nature of the PAO approach on *Physarum Learner* taken causes the algorithm to conduct the evaluating process for new candidate structures regardless on whether or not the score has reached a plateau. Given the nature of super-structures from the learning process exclusively from data, one could interrupt the learning process once small changes in the network provide a negligible change in score, or none at all (Wang et al., 2021).

In this paper we present a modified version of the *Physarum Learner* algorithm, which is comprised of three new steps in order to tackle some of the perceived issues on the original method. First, our approach is to abort the search procedure once the score has stagnated for a number of iterations. Second, as the main goal is to minimise a given score, we include one step to change the direction or presence of edges after each iteration, to cover a larger search space for a suitable super-structure. Third, in order to minimise the issue of premature convergence to a local optima, our approach uses different starting points across iterations, which helps to create diverse starting points to the subsequent search-and-score steps.

Our method is shown to achieve promising results on small and medium benchmark data sets, with a fast convergence and sound structures when compared to the expected result. We also provide our code for our improved implementation of the *Physarum Learner* using *Python* in an open-source approach, as well as the data sets used for the testing and validation of the method.

Section 2 summarises the theoretical background on Bayesian networks, and presents the relevant basis on the original *Physarum Learner* method, as well as the description of the data sets used. Section 3 contains the algorithm description and the improvements made when compared to the original version. Section 4 shows the results of the algorithm applied to the previously mentioned data sets and the pertinent discussion. Finally, Section 5 contains the conclusions achieved from the presented results.

2. THEORY

2.1 Bayesian Network

Formally, the function of a Bayesian Network (BN) is to estimate a joint probability distribution (JPD) based on n discrete random variables, from a data set D taken from the observed system (Koller and Friedman, 2009).

A BN $B = (G, \Theta)$ is defined by a Directed Acyclic Graph (DAG) G, and a set of conditional probability tables Θ. These structures are referred to as BN's *qualitative* and *quantitative* parts, respectively (Neapolitan, 2004).

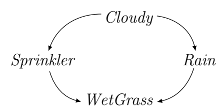
For the *qualitative* part of a BN, let G be a DAG comprised of a set of vertices $V = \{v_i\}_{i=1,2,...,n}$, and a set of directed edges $E = \{e_{ij}\}_{i,j=1,2,...,n; i \neq j}$. Each v_i represents a random variable from the data set D, and each $e_{ij} \in \{0, 1\}$ indicates the absence or presence of a directed edge pointing from the vertex v_i to the vertex v_j.

For the *quantitative* part, $\Theta = \{\theta_i\}_{i=1,2,...,n}$ is the set of conditional probability distributions for each one of the n observed variables. As such, $\theta_i = p(v_i|P_i)$, where P_i denotes the set of k vertices where $e_{ki} = 1$ from the aforementioned graph G, also known as the parents' set of the vertex v_i on the graph.

Figure 1 represents a Bayesian Network of a toy problem known as the "Sprinkler Network", presented in Murphy (1998). Each vertex represents a binary variable of interest, the edges illustrate associations among the variables, and every conditional probability table for every random variable is explicitly presented.

Cloudy = F	Cloudy = T
0.5	0.5

Cloudy	Sprinkler = F	Sprinkler = T
F	0.5	0.5
T	0.9	0.1



Cloudy	Rain = F	Rain = T
F	0.8	0.2
T	0.2	0.8

Sprinkler	Rain	WetGrass = F	WetGrass = T
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

Figure 1. Bayesian Network for the "Sprinkler" data set, adapted from Murphy (1998). The fact that all four variables are binary allow for a representation of false (F) or true (T) for the occurrence of each one. Every variable on this system has a conditional probability table associated with them, which represents the chance of a given outcome taking into account the variables associated with the one being observed and their outcomes, defined on the main text as $p(v_i|P_i)$.

Given both the *qualitative* and *quantitative* parts of the network, the joint probability distribution of the variables on B can be estimated through the Bayes' chain rule, defined as follows:

$$p(v_1, v_2, ..., v_n) = \prod_{i=1}^{n} p(v_i|P_i) \quad (1)$$

The main task of an algorithm designed to create a BN from data revolves around an optimisation problem: find a structure B whose joint probability distribution best fits the one from the given data set D provided.

Ref. Broom et al. (2012) shows that one approach to this optimisation problem is to find the DAG which best represents the association relationships between variables present on the data set. The adequacy of a given graphical structure can be estimated through a fitness function that associates a candidate graph with the provided data set. Examples of fitness functions for this scoring procedure include the *Bayesian Information Criterion* (BIC) (Schwarz, 1978), and the *Bayesian Dirichlet equivalence with uniform prior metric* (BDeu) (Buntine, 1991).

2.2 Physarum Learner

Physarum polycephalum ("multi-headed slime mould") belongs to the Amoebozoa group, in the class Myxomycetes (Stephenson and Stempen, 2000). During the active vegetative stage of its life cycle, called "plasmodium", it can extend for up to $30x30cm^2$, and move to a maximum speed of $4cm$ per hour, as a multi-nuclear unicellular being (Vogel and Dussutour, 2016).

The seminal work in Nakagaki et al. (2000) presented the ability of the *Physarum polycephalum* in consistently finding the shortest path between two sources of food in a maze, during its plasmodium stage. This behaviour kindled the interest of researchers on several areas into analyse, understand, and model the behaviour of this brainless being (Oettmeier et al., 2020). A survey on the subject can be found in Gao et al. (2018).

One particular area that benefited from the maze-solving experiment was mathematical modelling and route optimisation. *Physarum Solver* is an adaptive mathematical model which simulates the foraging behaviour of the *P. polycephalum* as presented in Nakagaki et al. (2000), and it is first proposed in Tero et al. (2006).

Physarum Solver's original purpose is to find the optimal path between two vertices on a weighted undirected graph, similar to the classical Dijkstra's algorithm, and was proven to provide the optimal solution for a graph on a Riemannian surface (Miyaji and Ohnishi, 2008). As the work in Nakagaki et al. (2000) describes the behaviour of the *P. polycephalum* on a maze, the weighted undirected graph used on the *Physarum Solver* algorithm is referred to as *Physarum* maze, and the initial and final vertices on the path-finding problem are called source and sink nodes, respectively.

Physarum Learner (Schön et al., 2014) is an algorithm which employs the *Physarum Solver* modelling as an auxiliary structure to build a BN from a data set. Figure 2 presents a flow chart for the original *Physarum Learner* algorithm, containing its main steps.

Given a data set containing n variables, the *Physarum Learner* algorithm starts by creating two graph structures: a fully-connected undirected graph, which will act as the "maze" for the *Physarum Solver* computations, and a directed graph, upon which the BN will be built on. Both structures contain n nodes each.

For the initial set up of the *Physarum* maze, the parameters for the edges on the *Physarum* maze are set to standard values. This structure will serve as an "external memory" for the creation of the BN, and its parameters will be updated accordingly on the next steps.

As for the directed graph, the algorithm creates edges connecting every pair of nodes, on a random direction. This decision presents no issues from a structural standpoint, given that the structure will be rebuilt after the first iteration of the algorithm.

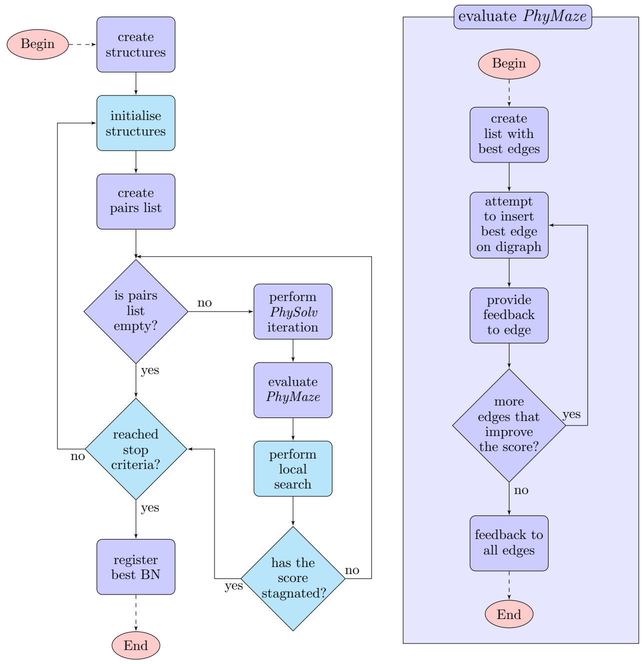


Figure 2. Flow chart depicting the *Physarum Learner* (Schön et al., 2014) algorithmic structure, with improvements presented in this article. On the left, the main structure is presented. The internal structure for the "evaluate *PhyMaze*" block is expanded on the right, as this block comprehends the interactions between the *PhyMaze* – an undirected graph – with the Bayesian Network been built. The boxes in blue are from the original algorithm, while the boxes in cyan show where the improvements from this work were made.

The original *Physarum Solver* algorithm finds the optimal path between a pair of variables on an undirected graph. For this reason, the *Physarum Learner* algorithm creates a list of every pair of nodes contained on its *Physarum* maze. This ensures that every pair of nodes and, consequently, every pair of variables, will be evaluated during execution.

One pair is chosen at random to represent the input source/sink nodes for the *Physarum Solver* algorithm. The parameters on the edges on the *Physarum* maze will be updated, reflecting the attempt to find the best path between this pair. The edge directly connecting the pair taken as source/sink from the previous step has its parameter set above a threshold value, even if the result of the *Physarum Solver* iteration did not set it as such. This ensures that all the edges will be considered at least once to be part of the BN on the next step.

Once the *Physarum Solver* iteration ends, the algorithm performs an *evaluation* on the resulting maze. This is the step which represents the interfacing between the undirected graph and the Directed Acyclic Graph (DAG), representing the BN.

The evaluation of the maze consists of creating a list containing every edge from the *Physarum* maze whose parameters surpasses the given threshold. All the other edges that don't comply with this criteria based on the *Physarum* maze are removed from the BN, if present.

For every edge present on the edges list, the algorithm attempts to insert the same edge on the DAG. Considering both directions for the edge, if the inclusion of it does not result in a cycle on the graph, and there is an improvement on the score from inserting this edge on given direction, this edge is considered as candidate to be permanently set on the DAG. Once the best connection is found among all the edges on the list, this connection is removed from the list, and an edge is set on the DAG considering the direction which provides the best improvement.

If a connection that provides the best improvement is found, then a feedback is applied to this edge on the *Physarum* maze. This feedback ensures that this edge will be obtained for the next evaluation step, and will not be removed from the DAG structure.

Finally, when there are no more edges on the list that provide a score improvement for the BN, a feedback function is applied to all of them, lowering their parameters in order to prevent the *Physarum Solver* algorithm to include those edges on the shortest path for the next iteration.

One iteration of the algorithm is defined by the list of every pair of nodes on the *Physarum* maze being depleted. Once an iteration has end, the *Physarum* maze as the "external memory" will contain the best edges found so far on the BN, coded on the parameter values for its edges. For the next iteration a new pairs list is created, and this time the learning process take into account the edges already identified as "best" for the BN from the previous iteration.

Let T be the total number of iterations on the *Physarum Learner* algorithm. Considering one time unit as the time for the algorithm to compute all the steps for a single pair of nodes on the complete pairs list, one can describe the time complexity for the algorithm $O(n)$ as:

$$\begin{aligned} O(n) &\approx T \cdot n^2 \\ O(n) &\propto n^2 \end{aligned} \tag{2}$$

As the algorithm performs all T iterations over the entire pairs list, the best case is the same as the worst case, as is the average case.

3. MATERIAL AND METHOD

3.1 Improved Physarum Learner

The original *Physarum Learner* algorithm exhibits good performance to build a Bayesian Network (BN) from a data set. Nevertheless, some of its presented features lead to excessive time consumption to perform the task at hand, as can be observed from the time complexity estimated on Equation 2.

The blocks coloured in cyan on Figure 2 represent steps modified from the original or entirely new blocks, on the algorithm here named *Improved Physarum Learner*

First, the evaluation step on the original article assumes a "greedy" approach to build a BN based on the *Physarum* maze, as it considers the best edge one at a time to include on the DAG. This procedure, although similar to the classical algorithm K2, does not contemplate the fact that, according to Equation 1, the joint probability is function of *every* parent for every variable on the network.

This definition implies that, once a new set of edges is inserted onto the graphical structure, a local search should be performed over the obtained graph. This extra step is required to determine the best direction for each edge, given that all the other edges are already present on the directed graph.

The implementation here proposed tries to change the direction of every edge on the obtained BN, on a random order, searching in the graph neighbourhood for a structure with a better score.

Second, the original algorithm ends an iteration only when its pairs list is empty. This ensures that every one of the approximately n^2 edges is considered to be inserted on the BN at least once. This approach does not take into account that the best structure might have already been found during previous computations, achieving a local optima super-structure.

A procedure to check for the eventual structure stagnation across several evaluation steps is to verify if the score for the obtained structure is kept constant, or within an error margin, for a given number of pairs analysed. The *Improved Physarum Learner* contains a structure which checks for this score stagnation. If the score is identified as constant, or within an error margin, after a percentage of pairs have been analysed, the current iteration is then finished.

Third, the original algorithm relies on the analysis of nodes on the pairs list in a random order, and both the external memory and the obtained BN are kept constant across iterations. The order in which the pairs are considered during an iteration may result in a bias for both structures, to be carried across iterations.

In order to lessen this eventual bias, and to enable new structures to be found, *Improved Physarum Learner* resets the *Physarum* maze as well as the BN structure to their original conditions. The best BN found so far is stored on a separate structure.

Lastly, the original algorithm proposes the edges on the first DAG to be set between all pairs of nodes, in random directions. The computational cost involved in creating all those edges and picking a random direction for each one is prohibitive, specially when dealing with networks with larger number of nodes.

Improved Physarum Learner creates its DAG structures containing only the nodes for the random variables, and no edges connecting them. This is a valid approach due to the structure been overridden at each evaluation step.

This version of the algorithm is implemented in the *Python3* language, due to its broad usage on scientific research, as well as to its open-source licensing. The complete code is readily available upon request.

3.2 Data sets

Here we describe the data sets used as benchmarks for the testing of our improved method.

Asia network This Bayesian Network is proposed in Lauritzen and Spiegelhalter (1988), as a tool to illustrate how the method on the original paper works. The original structure is reproduced here as Figure 3a. The structure consists of eight nodes, and eight edges. Each node codifies a binary variable, which can only assume the values of *True* or *False*.

We chose this structure as one of our benchmarks due to its small number of nodes, and the fact that it is a weakly connected graph, i. e. there is a path between every pair of nodes if the direction of the edges is omitted.

Sachs network This BN is presented in Sachs et al. (2005), where the authors study possible causal relationships on a protein-signalling environment. The structure obtained on the original article is reproduced here as Figure 5a.

This structure is a good candidate to test the method based on its relatively small size, and the fact that this graph is not connected as the one presented in the *Asia* structure. This is an interesting characteristic due to the usage of a fully-connected undirected graph as an auxiliary structure for the method.

Alarm network This network is presented in Beinlich et al. (1989), and encodes a medical monitoring system, built from the authors' expertise and medical textbook literature.

The graph contains thirty-seven nodes, each representing one out of three category of variable: a diagnostic, a measurement, or an intermediate variable. The values for all nodes are represented as a discreet value, with two or three levels each.

This is a particularly interesting network to be used as a benchmark due to the expert knowledge that it encodes. Each of the "diagnostic" type of variables is independent from the others, and are supposed to have no parents on the graphical structure, given the assumption that an identified condition characterised by a diagnostic is the cause of the variables associated to it. As the method here presented deals primarily with a data set, this *a priori* knowledge from the structure is omitted, being used afterwards to analyse the obtained structure.

4. RESULTS AND DISCUSSION

This section is reserved to the presentation and discussion of the results obtained from applying the *Improved Physarum Learner* on the benchmark data sets already presented. The learning process was undertaken on a desktop computer with Fedora OS, processor Intel® Core™ i7-4970 CPU at 3.60GHz, and 32GB of RAM at 1600MHz.

The Bayesian Networks (BNs) super-structures obtained are presented here on their graphical structure, with edges direction purposefully omitted. The colour scheme for the colouring of the edges is as follows: a green edge represents a correct edge, pointing in the correct direction; a blue edge represents a correct edge, but pointing in the opposite direction of the original; and a red edge represents an edge that should not exist. This correctness is based on the original graphical structures for each data set.

The justification for this representation is based primarily on the works of Pearl and Mackenzie (2018) and Cox Jr. (2018). Ref. Pearl and Mackenzie (2018) arguments that a finite data set can never encompasses all of the relationships of association from a system within itself. This also includes the direction of the edges in the obtained structure.

Furthermore, Cox Jr. (2018) argues that a Bayesian Network when built exclusively from data should be used as a starting point for an expert to analyse the system, as a result of the data set insufficiency on incorporating all of the correct associations.

4.1 Results for the Asia data set

Figure 3b presents the graph obtained for the *Asia* data set. As the data set with the least variables, the method was able to find all correct edges, on the correct directions, with exception of the edge *Smoking* → *Lung cancer*, which was found to be on the opposite direction.

The extra edge *Asia* → *Either* can be validated as an expected edge from the concept of "d-separation". The same reasoning can be applied to the extra edges *Tuberculosis* → *Xray*, and *Lung cancer* → *Xray*.

Ref. Koller and Friedman (2009) defines d-separation as a collateral edge resulting from the setting of the middle random variable on a association chain. Consider the chain $X \to Y \to Z$, with X, Y, and Z random variables. As X influences Y, and Y influences Z, if the variable Y has its value fixed, the influence from X is "felt" by Z as a collateral effect.

Although the extra edge *Tuberculosis* → *Bronchitis* can not be directly explained using the concept of d-separation, the association between these medical conditions is well established on the medical literature (Martin et al., 1968).

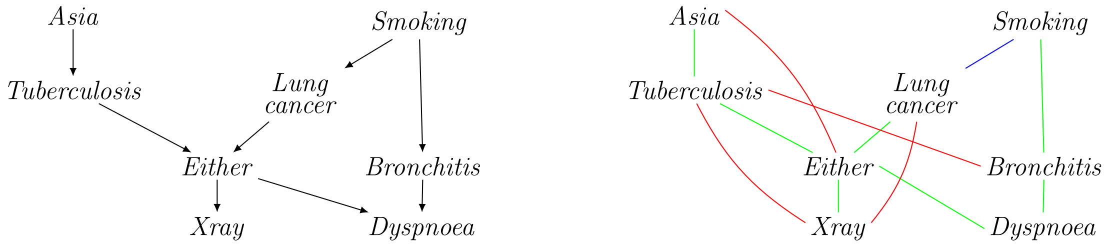
For the fast convergence of the method, Figure 4 presents the convergence of the BDeu score for the best BN structure obtained after each iteration of the algorithm, in red, compared to the score for the original structure, in blue. The convergence was obtained for a score lower than the expected from the original after only three iterations, whereas the original algorithm would require at least twenty-eight iterations to provide an answer for a network with eight nodes. The structure learning process for this structure took on average less than one minute across several tests.

4.2 Results for the Sachs data set

Figure 5b presents the graphical structure obtained for the *Sachs* data set. The *Improved Physarum Learner* algorithm managed to find the same number of edges from the original structure, although some edges were identified in the wrong direction, compared to the original.

Despite the algorithm employing a fully-connected maze as its external memory, it was able to allow the variables *Plcg*, *PIP3*, and *PIP2* to be disconnected from the remaining graph, as the original structure requires.

Figure 6 present the score convergence for the best network obtained after each iteration of the algorithm, in red. In



(a) Original structure.

(b) Obtained structure.

Figure 3. Bayesian networks for the *Asia* data set. On the left, the original structure as described in Lauritzen and Spiegelhalter (1988). On the right, the structure found with the improved method. Although the obtained structure has more edges than the original, the absolute value of its BDeu score is marginally lower (better) than the score from the original structure, given the same data set.

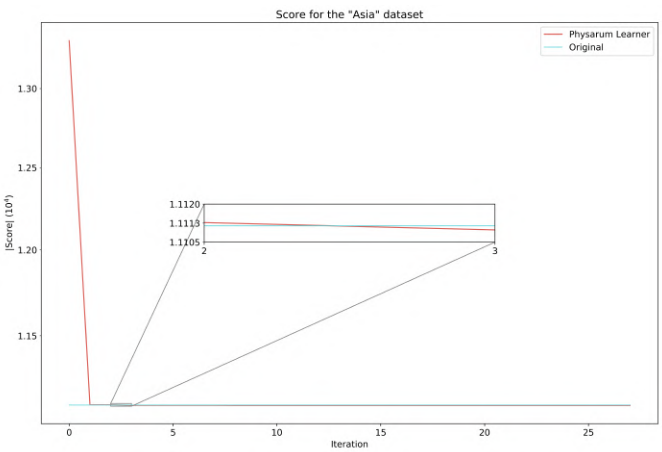


Figure 4. Semi-log representation for the convergence of the BDeu score absolute value for the *Improved Physarum Learner* applied to the *Asia* data set, in red. In blue, the score for the original structure, applied to the same data set. The detail presents the iteration after which the obtained structure presents a score marginally better than the original.

blue, the score for the original network on the same data set. After the thirty-seventh iteration, the score for both structures were exactly the same. The original algorithm requires at least fifty-five iterations to provide an answer for this network. The structure learning process for this structure took on average fifteen minutes across several tests.

This result raises a question concerning the relevance of the direction of the edges when obtaining a score for given structure. From the super-structure approach, both the structures are the best from a score standpoint, requiring an expert in the modelled system to point out which one is the structure which more accurately comprises the associations observed in the real world.

4.3 Results for the Alarm data set

For the *Alarm* data set, our method obtained twenty-four correct edges with the right orientation, eighteen correct edges with the wrong orientation, and eighteen edges that are not present in the original structure. The fact that it contains several extra edges reinforce the argument from Pearl and Mackenzie (2018), considering the details from this data set that were presented in Section 3.2.3. The algorithm did not take into account the fact that, in the real world, one diagnostic influences another diagnostic, although the expert on this medical system presented this constraint on the original article for the network, for instance.

Several instances of wrong edges present on the obtained structure could be dismissed from an expert on the system. For instance, the associations obtained between different diagnostics due to the knowledge that the events represented by these variables should not directly interact with one another. On a similar manner, some instances of wrong edges might be dismissed based on the concept of d-separation presented before.

The obtained model did not encompass only four edges present on the original network. To evaluate the correctness of the structure missing these four edges, conditional independence tests were run on each pair of variables, given the values on the data set. For the four cases the test pointed to independence between the variables within a threshold, which illustrate the impact that the data can have on the perceived best super-structure.

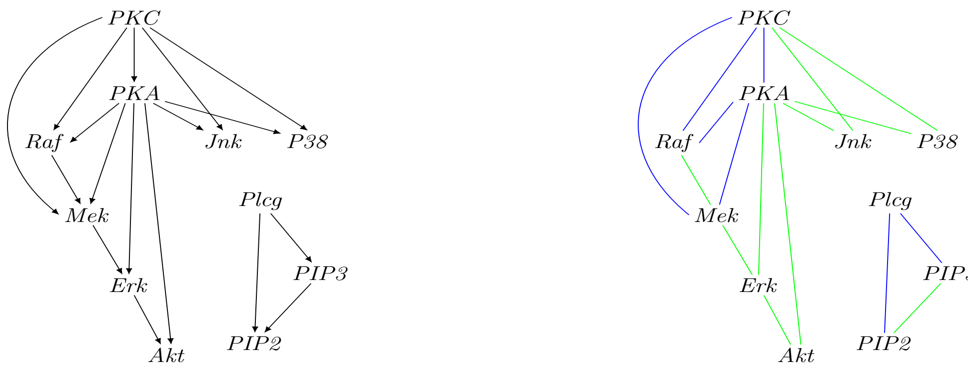
Figure 7 presents in red the score for the best structure found after each iteration, compared to the score for the original structure, presented in blue. As a result of the high number of variables on this data set, the score achieved several plateau levels during the execution. The structure learning process took on average seventy-two hours for this data set. This jump in computational time is expected as the internal *Physarum Solver* algorithm takes on a complete graph containing thirty-seven nodes, and the list of source/sink pairs contains six hundred and sixty-six different combinations.

The convergence test implemented was most valuable for this data set. It allows for the reset of structures after several iterations have passed with stagnated scores, instead of waiting for the end of the pairs list, which may take up to almost six hundred and seventy iterations.

5. CONCLUSION

In this paper we presented a version of the *Physarum Learner* algorithm re-implemented and improved to tackle computational weaknesses presented in the original method.

The *Improved Physarum Learner* achieved sound results when obtaining the Bayesian Networks from benchmark



(a) Traditional structure

(b) Obtained structure.

Figure 5. Bayesian networks for the *Sachs* data set. On the left, the original structure as proposed in Sachs et al. (2005). On the right, the structure found by us. Even with some of the edges on the opposite direction from the original, both structures present equal BDeu score given the same data set. The method was also able to produce a disconnected graph, in conformity to the original structure.

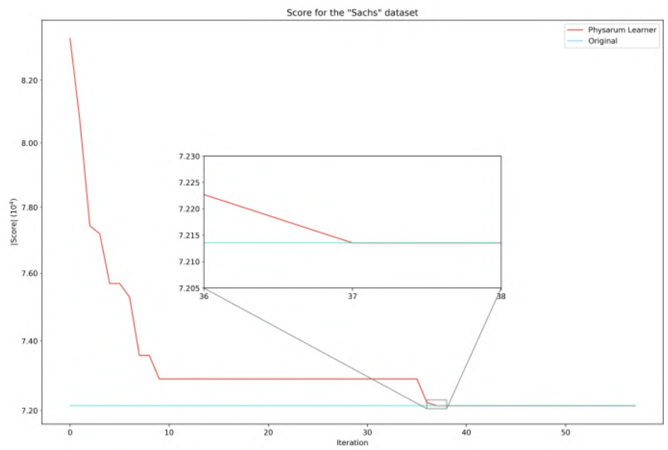


Figure 6. Semi-log representation for the convergence of the BDeu score absolute value for the *Improved Physarum Learner* applied to the *Sachs* data set, in red. In blue, the score for the original structure, applied to the same data set. The detail presents the iteration after which the obtained structure presents the exact same value for score as the original.

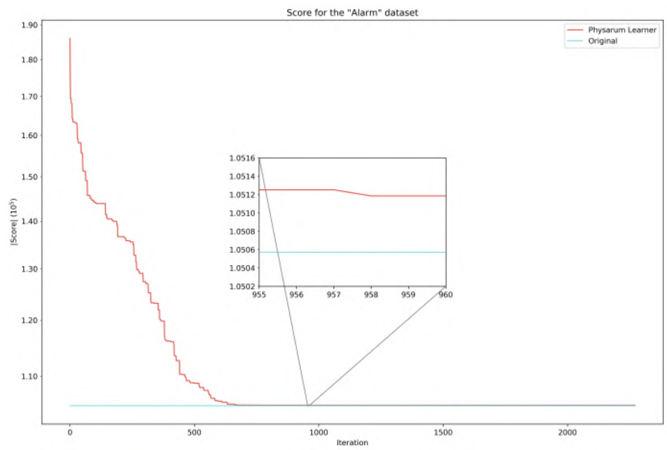


Figure 7. Semi-log representation for the convergence of the BDeu score absolute value for the *Improved Physarum Learner* applied to the *Alarm* data set, in red. In blue, the score for the original structure, applied to the same data set. The score for the obtained structured become stable on several plateaus, before becoming stable on a score slightly higher than the score for the original structure.

data sets, in a reasonable amount of iterations when compared to the original method.

Our results also provides evidence for the intrinsic association between the quality of a structure and the data set used to learn it. Results in the *Asia* network show that is possible to obtain a *marginally* better score with the structure here presented when compared to the structure on the original article. Results for the *Alarm* data set show that a Bayesian Network is only as good as the data allows it to be, given the perceived conditional independence on a subset of variables that have an edge in the original structure.

The results also corroborates the notion that multiple directed acyclic graphs can represent the same association structure from a scoring standpoint. In our case, both the expected and obtained DAG are slightly different on structure – direction of edges – but both present the exact same score given the same data set.

As limitation of the presented method, and future works within this context, we perceive a necessity for a mechanism that reduces the number of the nodes on the *Physarum maze* for each iteration of the *Physarum Solver*, as the total time required for this step across all iterations is exponential according to the number of nodes on the maze.

ACKNOWLEDGEMENTS

Vitor P. Ribeiro and Jose Antonio P. Balestieri are grateful to the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq); Carlos Maciel is partial supported in processes 142390/2019-4, 301853/2018-5. FAPESP 2014/50851-0, CNPq 465755/2014-3 and BPE Fapesp 2018/19150-6.

REFERENCES

Beinlich, I.A., Suermondt, H.J., Chavez, R.M., and Cooper, G.F. (1989). The alarm monitoring system: A case study with two probabilistic inference techniques for belief networks. In J. Hunter, J. Cookson, and J. Wyatt (eds.), *AIME 89*, 247–256. Springer Berlin Heidelberg, Berlin, Heidelberg.

Broom, B.M., Do, K.A., and Subramanian, D. (2012). Model averaging strategies for structure learning in bayesian networks with limited data. *BMC Bioinformatics*, 13(S13), S10.

Buntine, W. (1991). Theory refinement on bayesian networks. In *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'91, 52–60. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Chickering, D.M., Heckerman, D., and Meek, C. (2004). Large-sample learning of bayesian networks is np-hard. *Journal of Machine Learning Research*, 5, 1287–1330.

Chickering, D.M. (2013). Learning equivalence classes of bayesian networks structures. *arXiv:1302.3566*.

Cooper, G.F. and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4), 309–347.

Cox Jr., L.A. (2018). Modernizing the bradford hill criteria for assessing causal relationships in observational data. *Critical Reviews in Toxicology*, 48(8), 682–712.

Dai, J., Ren, J., Du, W., Shikhin, V., and Ma, J. (2018). An improved evolutionary approach-based hybrid algorithm for bayesian network structure learning in dynamic constrained search space. *Neural Computing and Applications*.

Gao, C., Liu, C., Schenz, D., Li, X., Zhang, Z., Jusup, M., Wang, Z., Beekman, M., and Nakagaki, T. (2018). Does being multi-headed make you better at solving problems? A survey of Physarum-based models and computations.

Gross, T.J., Bessani, M., Darwin Junior, W., Araújo, R.B., Vale, F.A.C., and Maciel, C.D. (2019). An analytical threshold for combining Bayesian Networks. *Knowledge-Based Systems*.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*.

Lauritzen, S.L. and Spiegelhalter, D.J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 50(2), 157–224.

Mao, J., Pan, Q., Miao, Z., and Gao, L. (2021). An effective multi-start iterated greedy algorithm to minimize makespan for the distributed permutation flowshop scheduling problem with preventive maintenance. *Expert Systems with Applications*, 169, 114495.

Martin, C.J., Cochran, T.H., and Katsura, S. (1968). Tuberculosis, emphysema, and bronchitis. *American Review of Respiratory Disease*, 97(6P1), 1089–1094.

Miyaji, T. and Ohnishi, I. (2008). Physarum can solve the shortest path problem on riemannian surface mathematically rigorously. *International Journal of Pure and Applied Mathematics*, 47(3), 353–369.

Murphy, K. (1998). A Brief Introduction to Graphical Models and Bayesian Networks. URL `https://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html`.

Nakagaki, T., Yamada, H., and Tóth, (2000). Maze-solving by an amoeboid organism. *Nature*, 407(6803), 470–470.

Natori, K., Uto, M., Nishiyama, Y., Kawano, S., and Ueno, M. (2015). Constraint-based learning bayesian networks using bayes factor. In J. Suzuki and M. Ueno (eds.), *Advanced Methodologies for Bayesian Networks*, 15–31. Springer International Publishing, Cham.

Neapolitan, R.E. (2004). *Learning Bayesian networks*. Prentice Hall series in artificial intelligence. Pearson Prentice Hall.

Oettmeier, C., Nakagaki, T., and Döbereiner, H.G. (2020). Slime mold on the rise: the physics of physarum polycephalum. *Journal of Physics D: Applied Physics*, 53(31), 310201.

Pearl, J. and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic Books, first edition edition.

Ruiz-Tagle, A., Lewis, A.D., Schell, C.A., Lever, E., and Groth, K.M. (2022). Bantera: A bayesian network for third-party excavation risk assessment. *Reliability Engineering & System Safety*, 223, 108507.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A., and Nolan, G.P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721), 523–529.

Schön, T., Stetter, M., Tomé, A.M., Puntonet, C.G., and Lang, E.W. (2014). Physarum Learner: A bio-inspired way of learning structure from data. *Expert Systems with Applications*, 41(11), 5353–5370.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.

Shen, J., Liu, F., Xu, M., Fu, L., Dong, Z., and Wu, J. (2022). Decision support analysis for risk identification and control of patients affected by covid-19 based on bayesian networks. *Expert Systems with Applications*, 196, 116547.

Stephenson, S.L. and Stempen, H. (2000). *Myxomycetes: a handbook of slime molds*. Timber Press.

Tero, A., Kobayashi, R., and Nakagaki, T. (2006). Physarum solver: A biologically inspired method of road-network navigation. *Physica A: Statistical Mechanics and its Applications*, 363(1), 115–119.

Tero, A., Kobayashi, R., and Nakagaki, T. (2007). A mathematical model for adaptive transport network in path finding by true slime mold. *Journal of Theoretical Biology*, 244(4), 553 – 564.

Uusitalo, L. (2007). Advantages and challenges of bayesian networks in environmental modelling. *Ecological Modelling*, 203(3), 312 – 318.

Vogel, D. and Dussutour, A. (2016). Direct transfer of learned behaviour via cell fusion in non-neural organisms. *Proceedings of the Royal Society B: Biological Sciences*, 283(1845).

Wang, S., Patwary, A., Huang, W., and LO, H.K. (2022). A general framework for combining traffic flow models and bayesian network for traffic parameters estimation. *Transportation Research Part C: Emerging Technologies*, 139, 103664.

Wang, Z., Gao, X., Tan, X., and Liu, X. (2021). Determining the direction of the local search in topological ordering space for bayesian network structure learning. *Knowledge-Based Systems*, 234, 107566.

ANNEX B – *BAYESIAN NETWORK FOR HYDROLOGICAL MODEL: AN INFERENCE APPROACH*

This Annex contains the complete paper presented at the 2022 International Joint Conference on Neural Networks (IJCNN), held at the 2022 IEEE World Conference on Computational Intelligence.

This paper presents a Bayesian Network tailored to perform statistical inferences for the evapotranspiration (Et_0) metric over complete or incomplete climatological data, with the Penman-Monteith method as the gold standard.

Another contribution is a brief discussion about the importance of a stochastic modelling to perform the inference of a metric traditionally estimated with analytical approximations.

The proposed method also has the property of allowing for estimates even in a scenario where not all the climatological variables required by the gold standard method are available.

Bayesian Network for Hydrological Model: an inference approach

Vitor P. Ribeiro
School of Engineering
São Paulo State University (UNESP)
Guaratinguetá, SP, Brazil
vitor.p.ribeiro@unesp.br

Ângela S. M. Cunha
Department of Biosystems Engineering
São Paulo University (USP/IBM/C4AI)
Piracicaba, SP, Brazil
angelasilviane@usp.br

Sérgio N. Duarte
Department of Biosystems Engineering
São Paulo University (USP)
Piracicaba, SP, Brazil
snduarte@usp.br

Carlos R. Padovani
Geoprocessing Laboratory
Brazilian Agricultural Research Corporation (EMBRAPA)
Corumbá, MS, Brazil
carlos.padovani@embrapa.br

Patricia A. A. Marques
Department of Biosystems Engineering
São Paulo University (USP)
Piracicaba, SP, Brazil
paamarques@usp.br

Carlos D. Maciel
Department of Electrical Engineering
São Paulo University (USP)
São Carlos, SP, Brazil
carlos.maciel@usp.br

José Antônio P. Balestieri
School of Engineering
São Paulo State University (UNESP)
Guaratinguetá, SP, Brazil
jose.perrella@unesp.br

***Abstract*—According to the Food and Agriculture Organisation, there are growing concerns about the availability and use of water in agriculture. The hydrological model generates a water balance and the resulting value indicates the amount of available water in a given area. The calculation of the water balance is fundamental for the development of new strategies for the management of water resources. One of its main adversities is the estimation of evapotranspiration, which may be considered a fundamental component. This factor considers climatological variables collected from weather stations that are spread over large areas. However, there are frequent cases of long periods of missing data. We evaluated the performance of a Bayesian Network inference model for estimating evapotranspiration in a large agricultural region in Brazil. To this end, the method considered factors such as accuracy, missing data, and model portability. The results indicate that the model achieves up to 86% accuracy when comparing estimated values to expected values derived from the Penman-Monteith equation. The results show that wind speed and relative humidity are the most critical climatological variables for accurate estimation.**

***Index Terms*—Evapotranspiration, Water Balance, Bayesian network, Bayesian Inference**

I. Introduction

According to the Food and Agriculture Organisation of the United Nations (FAO-UN), there are growing concerns about water availability and its usage in agriculture, which can have a serious impact on food security [1]. Moreover, the sustainable and conscious usage of water resources permeates directly or indirectly several of the Sustainable Development Goals (SDGs) proposed by the UN, including “SDG 2 - Zero Hunger” and “SDG 6 - Clean Water and Sanitation” [2].

One useful approach to assess the dynamic of water resources from a region is the water cycle quantification [3]. This method comprises several dynamic factors, including precipitation, irrigation, water percolation into the ground, evaporation from the soil and wet vegetation, and plant transpiration [4]. The last two factors are usually combined into a single metric, defined as evapotranspiration and is often considered to be the most relevant factor in the water cycle [5].

Given its dynamic relation with climatological factors, the evapotranspiration variable is difficult to measure directly [6]. According to the specialised literature [7]–[9], the empirical method considered a benchmark for observed evapotranspiration is the Penman-Monteith method. The main drawback of employing the Penman-Monteith method to estimate the evapotranspiration is that it relies on several climatological variables [10], such as net solar radiation, vapour water pressure, wind speed, relative humidity, and temperature.

Several additional empirical methods requiring fewer input data have been proposed. For example, the Hargreaves-Samani method [11] uses only temperature and extraterrestrial radiation, the latter of which is algebraically estimated; and the Benavides & Lopez method [12] uses temperature and relative humidity. Additionally, models have been proposed that take into account the climatological characteristics of a particular region, such as the Turc method [13], which was proposed for the western European climate, and the Linacre method [14], which was calibrated for the Australian climate.

Several studies estimate evapotranspiration using machine learning techniques. Recent work includes [15], in which the authors used Convolutional Neural Networks (CNNs) and ensembles to estimate reference evapotranspiration and obtained superior results in terms of variance and precision when compared to Seasonal AutoRegressive Integrated Moving Average (SARIMA) and Seasonal Naive approaches. While CNNs are frequently effective as predictive models, they do not provide an easily understandable model for humans [16].

Another recent example is presented in [17], where the authors employ a hybrid model composed of Gray Wolf Optimiser (GWO), Support Vector Machine (SVM), and data preprocessing methods to estimate evapotranspiration solely from this variable time series'. They obtained favourable results for the observed region despite the absence of direct climatological data. This method is only applicable if the preceding steps in the evapotranspiration time series are available.

A Bayesian Network is a probabilistic graphical model that encapsulates statistical relationships between variables in an observed system [18]. This model is applicable to a variety of disciplines, including project management [19], safety science [20], and traffic engineering [21]. Ecology is another field that has benefited from this machine learning approach, whether for inference over complex ecological variables [22] or for establishing relationships between native species and their habitat [23].

Due to the inherently probabilistic nature of Bayesian Network modelling, it deals with stochastic variables [24]. Additional characteristics include its robustness in the presence of missing data [18] and its structure provides an explainable model that produces a comprehensive output for stakeholders [25].

The present study uses a Bayesian Network model to estimate evapotranspiration from meteorological stations located in the area of five cities in a large agricultural region in Brazil. The stochastic nature of the climatological variables used in the Penman-Monteith method is incorporated into this model. We show that our model achieves an accuracy of between 80% and 86% in each area when all variables are present, compared to the expected results from the Penman-Monteith method, corroborating the proposed model's quality and portability. Additionally, we provide estimates with missing climatological data, showing that the model can still produce high-accuracy estimates without some variables required for the Penman-Monteith method. Our findings indicate that wind speed and minimum relative humidity are the most critical variables for the proposed model.

The database used in this article contains climate time series from various locations throughout Brazil's MATOPIBA region. This location is frequently mentioned as one of the leading agricultural producers in the country [26]. The National Institute of Meteorology (INMET) provides the original raw data, which is maintained by the Brazilian government in a public and open database.

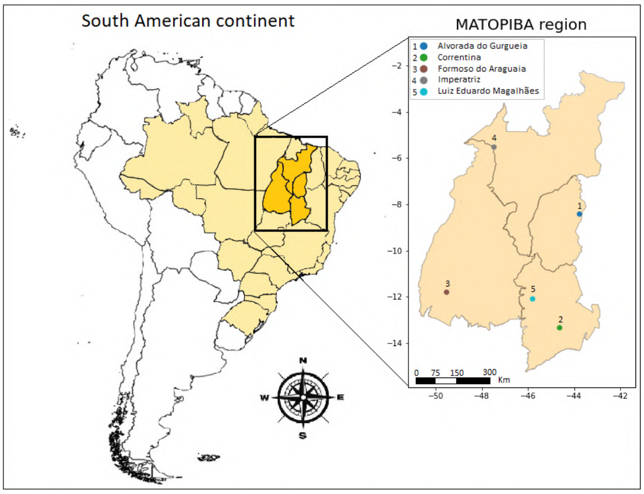


Fig. 1. Geographical location of the MATOPIBA region in Brazil. On the detail, the localisation for the cities whose climatological data was used for this study. Detail axis represent latitude and longitude (in degrees).

TABLE I
SELECTED CITIES FROM THE MATOPIBA REGION.

City	Coordinates	Station code (INMET)
Luís Eduardo Magalhães	12° 05' 31" S 45° 48' 18" W	A404
Correntina	13° 20' 34" S 44° 38' 13" W	A416
Imperatriz	5° 31' 33" S 47° 28' 33" W	A225
Formoso do Araguaia	11° 48' 09" S 49° 31' 42" W	A039
Alvorada do Gurgueia	8° 25' 26" S 43° 46' 37" W	A336

II. Material and Method

This Section contains the theoretical basis required to substantiate the results, discussion and conclusion presented in this paper.

A. *Brazilian MATOPIBA region*

MATOPIBA is a region in the north-eastern Brazilian region that includes parts of the states of Maranhão (MA), Tocantins (TO), Piauí (PI), and Bahia (BA) (Fig. 1). This region accounts for a sizeable portion of the soybean and livestock production in Brazil [27]. According to the Köppen climate classification, the majority of the MATOPIBA region has a Tropical Savanna (Aw) climate. The climate is semi-humid, with a dry winter (May to September) and rainy summer (October to April), with annual precipitation ranging between 1000 and 1600 mm [28], [29]. The cities of MATOPIBA whose data were used in this study are listed in Table I.

According to the Brazilian Institute of Geography and Statistics (IBGE), the municipality of Luís Eduardo Magalhães in Bahia has the state's seventh-largest economy [30], in addition to being a major producer of soybeans, cotton, and corn [31]. Correntina, also located in Bahia, is a region with significant agricultural potential but is still in its infancy [32]. Imperatriz, in Maranhão, is the state's second-largest economy [33]. Formoso do Araguaia, located in Tocantins, is the state's largest municipality and boasts the world's largest continuous area of irrigated rice, in addition to favourable climatic conditions for high yields [34]. Finally, the city of Alvorada do Gurgueia in Piauí is notable for its robust economy, agricultural potential, and abundant natural water resources [35]. These cities were chosen for their importance to the MATOPIBA region and will be used to evaluate the method described in this document.

B. Reference evapotranspiration

Evapotranspiration ET is defined as a sum of plant transpiration with the evaporation of water from the soil and wet vegetation. This phenomenon is nonlinear and complex [4], [5], consequently, robust nonlinear methods usually are used for modelling [5].

In irrigation planning and water distribution management, it is necessary to estimate ET. As the evapotranspiration dynamic is distinct from each crop species and its development stage as well as from each substrate, a reference evapotranspiration ET_0 metric is commonly employed [36]. The benchmark conditions for this modelling are a hypothetical reference crop with albedo of 0.23, on a soil with fixed surface resistance of 70 $s \cdot m^{-1}$ [36].

Among the various existing ET_0 methods, the application of the FAO-56 Penman-Monteith equation [37] is widely used and considered a standard norm [37]–[39]. This equation has two advantages over many other methods. The first advantage indicates that it is a predominantly physical approach and can be used globally without the imposition to estimate additional parameters. The second advantage shows that it is a well-documented method, implemented in various software and tested using a variety of lysimeters [37], [38], [40].

In agreement with the Penman-Monteith method, ET_0 can be estimated using Eq. 1 [36].

$$ET_0 = \frac{0.408\Delta(R_n - G) + \gamma\left(\frac{900}{T+273}\right)U_2(e_s - e_a)}{\Delta + \gamma(1 + 0.34U_2)} \tag{1}$$

From Eq. 1, ET_0 is estimated in $mm \cdot day^{-1}$ unit. The variables presented on the Eq. 1 are as follows:

- $\Delta\left[kPa \cdot {}^{\circ}C^{-1}\right]$ is the slope for the vapour pressure curve, obtained from the daily mean temperature and the saturation vapour pressure from that temperature;
- $R_n\left[MJ \cdot m^{-2} \cdot day^{-1}\right]$ is the net solar radiation, which is the balance of the net incoming shortwave radiation and the net outgoing longwave radiation. The estimate of this parameter takes into account the latitude and altitude for the observed site, as well as minimum, maximum, and mean temperature, the actual vapour pressure and the total incident solar radiation;
- $G\left[MJ \cdot m^{-2} \cdot day^{-1}\right]$ is the soil heat flux density, that may be considered null given a sufficient long time frame;
- $\gamma\left[kPa \cdot {}^{\circ}C^{-1}\right]$ is the psychrometric constant, a relation between the partial water pressure in air to the temperature;
- $T\left[{}^{\circ}C\right]$ is the mean daily temperature;
- $U_2\left[m \cdot s^{-1}\right]$ is the on-site wind speed, measured at 2 meters height;
- $e_s\left[kPa\right]$ and $e_a\left[kPa\right]$ are the saturation and actual vapour pressure, respectively.

Except for the variables T and U_2, all of the above factors were obtained through intermediate equations and approximations. Ref. [36] presents a detailed discussion on the intermediate steps.

For this paper, the climatological data used for the ET_0 estimate were minimum, maximum, and mean temperature, minimum and maximum relative humidity, wind speed, and total incident surface radiation.

C. Bayesian Networks

A Bayesian Network is a probabilistic graphical model that expresses the Joint Probability Distribution (JPD) across data using a direct acyclic graph. This model describes the probabilistic dependence of variables in an observable system [41]. The relationships presented in this model are derived entirely from an optimisation procedure and raw data [42]. In this case, edges from the proposed models represent probabilistic dependencies and not causal links.

Rigorously, a Bayesian Network B comprehends a structural element in the form of a directed acyclic graph (DAG) G and a set of Conditional Probability Tables (CPTs) Θ, referred to as the parameters of the structure [43].

Each vertex v_i from G represents a variable from the observed system, among all n variables. A directed edge from vertex v_j to vertex v_i ($i \neq j$) express an association relationship from the former to the latter. The set of all vertices that have an edge that ends at vertex v_i comprise the parents' set for v_i, denoted $Pa(v_i)$. Accordingly, each vertex that has v_i on its parents' set is called a child vertex for v_i.

For each observed variable v_i, there is a CPT $\theta_i \in \Theta$ that describes the probability of each possible state for v_i, conditioned to each combination of states from all variables on $Pa(v_i)$. In a short notation, the CPT for each v_i variable can be denoted as $p(v_i|Pa(v_i))$, for each state of v_i and each combination of states from the variables in $Pa(v_i)$.

As described in [44], the task of finding the Bayesian Network that best encapsulates the JPD $p(v_1, v_2, ..., v_n) = \prod_{i=1}^{n} p(v_i|Pa(v_i))$ exclusively from data can be approached as a DAG optimisation problem. One common technique to solve this combinatorial NP-hard problem is to employ a score-and-search algorithm, where a viable metric is assigned to each possible structure according to the provided data set, and the algorithm is required to change the graph topology, i.e. remove, add, and/or change edges direction to obtain the structure with the best score.

Examples of scoring functions to evaluate adherence from the DAG to data are the Bayesian Information Criterion (BIC) [45], the Minimum Description Length (MDL) [46], and the Bayesian Dirichlet equivalence with uniform prior metric (BDeu) [47]. A version of the classical Hill Climbing approach is commonly used to explore the search space of possible graphical structures [48].

Once a good candidate for the DAG G for the BN is obtained, the next step is to populate the CPT for each variable, according to the configuration of edges. As each θ_i encodes the probability for each state of v_i given the combination of values possible from $Pa(v_i)$, it is possible to perform inference tasks [49]. Given a subset of observed states from $Pa(v_i)$, the BN can be inspected for the most probable outcome for v_i. From a different approach, the BN can be examined for the most probable state from v_i that results simultaneously on each observed state from the children of v_i according to G.

D. Data availability

The data used for the experiments on this paper were obtained from the INMET, which maintains a public data base containing meteorological data from data collections stations scattered throughout the Brazilian territory.

The download for original data from INMET can be found at the address *https://tempo.inmet.gov.br/TabelaEstacoes/xxxx*, where *xxxx* refers to the code for the required station. The codes for stations used throughout this paper are presented as the third column in Table I.

III. Results and Discussion

In this Section, we present the result of obtaining a Bayesian Network for the climatological data set from Luís Eduardo Magalhães, as well as results from the inference of ET_0 from the graphical model. We also present discussions regarding the adherence of the model with climatological interactions, the relevance of variables from the data set, accuracy performance, and model portability.

A. Bayesian Network for climatological data

In order to study probabilistic associations between variables from the original data set, the following climatological components were extracted: mean, maximum, and minimum temperature, maximum and minimum relative humidity, total incident surface radiation, and wind speed, each one combined into daily values. Besides, the reference evapotranspiration from the Penman-Monteith method was obtained. Each entry on the final data set is comprised of the daily climatological data as cited and the evapotranspiration calculated.

Fig. 2 presents the graphical structure for the Bayesian Network (BN) which best represents the association relationships from the climatological variables and reference evapotranspiration obtained from Luís Eduardo Magalhães data set. This result was obtained from the *HillClimb* score-and-search algorithm and BDeu scoring method, both from the *pgmpy* Python package [50]. The Conditional Probability Tables were omitted due to the high number of states from each variable

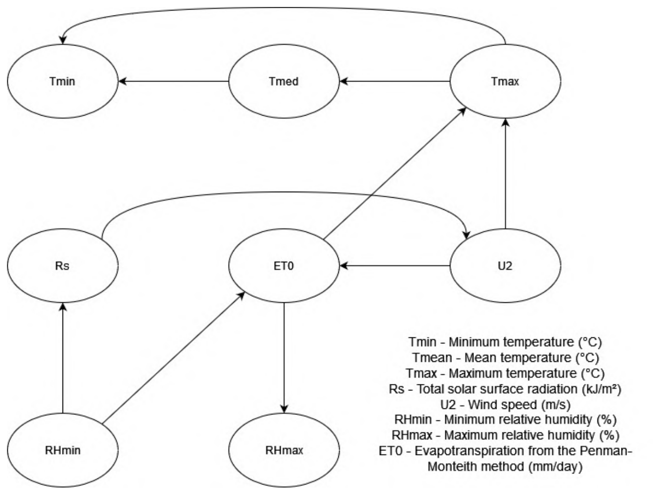


Fig. 2. Bayesian network obtained for the reference evapotranspiration (ET_0), considering the variables for the Penman-Monteith method applied on the data set from Luís Eduardo Magalhães. The climatological data represented by each node is described on the bottom right of the figure.

presented. The complete data set for this location contains 6466 entries, and the mean time to obtain such graphical structure was 1.01 seconds across 150 runs.

The associations presented by the edges on the BN represented in Fig. 2 are reasonable within the context of the calculations and climatological models. For instance, the BN managed to capture the relations between all three temperature variables ($Tmin$, $Tmax$, and $Tmean$) which were expected since all three nodes represent the same metric through different perspectives.

One climatological association identifiable from the model is represented between total solar surface radiation Rs and the wind speed $U2$. The uneven absorption of solar radiation at the Earth's surface may result in spatial differences in atmospheric pressure, leading to wind formation [51]. The associations between the evapotranspiration $ET0$, minimum and maximum relative humidity, respectively $RHmin$ and $RHmax$ variables were expected as $ET0$ estimates water loss to the atmosphere from soil and vegetation.

From the graphical structure for the BN in Fig. 2, the possible values for the $ET0$ are conditioned to each combination of values from the variables $RHmin$ and $U2$, since both compose the parents' set for this target variable. This result suggests that the most relevant climatological data for a good inference for the $ET0$ estimate are both relative humidity and wind speed.

Another characteristic that stands out from the obtained model is the lack of direct association between the total surface solar radiation variable and evapotranspiration. This suggests that for this particular model, the inferred value of $ET0$ is independent of Rs, conditioned to the input of $U2$.

TABLE II

INFERENCE ACCURACY (%) FOR THE $ET0$ VARIABLE FROM COMPLETE AND MISSING DATA, ACROSS SEVERAL CITIES FROM THE MATOPIBA REGION.

Removed variable(s)	Luís Eduardo Magalhães	Correntina	Imperatriz	Formoso do Araguaia	Alvorada do Gurgueia
–	82.5	86.3	85.5	83.1	80.2
$U2$	65.7	84.8	79.5	78.6	70.6
$RHmin$	79.2	80.4	80.8	79.1	73.8
$Tmax$	83.0	86.5	85.5	83.1	79.8
$RHmax$	80.7	84.2	87.2	78.0	76.3
All parents	64.3	79.3	74.2	75.6	66.7
All children	80.9	84.3	87.1	77.8	77.5
Both parents and children	53.0	67.8	73.7	59.9	54.8

B. *Inference from complete and incomplete data*

Once the graphical structure for the BN was obtained (Fig. 2), the next step was to evaluate the quality of the estimates for the $ET0$ variable, under two different scenarios: inference considering all the climatological data known, and the same operation with fewer variables available.

Table II presents the results for the inference tasks. A k-fold cross-validation was performed on each set of experiments, with the iterator provided by the *scikit-learn* Python package [52], and k value set to 5 across all experiments. A k-fold cross-validation approach splits the data set into k disjoint and evenly sized sets, to employ one partition as test set and the remainder k-1 partitions as training set. This test/training division is repeated k times until every partition is used once as test set. Finally, the CPTs were populated at each iteration considering the training set indexed by the iterator.

The second column on Table II presents the mean accuracy for the $ET0$ estimates, considering the climatological data set from Luís Eduardo Magalhães.

The second row shows the accuracy for the first set of experiments, where the expected $ET0$ value is compared to the most probable value from the inference over the probabilistic structure considering known the values for each climatological variable. Under this scenario with all data available for the Penman-Monteith method, the model achieved a mean accuracy of 82.5%, with the wrong estimates being slightly more often overestimates considering the expected value.

Subsequent rows from the second column present the mean accuracy for $ET0$ estimates after removing each combination of climatological variables required for the Penman-Monteith method from the inference query. The removed elements were the parents from the $ET0$ variable according to the graphical structure in Fig. 2, i. e. $U2$ and $RHmin$, as well as the children from the target variable, these being $Tmax$ and $RHmax$.

The results show that wind speed and minimum relative humidity are the most critical variables to an accurate estimate for the $ET0$ variable, with the lack of the former (third row) resulting in the lowest accuracy when removing single variables from the inference query. Accordingly, the removal of both parents results in a worse accuracy performance than removing both children (seventh and eighth row, respectively).

Finally, although the removal of each individual and both children variables from the inference query (fifth, sixth, and eighth rows, respectively) show no significant deterioration of mean accuracy performance compared to the query with all variables, the importance of these elements can be seen when comparing scenarios with both parents removed (seventh row) with that of parents and children removed (ninth row).

C. *Inference for different cities' data set*

The next sets of experiments deal with the accuracy performance for the inference of $ET0$ variable considering the graphical model obtained from Luís Eduardo Magalhães data applied to other cities in the MATOPIBA region. These experiments were carried to assess the portability of the graphical model across cities, albeit from the same climate profile.

The experiments for each city followed the same design as for Luís Eduardo Magalhães, with k-fold cross-validation and the CPTs learnt for each test set over the same graphical structure presented in Fig. 2. The data set for each city consisted of at least 4195 entries (Alvorada do Gurgueia) and at most 5058 (Imperatriz).

Columns three to six from Table II present the accuracy for estimates of $ET0$ variable, under the same data availability conditions from the previous experiment.

Considering the scenario where all the climatological variables required for the Penman-Monteith method were present, the accuracy for the estimates across all four cities was observed to be at least 80%. For all these locations the wrong estimates showed a bias towards underestimates, which points to a conservative inference from the model when applied to a different data set from which it was originally built.

From the experiments where a single variable was removed from the inference query, the inference results for all cities show that wind speed is the most critical variable, except for Correntina, which presents the lowest accuracy when minimum relative humidity is removed. This is a coherent result, given that both wind speed and minimum relative humidity are parent variables for $ET0$ on the graphical structure.

Across all data sets the inference performance was worse when removing both parents than removing both children. This result suggests a similar level of association between these five variables – $U2$, $RHmin$, $RHmax$, $Tmax$, and $ET0$ – across all data sets.

IV. Conclusion

In this article, we presented a Bayesian Network to evaluate relationships between climatological variables and evapotranspiration estimates, and to assess the performance of inference tasks on this probabilistic nature, under both scenarios of complete data and missing variables.

Our results have shown that the model achieves up to 86% accuracy for estimates of the $ET0$ variable, and that wind speed and minimum relative humidity are the most critical variables for a high accuracy estimate.

We also have shown a consistent performance of estimates when porting the model obtained for a city to other cities, albeit from the same climate profile. This result suggests that the model managed to capture the association between climatological variables, and these relationships are persistent at least within the same climate.

Future work on this subject include external validation for this model when applied to regions with more dissimilar climates, and a support framework to estimate the climatological variables needed for the Penman-Monteith method via a stochastic approach, such as Monte Carlo Markov Chain (MCMC) based models.

Acknowledgment

This work was partially supported by following agencies: grants #2014/50851-0, #2021/02464-0, São Paulo Research Foundation (FAPESP), grants #465755/2014-3, #301853/2018-5, #142390/2019-4, Brazilian National Council for Scientific and Technological Development (CNPq), and grant #2018/19150-6 BPE FAPESP.

References

[1] FAO, *The State of Food and Agriculture 2020*, ser. The State of Food and Agriculture. FAO, 2020.

[2] United Nations Department for Economic and Social Affairs, *Sustainable Development Goals Report 2020*. United Nations, 2020. [Online]. Available: https://unstats.un.org/sdgs/report/2020/The-Sustainable-Development-Goals-Report-2020.pdf

[3] T. G. Huntington, P. K. Weiskel, D. M. Wolock, and G. J. McCabe, "A new indicator framework for quantifying the intensity of the terrestrial water cycle," *Journal of Hydrology*, vol. 559, pp. 361–372, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022169418301276

[4] L. Wu, H. Zhou, X. Ma, J. Fan, and F. Zhang, "Daily reference evapotranspiration prediction based on hybridized extreme learning machine model with bio-inspired optimization algorithms: Application in contrasting climates of China," *Journal of Hydrology*, vol. 577, p. 123960, 07 2019.

[5] M. S. Zakeri, S.-F. Mousavi, S. Farzin, and H. Sanikhani, "Modeling of reference crop evapotranspiration in wet and dry climates using data-mining methods and empirical equations," *Journal of Soft Computing in Civil Engineering*, vol. 6, no. 1, pp. 1–28, 2022.

[6] I. Ghiat, H. R. Mackey, and T. Al-Ansari, "A review of evapotranspiration measurement models, techniques and methods for open and closed agricultural field applications," *Water*, vol. 13, no. 18, 2021. [Online]. Available: https://www.mdpi.com/2073-4441/13/18/2523

[7] M. Nouri and M. Homaee, "Reference crop evapotranspiration for data-sparse regions using reanalysis products," *Agricultural Water Management*, vol. 262, p. 107319, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378377421005965

[8] D. K. Vishwakarma, K. Pandey, A. Kaur, N. Kushwaha, R. Kumar, R. Ali, A. Elbeltagi, and A. Kuriqi, "Methods to estimate evapotranspiration in humid and subtropical climate conditions," *Agricultural Water Management*, vol. 261, p. 107378, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378377421006557

[9] A. Subedi and J. L. Chávez, "Crop evapotranspiration (et) estimation models: A review and discussion of the applicability and limitations of et methods," *Journal of Agricultural Science*, vol. 7, no. 6, p. p50, May 2015. [Online]. Available: http://www.ccsenet.org/journal/index.php/jas/article/view/45723

[10] G. D. Silva, S. H. Dias, L. Ferreira, J. Santos, and F. Cunha, "Performance of different methods for reference evapotranspiration estimation in Jaíba, Brazil," *Revista Brasileira de Engenharia Agrícola e Ambiental*, vol. 22, 02 2018.

[11] G. H. Hargreaves and Z. A. Samani, "Reference crop evapotranspiration from temperature," *Applied engineering in agriculture*, vol. 1, no. 2, pp. 96–99, 1985.

[12] J. G. Benavides and D. Lopez, "Formula para el calculo de la evapotranspiracion potencial adaptada al tropico (15º n - 15º s)," *Agronomia Tropical*, vol. 20, no. 5, pp. 335–345, 1970.

[13] L. Turc, "Estimation of irrigation water requirements, potential evapotranspiration: A simple climatic formula evolved up to date," *Annales Agronomiques*, vol. 12, no. 1, pp. 13–49, 1961.

[14] E. T. Linacre, "A simple formula for estimating evaporation rates in various climates, using temperature data alone," *Agricultural Meteorology*, vol. 18, no. 6, pp. 409–424, 1977. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0002157177900073

[15] P. de Oliveira e Lucas, M. A. Alves, P. C. de Lima e Silva, and F. G. Guimarães, "Reference evapotranspiration time series forecasting with ensemble of convolutional neural networks," *Computers and Electronics in Agriculture*, vol. 177, p. 105700, Oct 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S016816992031262X

[16] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, aug 2018. [Online]. Available: https://doi.org/10.1145/3236009

[17] T. Fu, X. Li, R. Jia, and L. Feng, "A novel integrated method based on a machine learning model for estimating evapotranspiration in dryland," *Journal of Hydrology*, vol. 603, p. 126881, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022169421009318

[18] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 01 2009.

[19] A. L. R. de Sousa, C. R. B. de Souza, and R. Q. Reis, "A 20-year mapping of bayesian belief networks in software project management," *IET Software*, vol. 16, no. 1, pp. 14–28, 2022. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/sfw2.12043

[20] M. Gödel, N. Bode, G. Köster, and H.-J. Bungartz, "Bayesian inference methods to calibrate crowd dynamics models for safety applications," *Safety Science*, vol. 147, p. 105586, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925753521004276

[21] Y. Wu and J. J. Yu, "A bayesian learning network for traffic speed forecasting with uncertainty quantification," in *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–7.

[22] E. Hui, R. Stafford, I. M. Matthews, and V. A. Smith, "Bayesian networks as a novel tool to enhance interpretability and predictive power of ecological models," *Ecological Informatics*, vol. 68, p. 101539, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574954121003307

[23] E. G. Mitchell, M. I. Wallace, V. A. Smith, A. A. Wiesenthal, and A. S. Brierley, "Bayesian network analysis reveals resilience of the jellyfish aurelia aurita to an irish sea regime shift," *Scientific Reports*, vol. 11, no. 1, p. 3707, Dec 2021. [Online]. Available: http://www.nature.com/articles/s41598-021-82825-w

[24] M. Bessani, J. A. Massignan, T. M. Santos, J. B. London, and C. D. Maciel, "Multiple households very short-term load forecasting using bayesian networks," *Electric Power Systems Research*, vol. 189, p. 106733, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378779620305368

[25] C. Moreira, Y.-L. Chou, M. Velmurugan, C. Ouyang, R. Sindhgatta, and P. Bruza, "Linda-bn: An interpretable probabilistic approach for demystifying black-box predictive models," *Decision Support Systems*, vol. 150, p. 113561, 2021, interpretable Data Science For Decision Making. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167923621000713

[26] R. L. A. Maranhão, O. A. de Carvalho Júnior, P. M. Hermuche, R. A. T. Gomes, C. M. McManus Pimentel, and R. F. Guimarães, "The spatiotemporal dynamics of soybean and cattle production in Brazil," *Sustainability*, vol. 11, no. 7, p. 2150, Apr 2019. [Online]. Available: https://www.mdpi.com/2071-1050/11/7/2150

[27] S. P. Polizel, R. M. da Silva Pinto Vieira, J. Pompeu, Y. da Cruz Ferreira, E. R. de Sousa-Neto, A. A. Barbosa, and J. P. H. B. Ometto, "Analysing the dynamics of land use in the context of current conservation policies and land tenure in the cerrado – Matopiba region (Brazil)," *Land Use Policy*, vol. 109, p. 105713, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0264837721004361

[28] L. Silva, D. Casaroli, A. Pêgo Evangelista, J. Júnior, and R. Battisti, "Rainfall intensity-duration-frequency relationships for risk analysis in the region of Matopiba, Brazil," *Revista Brasileira de Meteorologia*, vol. 34, pp. 247–254, 06 2019.

[29] P. Bispo, P. Rodriguez-Veiga, B. Zimbres, S. Miranda, C. Cezare, S. Fleming, F. Baldacchino, V. Louis, D. Rains, M. García, F. Espírito-Santo, I. Roitman, A. Pacheco-Pascagaza, Y. Gou, J. Roberts, K. Barrett, L. Ferreira, J. Zanin Shimbo, A. Alencar, and H. Balzter, "Woody aboveground biomass mapping of the brazilian savanna with a multi-sensor and machine learning approach," *Remote Sensing*, vol. 12, p. 2685, 08 2020.

[30] Brazilian Institute of Geography and Statistics (IBGE), "Luís Eduardo Magalhães: Panorama," https://cidades.ibge.gov.br/brasil/ba/luis-eduardo-magalhaes/panorama, 2019.

[31] Prefeitura, "Luís Eduardo Magalhães (BA)," https://luiseduardomagalhaes.ba.gov.br/a-cidade/, 2019/2020.

[32] Brazilian Institute of Geography and Statistics (IBGE), "Correntina: Panorama," https://cidades.ibge.gov.br/brasil/ba/correntina/panorama, 2019.

[33] ——, "Imperatriz: Panorama," https://cidades.ibge.gov.br/brasil/ma/imperatriz/panorama, 2019.

[34] ——, "Formoso do Araguaia: Panorama," https://cidades.ibge.gov.br/brasil/to/formoso-do-araguaia/panorama, 2019.

[35] ——, "Alvorada do Gurguéia: Panorama," https://cidades.ibge.gov.br/brasil/pi/alvorada-do-gurgueia/panorama, 2019.

[36] R. Allen, L. Pereira, D. Raes, and M. Smith, *Crop evapotranspiration: guidelines for computing crop water requirements*, ser. FAO irrigation and drainage paper 56. Food and Agriculture Organization of the United Nations, 1998.

[37] P. Droogers and R. Allen, "Estimating reference evapotranspiration under inaccurate data conditions," *Irrigation and Drainage Systems*, vol. 16, pp. 33–45, 02 2002.

[38] X. Hou, J. Fan, F. Zhang, W. Hu, F. Yan, C. Xiao, Y. Li, and H. Cheng, "Determining water use and crop coefficients of drip-irrigated cotton in south Xinjiang of China under various irrigation amounts," *Industrial Crops and Products*, vol. 176, p. 114376, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0926669021011419

[39] I. Walter, R. Allen, R. Elliott, M. Jensen, D. Itenfisu, T. Howell, R. Snyder, P. Brown, and S. Echings, "The asce standardized reference evapotranspiration equation," *Final Report of the Task Committee on Standardization of Reference Evapotranspiration*, 05 2001.

[40] C.-Y. Xu and V. Singh, "Cross comparison of empirical equations for calculating potential evapotranspiration with data from Switzerland," *Water Resources Management*, vol. 16, pp. 197–219, 06 2002.

[41] R. Neapolitan, *Learning Bayesian Networks*. Pearson, 01 2003.

[42] L. A. C. Jr., "Modernizing the bradford hill criteria for assessing causal relationships in observational data," *Critical Reviews in Toxicology*, vol. 48, no. 8, pp. 682–712, 2018. [Online]. Available: https://doi.org/10.1080/10408444.2018.1518404

[43] T. J. Gross, M. Bessani, W. Darwin Junior, R. B. Araújo, F. A. C. Vale, and C. D. Maciel, "An analytical threshold for combining Bayesian Networks," *Knowledge-Based Systems*, 3 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0950705119301352

[44] B. M. Broom, K.-A. Do, and D. Subramanian, "Model averaging strategies for structure learning in bayesian networks with limited data," *BMC Bioinformatics*, vol. 13, no. S13, p. S10, Aug 2012.

[45] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978. [Online]. Available: http://www.jstor.org/stable/2958889

[46] J. Suzuki, *A Construction of Bayesian Networks from Databases Based on an MDL Principle*. Elsevier, 1993, p. 266–273. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/B9781483214511500378

[47] W. Buntine, "Theory refinement on bayesian networks," in *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*, ser. UAI'91. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1991, p. 52–60.

[48] M. Scutari and R. Nagarajan, "Identifying significant edges in graphical models of molecular networks," *Artificial Intelligence in Medicine*, vol. 57, no. 3, p. 207–217, Mar 2013. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0933365712001546

[49] A. Alexiou, V. D. Mantzavinos, N. H. Greig, and M. A. Kamal, "A bayesian model for the prediction and early diagnosis of alzheimer's disease," *Frontiers in Aging Neuroscience*, vol. 9, Mar 2017. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fnagi.2017.00077/full

[50] A. Ankan and A. Panda, "pgmpy: Probabilistic graphical models using python," in *Proceedings of the 14th Python in Science Conference (SCIPY 2015)*. Citeseer, 2015.

[51] M. Pidwirny, *Forces Acting to Create Wind. Fundamentals of Physical Geography*, 2nd ed., 2006. [Online]. Available: https:http://www.physicalgeography.net/fundamentals/7n.html

[52] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

ANNEX C – *A STOCHASTIC BAYESIAN ARTIFICIAL INTELLIGENCE FRAMEWORK TO ASSESS CLIMATOLOGICAL WATER BALANCE UNDER MISSING VARIABLES FOR EVAPOTRANSPIRATION ESTIMATES*

This Annex contains the complete paper published in the Agronomy journal in Nov/23.

agronomy

Article

A Stochastic Bayesian Artificial Intelligence Framework to Assess Climatological Water Balance under Missing Variables for Evapotranspiration Estimates

Vitor P. Ribeiro [1,*], Luiz Desuó Neto [2], Patricia A. A. Marques [3], Jorge A. Achcar [4], Adriano M. Junqueira [1], Adilson W. Chinatto, Jr. [5], Cynthia C. M. Junqueira [5], Carlos D. Maciel [2] and José Antônio P. Balestieri [1]

1 School of Engineering and Sciences, São Paulo State University (UNESP), Guaratinguetá 12516-410, SP, Brazil; adriano.junqueira@unesp.br (A.M.J.); jose.perrella@unesp.br (J.A.P.B.)
2 Department of Electrical and Computer Engineering, University of São Paulo (USP), São Carlos 13566-590, SP, Brazil; luiz.desuo.neto@usp.br (L.D.N.); carlos.maciel@usp.br (C.D.M.)
3 Department of Biosystems Engineering, University of São Paulo (USP), Piracicaba 13418-900, SP, Brazil; paamarques@usp.br
4 Medical School, University of São Paulo (USP), Ribeirão Preto 14049-900, SP, Brazil; achcar@fmrp.usp.br
5 Espectro Ltd., Campinas 13084-012, SP, Brazil; chinatto@espectro-eng.com.br (A.W.C.J.); cyjunqueira@uol.com.br (C.C.M.J.)
* Correspondence: vitor.p.ribeiro@unesp.br

Abstract: The sustainable use of water resources is of utmost importance given climatological changes and water scarcity, alongside the many socioeconomic factors that rely on clean water availability, such as food security. In this context, developing tools to minimize water waste in irrigation is paramount for sustainable food production. The evapotranspiration estimate is a tool to evaluate the water volume required to achieve optimal crop yield with the least amount of water waste. The Penman-Monteith equation is the gold standard for this task, despite it becoming inapplicable if any of its required climatological variables are missing. In this paper, we present a stochastic Bayesian framework to model the non-linear and non-stationary time series for the evapotranspiration estimate via Bayesian regression. We also leverage Bayesian networks and Bayesian inference to provide estimates for missing climatological data. Our obtained Bayesian regression equation achieves 0.087 mm $\cdot$ day^{-1} for the RMSE metric, compared to the expected time series, with wind speed and net incident solar radiation as the main components. Lastly, we show that the evapotranspiration time series, with missing climatological data inferred by the Bayesian network, achieves an RMSE metric ranging from 0.074 to 0.286 mm $\cdot$ day^{-1}.

Keywords: irrigation planning; decision support system; artificial intelligence; stochastic modeling; Bayesian inference

check for updates

Citation: Ribeiro, V.P.; Desuó Neto, L.; Marques, P.A.A.; Achcar, J.A.; Junqueira, A.M.; Chinatto, A.W., Jr.; Junqueira, C.C.M.; Maciel, C.D.; Balestieri, J.A.P. A Stochastic Bayesian Artificial Intelligence Framework to Assess Climatological Water Balance under Missing Variables for Evapotranspiration Estimates. *Agronomy* **2023**, *13*, 2970. https://doi.org/10.3390/agronomy13122970

Academic Editors: Mo Li and Fan Zhang

Received: 28 September 2023
Revised: 30 October 2023
Accepted: 2 November 2023
Published: 30 November 2023

1. Introduction

Water maintains life. From a broad perspective, water resources are at the core of several vital elements such as ecology [1], agriculture [2], transportation [3], electricity generation [4], and thirst quenching [5]. Water resources also impact many sociopolitical and economic areas and are essential for human survival, alongside other natural and human resources. Given the importance of these factors for the development of humanity, the United Nations (UN) proposed sustainable development goals (SDGs). These goals consist of seventeen components with their underlying objectives proposed as guidelines for the rational and sustainable development of all UN member states [6]. In Brazil, many socioeconomic factors are directly tied to the use of water resources. For example, hydroelectric plants generated the majority of the country's electricity generation capacity of 181.6 GW for 2022, with a share of 60.2% [7]. Brazil is also a key contributor to global food security, ranking in 2021 as the world's largest producer of coffee beans, oranges, and soybeans; the second-largest in cattle meat; and the third in maize and chicken meat, among several other agricultural and livestock products [8]. The intricate interactions between soil, plants, and atmospheric systems play a pivotal role in food production [9]. Therefore, techniques that aid in producing nutritious food based on limited components of soil and water resources are paramount due to climate change and the scarcity of water resources [10].

The productivity of crops is one of the predominant factors in the maintenance of food security. The total yield of a crop is highly correlated with factors such as nutrient satisfaction, the application of pesticides, sunshine, and water availability [11]. Wheat is one of the three main cereals that underpin global security—alongside maize and rice—and serves as an example to highlight the amount of water required to produce a significant quantity of such an essential commodity [12]. In 2021, mainland China was the world's largest wheat producer, with an estimated harvested area of 23.5 million hectares, according to the UN's Food and Agriculture Organization (FAO-UN) [8]. Based on a rough approximation where the estimated area is composed solely of winter wheat, which requires approximately 120 mm (1200 $m^3 \cdot ha^{-1}$) for optimal production under regular climate conditions [13], the water required solely for this crop during its production cycle amounts to about 2.828×10^{10} mm as a volume measurement over a standardized area of 1 m^2. This water volume for the consumptive use of one crop in one country underscores the importance of policies that promote the sustainable and rational use of water resources

One method to minimize the waste of water in irrigation is to perform the climatological water balance for the cultivated area. This task consists of monitoring climatological events—mainly evapotranspiration and precipitation—to estimate the availability of water resources to the crops in the area [14]. The Penman–Monteith equation is the gold standard for estimating a reference evapotranspiration metric under standardized conditions [15]. The method demands the acquisition of climatological variables—net solar radiation, wind speed, temperature, and relative humidity—whose measurements may not be available for a given area [16]. Several other empirical equations are present in the literature, most as simplifications from the gold standard for specific regions or to deal with the lack of a given climatological variable [17]. Many authors have presented different approaches to tackle the issue of evapotranspiration estimates from data leveraging computational power and algorithmic resources [18–22]. Moreover, the method proposed in [23] models and discusses evapotranspiration on a global scale unaided by soil and vegetation data, and the review in [24] covers many evaporation and evapotranspiration estimate methods from the standpoint of several applications, their requirements, and characteristics.

Reliably assessing the parameters of the climatological water balance is paramount for achieving sustainable water usage and providing food security for the population [25]. Despite the multitude of techniques and approaches to estimate these parameters—mainly evapotranspiration—most methods presented in the literature lack flexibility when a required climatological factor is unavailable, becoming inapplicable [18]. Another characteristic of many methods that provide forecasts based on climatological or agrometeorological data is their inability to cope with the nonlinearity and non-stationarity inherent to these elements [26,27]. As such, the specialized literature reveals a knowledge gap for stochastic methods with a degree of flexibility that allows them to perform even with incomplete information.

A method class that resonates with the purported characteristics is the stochastic Bayesian regression analysis, which estimates the probability distribution functions that best fits the data [28]. This process has the innate ability to assimilate the non-deterministic and non-stationary characteristics in the data and provide a robust tool for modeling real-world data in several fields, such as medicines [29] and power planning [30]. Another method based on Bayesian statistics is the Bayesian network (BN), which is an approach based on probabilistic graph models encompassing association relationships between stochastic variables, which allows for the query of the most probable outcome of a given variable based on a complete or incomplete observation of the other variables in the model [31]. Examples of fields of study that benefit from BN modeling include environmental science [32], supply chain analysis [33], and engineering reliability [34]. The main drawback of this method relies on finding the directed acyclic graph that presents the best adherence to the data, which is a super-exponential problem based on the number of observed variables and belongs to the NP-hard class of problems [35].

The contributions of this paper are three-fold: first, we present and discuss a fully stochastic computational framework to determine the climatological variables that are most relevant to estimate daily evapotranspiration exclusively from data based on semi-parametric modeling, Bayesian statistics, and MCMC methods. Second, we establish a Bayesian logistic regression equation to incorporate climatological data characteristics of nonlinearity and non-stationarity. Thirdly, we present a BN structure to support the inference task of missing variables, which provides flexibility to the proposed model and allows for its usage under different scenarios of data availability. We also show the performance of the proposed BN to infer any of the missing relevant climatological factors based on the evidence as samples of the other variables and achieve reasonable error metrics for evaporation estimate from the logistic regression with inferred data against the expected value. Lastly, we compare the behaviors of measured precipitation, estimated evapotranspiration—both from the Penman–Monteith equation and the proposed method—climatological water balance, and in situ tensiometer data to verify the adherence of the theoretical models to real-world applications.

The remainder of this paper is structured as follows: Section 2 provides the theoretical background relevant to the conducted experiments and discussions for both the climatological models and fundamentals of Bayesian modeling. Section 3 presents the employed data, evaluated models, and descriptions of the main algorithms. Section 4 presents the results and discussions pertinent to the proposed experiments. Lastly, Section 5 presents the conclusions of this paper, alongside remarks about the advantages and drawbacks of the proposed methodology and a few identified gaps for future works.

2. Theory

This section provides a theoretical background of the topics discussed throughout this work. Sections 2.1 and 2.2 present a theory for the climatological and agrometeorological components of the discussion, while Sections 2.3 and 2.4 focus on the Bayesian inference and BNs.

2.1. Reference Evapotranspiration—Penman–Monteith (FAO-56) Method

Estimating the amount of water evaporated and transpired by the culture and evaporated from the soil is vital to determine the irrigation volume required for optimal production. However, it is difficult—or in some cases, practically impossible—to measure directly due to various climatological factors.

In light of the significance of evapotranspiration, the literature offers diverse analytical methods [36–39] to estimate it from climatological variables based on various environmental conditions [17]. These traditional methods utilize regression analysis to model and extrapolate the evapotranspiration metric conditioned on the availability of specific climatological variables, such as the Hargreaves–Samani equation, which relies solely on mean temperature and extraterrestrial solar radiation [36]. Another instance is the Linacre method, calibrated according to the Australian climate [37].

The Food and Agriculture Organization [15] has designated the Penman–Monteith equation as the gold standard for estimating standardized evapotranspiration for a grass crop (vegetation height = 0.12 m, albedo = 23%) on soil with a fixed surface resistance of 70 $\text{S} \cdot \text{m}^{-1}$ without water restrictions. The original Penman–Monteith equation is presented here as Equation (1), which estimates the ET_0 reference evapotranspiration in millimeters per day.

$$ET_0 = \frac{0.408\Delta(R_n - G) + \gamma\left(\dfrac{900}{T_{mean} + 273}\right)U_2(e_s - e_a)}{\Delta + \gamma(1 + 0.34U_2)} \tag{1}$$

The Penman–Monteith equation necessitates several climatological measurements, including net solar radiation (R_n), mean daily temperature (T_{mean}), on-site wind speed at 2 m (U_2), and actual (e_a) and saturation (e_s) vapor pressure. The slope of the vapor pressure curve (Δ), the soil heat flux density (G), and the psychrometric constant (γ) are additional parameters.

Despite its high correlation with the observed value for the evapotranspiration metric, the primary limitation of this equation is its dependence on several climatological factors. Weather stations that can automatically provide the required data for the Penman–Monteith equation are unattainable to most rural producers, resulting in this method being unfeasible to estimate [40].

One analytical approach employed throughout the literature to estimate daily evapotranspiration based on fewer climatological variables is the Hargreaves–Samani method [36], presented here as Equation (2). This method requires the mean, maximum, and minimum daily temperatures—respectively T_{mean}, T_{max}, and T_{min}—as well as an estimate for the extraterrestrial incident solar radiation (R_a). Another common analytical approach is the Benavides and Lopez method [39], presented as Equation (3), whose equation only requires data on the mean daily temperature (T_{mean}) and mean daily relative humidity (RH_{mean}).

$$ET_{0Hg} = R_a \times 0.408 \times 0.0023(T_{mean} + 17.8)(T_{max} - T_{min})^{0.5} \tag{2}$$

$$ET_{0BL} = 1.21 \times 10\left(\frac{7.42T_{mean}}{234.7 + T_{mean}}\right)(1 - 0.01RH_{mean}) + 0.21T_{mean} - 2.30 \tag{3}$$

The variety of phytomorphology properties for crops used for plantations implies a high variance for the water amount evaporated and transpired by the plants. For this reason, and to provide a generic framework to estimate evapotranspiration under scenarios of different cultures and development stages, the analytical equations—including the Penman–Monteith approach—provide a standardized reference evapotranspiration ($ET_0 \triangleq ET$). To estimate the evapotranspiration value for a given culture and development stage, one can use the culture coefficient (k_c) to obtain the crop evapotranspiration (ET_c) according to Equation (4).

$$ET_c = k_c \cdot ET_0 \tag{4}$$

Throughout this text, the word "evapotranspiration" refers to the reference evapotranspiration $ET_0 \triangleq ET$ unless stated otherwise.

2.2. Climatological Water Balance and Water Requirements

The water balance for a given region obeys the mass conservation principle summarized by Equation (5), which describes the equilibrium of water inflows and outflows from a control volume [41].

$$P = ET + Q + \Delta S \tag{5}$$

In Equation (5), the water inflow from precipitation (P) balances the outflows from evapotranspiration (ET) and surface and subsurface runoff (Q), promoting a variation in soil moisture (ΔS).

When considering a plantation scenario of groundwater flow equilibrium, a linear or exponential function describes the surface runoff as a function of precipitation [42]. As such, Equation (6) describes the soil moisture dynamics as a function of evapotranspiration and precipitation.

$$P = ET + f(P) + \Delta S \tag{6}$$

Assessing soil water content plays a fundamental role in establishing the water available to crops, which impacts plantation production. The water amount required for optimal production depends on the culture, its development stage, and efforts to minimize water waste.

A simple approach to estimating the water content in a given region relies on climatological data for precipitation and evapotranspiration, the so-called climatological water balance, to estimate the available water deficit (*awd*), the percentage of water stored in soil (*st*), and the net water balance [43]. Algorithm 1 presents the heuristics to assess the *awd* and *st* metrics solely from the climatological factors.

Algorithm 1 Algorithm to estimate the accumulated water deficit *curr_awd* and water stored *curr_st* based on evapotranspiration ET, precipitation P, and the previous values of these estimates.

Require: ET, P, $prev\_awd$, $prev\_st$
if $ET > P$ **then**
  $curr\_awd \leftarrow prev\_awd + (P - ET)$
  $curr\_st \leftarrow 100 \cdot e^{curr\_awd/100}$
else
  $curr\_st \leftarrow prev\_st + (P - ET)$
  if $curr\_st > 100$ **then**
    $curr\_st \leftarrow 100$
  end if
  $curr\_awd \leftarrow 100 \cdot e^{curr\_st/100}$
end if
  return $curr\_awd$, $curr\_st$

Given all these nuances, several methods estimate soil water content based on readings of low-cost sampling equipment, such as tensiometers and capacitors, to provide decision-support for the irrigation responsible [44]. One of these methods is the van Genuchten equation, which estimates soil water content for a specific soil (θ, in percentage) based on soil moisture content estimated by suction pressure (ψ, in kPa) and experimental coefficients [45]. Equation (7) provides the generic expression for the van Genuchten equation.

$$vG(\psi) = \theta_r + \frac{\theta_s - \theta_r}{\left(1 + (\alpha|\psi|)^n\right)^m} \tag{7}$$

where θ_s is the saturated soil water content or field capacity, and θ_r is the residual soil water content or wilting point under the physical limitations of the tensiometers. The parameters $\alpha\ \left[\text{cm}^{-1}\right]$, n, and $m = 1 - 1/n$ are constants obtained experimentally based on the soil's characteristic water retention curve.

2.3. Bayesian Regression and Modeling

Regression analysis is a tool used to estimate the influence of a set of input variables over an output (response) variable [46]. Let y be the target variable and $X = (X_1, X_2, \ldots, X_n)$ be the set of components (or covariates) that may contribute to the values of the outcome. Equation (8) presents a multiple linear regression of y as a linear combination of elements X_i plus a constant value as the series' DC component. The main task of this regression is to estimate $A = \{a_0, a_1, \ldots, a_n\}$ as the set of regression parameters from the dataset.

$$y \approx a_0 + \sum_{i=1}^{n} a_i X_i + \varepsilon \tag{8}$$

A random variable usually assumed with a Gaussian $\mathcal{N}(0, \sigma^2)$ distribution models the regression error term ε (non-observed) in Equation (8). Under the frequentist approach, the estimators of the regression parameters $a_0, a_1, \ldots, a_n$ are usually obtained using the minimum squares method, i.e., the set of coefficients that minimizes the square sum of the errors [47]. The variance parameter σ^2 should also be estimated from the data.

Under a Bayesian approach (Bayesian regression model), we assume the same regression model from Equation (8), while considering Gaussian errors, with the regression parameters β_i representing random quantities with specified probability density functions—also denoted as prior distributions—with hyperparameters φ_i. These hyperparameters may be univariate or multivariate, whose values are usually known, as represented in Equation (9). If the hyperparameters are unknown, which comes from a hierarchical Bayesian approach, each hyperparameter also requires—over itself—a prior distribution.

$$y \approx \beta_0(\varphi_0) + \sum_{i=1}^{n} (\beta_i(\varphi_i) \cdot X_i) + \varepsilon \tag{9}$$

Bayes' theorem provides the means to obtain the posterior distributions for the regression parameters β_i [48]. Let D represent the dataset, $\pi(B)$ represent the joint prior distribution for the vector of regression parameters, $B = (\beta_0, \beta_1, \ldots, \beta_n)$, and $L(B \mid D)$ represent the likelihood function for B based on the data D. Under the assumption of Gaussian distribution for the error ε in Equation (9), the joint posterior distribution for B, denoted as $\pi(B \mid D)$, is given by the combination of the prior distribution $\pi(B)$ with the likelihood function $L(B \mid D)$, as described in Equation (10). The integral in the denominator represents multiple integrals with dimension *dim(B)*.

$$\pi(B \mid D) = \frac{\pi(B)L(B \mid D)}{\int_B \pi(B)L(B \mid D)dB} \tag{10}$$

The integral in Equation (10)'s denominator represents multiple integrals with dimension *dim(B)*, which presents the main drawback of the Bayesian regression analysis. An analytical approach to estimate the term $\int_B \pi(B)L(B \mid D)dB$ usually becomes unfeasible.

Popular approaches extensively used in applied Bayesian analysis are Markov Chain Monte Carlo (MCMC) method, such as Gibbs sampling or Metropolis–Hastings algorithms, to simulate samples for the joint posterior distribution of interest [48]. The MCMC method generates values from each conditional posterior distribution $\pi(\beta_i \mid \beta_{(i)}, D)$, where $\beta_{(i)}$ denotes the vector of all parameters, except β_i, and is available in existing Bayesian software.

From the simulated samples of the joint posterior distribution of interest, we obtain the posterior summaries of interest, such as Monte Carlo point Bayesian estimates for the parameters of the model (usually the posterior means) or 95% credible intervals for the parameters of interest.

The authors acknowledge the intricacies and hardships of the MCMC methods, such as chain convergence, the sampler chosen, and initial values for the parameters in the iterative method. These characteristics, albeit relevant, lie outside the scope of the present work and were omitted. Please refer to reference [49] for more details on MCMC methods.

2.4. Bayesian Networks

Let D represent the dataset sampled from a system, which comprises n random variables, i.e., variables whose values follow a probability distribution function. A BN is a probabilistic graph model representing the stochastic association relationships between all variables in a given system [31].

Let Θ be a characterization of the joint probability function, according to a product measure for a system's n random variables $X_i, i = \{1, 2, \ldots, n\}$ based on arbitrary ordering. Equation (11) provides a generalization for such a joint distribution.

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i \mid X_j = x_j, \forall j < i) \qquad j \in \mathbb{N}^* \tag{11}$$

It is possible to factor out the independent terms for each X_i variable in Equation (11), which relies on conditional probability distributions. Let $pa(X_j)$ represent the set of random variables whose pairwise joint probability function satisfies the relation in Equation (12), given an ordering for the system's variables. The set $pa(X_i)$ defines the set of *parent* variables for X_i.

$$X_i \in pa(X_j) \iff P(X_i = x_i, X_j = x_j) < P(X_i = x_i)P(X_j = x_j), \quad i > j; \quad i, j \in \{1, 2, \ldots, n\} \tag{12}$$

Equation (13) presents a simplified version of Equation (11), based on the variables X_i conditioned to their parents' set $pa(X_i)$.

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i \mid pa(X_i)) \tag{13}$$

A BN Θ is a probabilistic graphic model that represents the joint probability function in Equation (13), based on a directed acyclic graph (DAG) G [35]. The BN's graphical structure contains one vertex for each random variable X_i, and its edges correspond to the perceived mathematical dependency of a variable on its parents. Based on this definition, each node X_i has one directed path, starting from each $X_j \in pa(X_i)$ and ending on itself.

A relevant comment regarding the definition of the parent sets from Equation (12) that reflects on the direction of the edges in G is that it establishes an *ordering* for the random variables. Based on the symmetrical nature of the right-hand side of Equation (12)—$P(X_i)P(X_j) = P(X_j)P(X_i)$—whether X_i belongs to $pa(X_j)$ or X_j belongs to $pa(X_i)$ is entirely dependent on the ordering of the variables. Indeed, the edges in the graph of a BN do not imply a causal relationship between random variables but rather association relationships based on the variables' ordering [50].

Each vertex on the BN has an underlying representation of the probability distribution σ_i for its values, conditioned to each realization for the elements on its parent's set. This mechanism—usually a set of conditional probability tables (Σ) when all observed variables are discrete—allows for inference queries for the most probable outcome of any given variable, given a complete or partial observation of all the elements on this component's parent set. As such, both the DAG G and the collection of CPTs over the vertices Σ are factors that fully describe a BN $\Theta = (G, \Sigma)$ [31].

Equation (14) presents a generalization for Equation (10) based on Bayes' theorem, and the most right-handed side component is valid if the dataset contains only independent and identically distributed (i.i.d.) samples.

$$P(\Theta \mid D) = \frac{P(D \mid \Theta)P(\Theta)}{P(D)} \propto P(D \mid \Theta)P(\Theta) \tag{14}$$

The model that has the best adherence to the underlying system is the one that provides the best posterior probability based on the provided dataset under the (usually naive [51]) assumption that the provided data fully represent the system's behavior. As such, Equation (15) provides a mathematical expression based on the right-hand side Equation (14) to find the best model for the underlying system based on data.

$$\max P(\Theta \mid D) \propto \underset{\Theta}{\operatorname{argmax}}\, P(D \mid \Theta)P(\Theta) \tag{15}$$

In other words, Equation (15) stipulates that the model parametrized by Θ that provides the best likelihood of data given its parameters $P(D \mid \Theta)$ is the one that provides the best predictive posterior distribution $P(\Theta \mid D)$ for the system's components.

Finding the best BN parametrized by $\Theta = (G, \Sigma)$ that maximizes the likelihood of the data distribution according to the model, satisfying Equation (15), entails a graph optimization problem [52]. This characteristic is due to the structural component of a BN being a DAG, and the representation of the joint probability for each node relies on its parent set.

There are three main approaches to obtaining a BN to maximize the data likelihood of the model: score-and-search, constraint-based, and hybrid methods. Score-and-search methods employ a measure to score the model's adherence to the data and search on the graph space for a structure that improves this metric. Constraint-based methods rely on conditional independence tests over the data variables to establish the edges on G, similar to those presented in Equation (12). Lastly, hybrid methods start from a constraint-based approach to scale down the graph search space and, afterward, a score-and-search method to scour the reduced space.

The significant amount of time and computational power required to derive a Bayesian Network (BN) from data are among the main disadvantages of this method. Despite these substantial requirements, the inference queries on the BN structure rely on basic matrix operations considering the CPT nodes.

3. Material and Method

3.1. Time Series for Climatological Data

The data employed throughout this paper originate from the Brazilian National Institute of Meteorology (INMET—Instituto Brasileiro de Meteorologia). The dataset consists of time series of seven climatological variables, i.e., temperature, atmospheric pressure, relative humidity, dew point temperature, wind speed, incident solar radiation, and precipitation.

The particular monitored location is the city of Patrocínio in the state of Minas Gerais (18°59′48″ S, 46°59′09″ W), at 978.11 m above sea level, and a warm temperate with a hot summer climate profile (Cwa), according to the Köppen classification. Figure 1 presents the geographical location of the state of Minas Gerais (MG) within the Brazilian map and the Patrocínio municipality within the former, alongside four other municipalities whose climatological time series that the authors employed as validation sets. Patrocínio is Brazil's largest coffee producer and the fourth-largest producer of bovine milk, among other goods. The initial dataset consists of 5473 daily samples for each climatological variable from Patrocínio/MG. The data span from 22 August 2008 to 31 January 2022, with a 97% rate of valid elements.

Figure 2a–g present graphical representations of the time series, with temperature, atmospheric pressure, relative humidity, and dew point temperatures sub-divided into maximum, mean, and minimum values for each day. Figure 2h presents the estimated evapotranspiration according to the Penman–Monteith equation over the available climatological data.

An additional dataset with the same climatological variables from Patrocínio/MG acts as a test set to evaluate the proposed method. This extra set spans from 1 February to 31 December 2022, with a total of 334 samples. Furthermore, four supplementary datasets, each containing 334 samples from the same period previously described, from four different cities across Brazil, serve to evaluate the model when considering distinct regions and climate profiles. Table 1 presents the names of cities, their geographic coordinates, the climate profiles according to the Köppen classification, and their distances from Patrocínio/MG. The only criterion for choosing these cities is to provide good coverage across Brazil's territory.

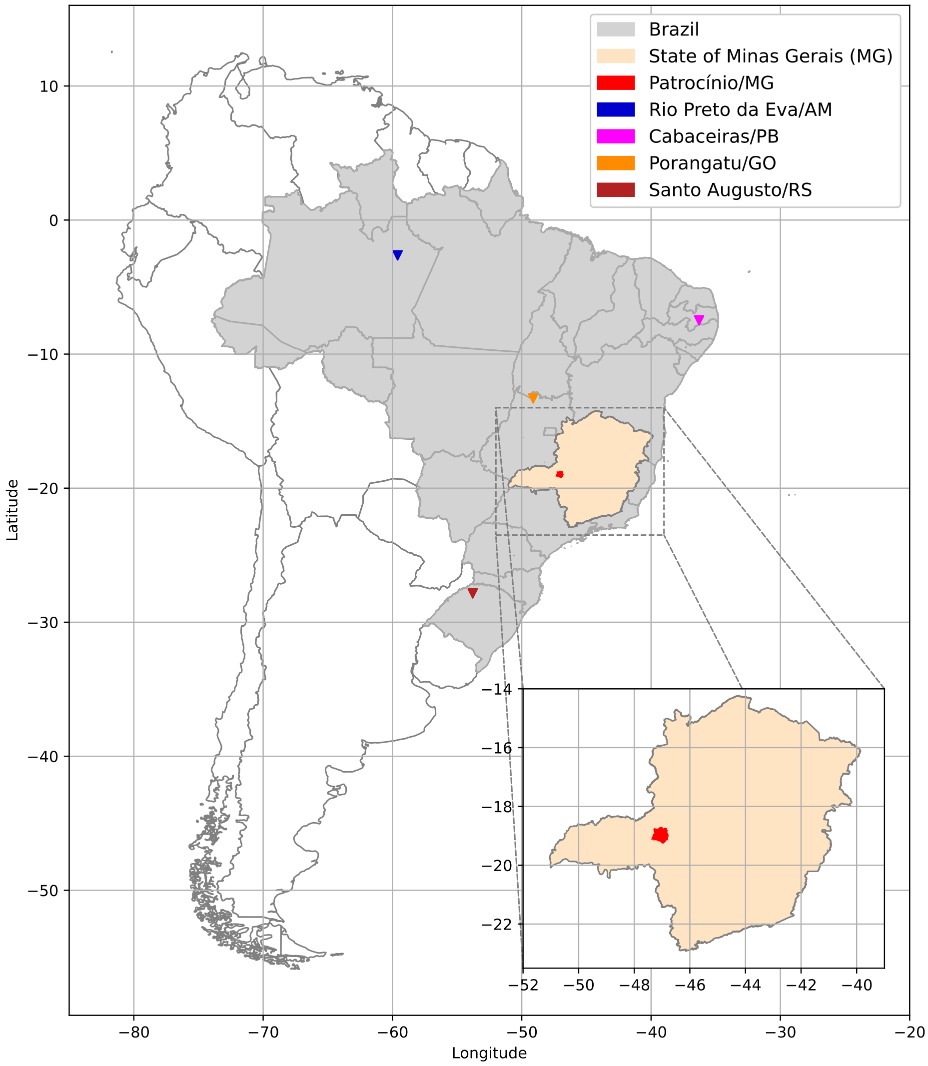
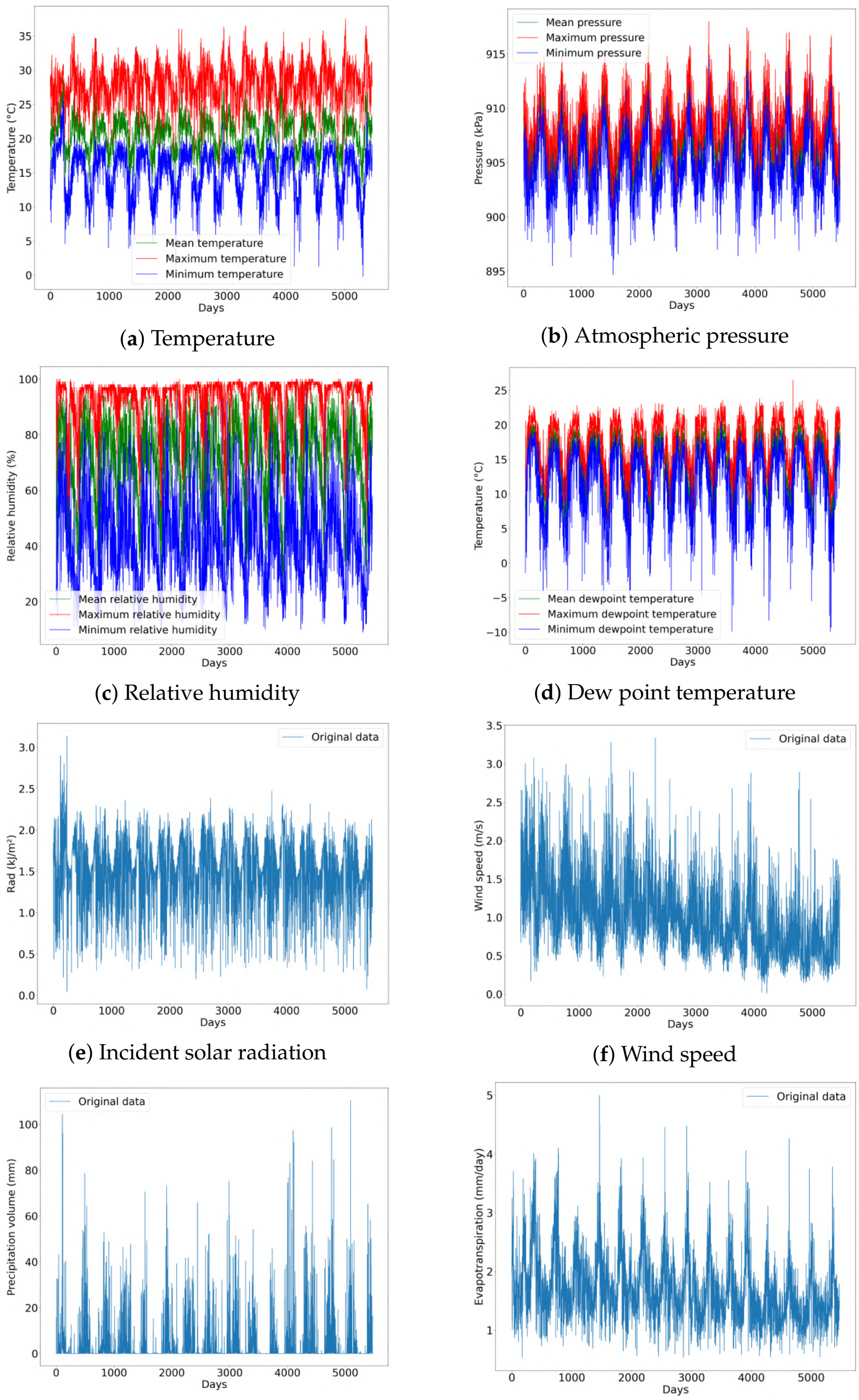


Figure 1. The geographical location of the Minas Gerais (MG) state is highlighted in yellow within Brazil's map. The red detail within the blue area highlights the Patrocínio/MG municipality's geographical location. The other four markers represent cities with different climate profiles from Patrocínio/MG's, whose climatological time series were employed by the authors to validate and deepen the discussion around the proposed model.

Table 1. Localization and profile climate of four different cities within Brazil, alongside their distances to Patrocínio/MG.

City's Name	Coordinates	Climate Profile	Distance to Patrocínio/MG (km)
Rio Preto da Eva/AM	2°41′56″ S 59°42′00″ W	Tropical rain forest (Af)	2278
Cabaceiras/PB	7°29′20″ S 36°17′13″ W	Hot semi-arid (BSh)	1723
Porangatu/GO	13°26′27″ S 49°08′56″ W	Tropical wet (Aw)	659
Santo Augusto/RS	27°51′03″ S 53°46′37″ W	Humid subtropical (Cfa)	1202



(**a**) Temperature (**b**) Atmospheric pressure

(**c**) Relative humidity (**d**) Dew point temperature

(**e**) Incident solar radiation (**f**) Wind speed

(**g**) Precipitation (**h**) Estimated evapotranspiration

Figure 2. Visualization of the complete time series for each of the fifteen climatological variables observed for the experiments in this paper, totaling 5473 samples; (**a**–**d**) follow the same colour pattern, where the red line represents the maximum value for that variable, green for the mean observed value, and blue for the minimum value; (**e**–**h**) represent the daily observations for each variable.

3.2. Data Collectors

INMET employs the commercial Vaisala HydroMat™ model MAWS301 as a standard climatological data collector for automatic stations, responsible for data acquisition and communication with the governmental systems [53]. The climatological factors monitored hourly by INMET are as follows:

- Temperature [°C], available at maximum, minimum, and instant values;
- Relative humidity [%], available at maximum, minimum, and instant values;
- Dew point temperature [°C], available at maximum, minimum, and instant values;
- Atmospheric pressure [hPa], available at maximum, minimum, and instant values;
- Wind speed and gust [$m \cdot s^{-1}$], measured at 10 m height;
- Wind direction [degrees];
- Incident solar radiation [$kJ \cdot m^{-2}$];
- Precipitation [mm].

INMET provides climatological data hourly within an open-access policy from their automatic collectors after acquisition, communication, and pre-processing. The final dataset is available at https://tempo.inmet.gov.br/TabelaEstacoes/xxxx (accessed on 1 July 2023) (site in Brazilian Portuguese), where xxxx is the required station's code. The station code for the automatic collector in Patrocínio/MG is $A523$.

3.3. Soil Characterisation

The soil characteristics around the Patrocínio/MG municipality correspond to those of the Ferralsol reference soil group, according to the International Union of Soil Sciences classification, with a high concentration of hematite [54,55].

The physical water parameters for this soil (field capacity (FC) and wilting point (WP)) alongside the water retention curve were obtained via the pressure plate method applied to three different soil samples, with one from each layer, i.e., from 0 cm to 20 cm, from 20 cm to 40 cm, and 40 cm to 60 cm.

The experimental results suggest uniform physical water parameters throughout the three layers. As such, Figure 3 presents the resulting water retention curve for the three strata.

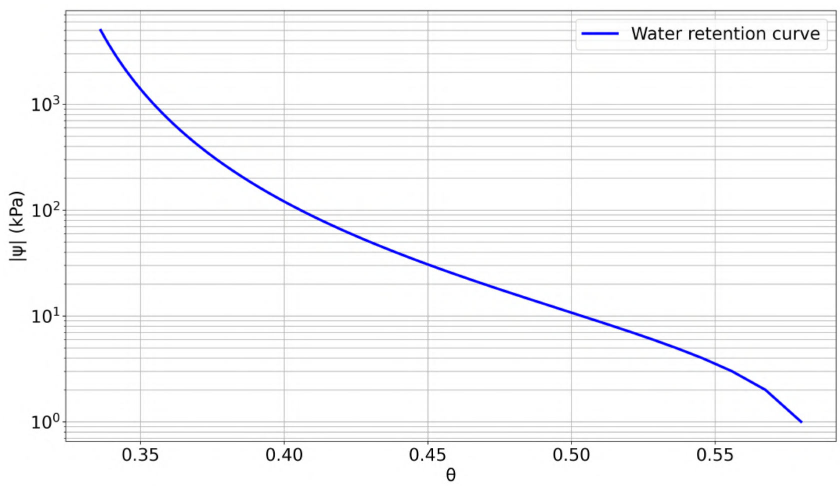


Figure 3. Semi-log representation for the water retention curve obtained from three soil samples from different depths near the Patrocínio/MG municipality. The uniform distribution of soil characteristics throughout the layers allows for the representation of a single water retention curve.

Lastly, Table 2 presents the van Genuchten equation parameters based on the observed soil retention curve in Figure 3. The authors include the parameters for each of the three layers as a matter of completeness despite their uniformity.

Table 2. Parameters for the van Genuchten equation obtained via the water retention curve and pressure plate method.

Layer	θ_s ($cm^3 \cdot cm^{-3}$)	θ_r ($cm^3 \cdot cm^{-3}$)	α (cm^{-3})	n	m
0–20 cm	0.59	0.31	0.25	1.33	0.25
20–40 cm	0.59	0.31	0.25	1.33	0.25
40–60 cm	0.59	0.31	0.25	1.33	0.25

3.4. Stochastic Modeling

The statistical models employed in this paper address three factors to assess the quality of the Bayesian regression for the evapotranspiration metric: (1) Which mathematical transformation applied to the target variable yields regression results with the lowest error metric? (2) The influence of autoregressive components on the quality of the regression. (3) An evaluation concerning the impact of the sample size on the regression learning process and its results.

The experiments evaluate five distinct operations over the evapotranspiration metric: linear—$f(x) = x$, Box-Cox transformation, square root, natural logarithm, and exponential. Equation (16) presents the definition of the one-parameter Box-Cox transformation [56] with an estimation of the value of the parameter λ from the dataset, which leads to better normality of the transformed series or is assumed to be known in each application. The number of autoregressive elements is, at most, three. Lastly, experiments employed the last 1000 or 2500 samples or the entire set with 5473 entries. A complete permutation of these components results in a total of sixty experiments.

$$f(x, \lambda) = \begin{cases} \frac{(x\lambda - 1)}{\lambda}, & \lambda \neq 0 \\ log(x), & \lambda = 0 \end{cases} \tag{16}$$

Equations (17)–(23) establish the experiments' framework based on the posterior distribution in Equation (10). Let $N \in \{1000, 2500, 5473\}$ represent the number of samples for each experiment and X_i represent each of the fifteen climatological variables employed, according to Section 3.1. Equation (17) establishes the target series y composed of the evapotranspiration series after applying one of the five equations presented, while Equation (18) provides a prior probability distribution assumed for this series based on a Gaussian density function.

Equations (19a) and (19b) describe the parametrization of the hyperparameter μ_i as a linear combination of the climatological elements. Equation (19a) applies when no autoregressive window is present, while Equation (19b) denotes the experiments with $w \in \{1, 2, 3\}$ of such components.

Lastly, Equations (20)–(23) provide initial, approximately non-informative prior distributions for the parameter τ and regression parameters β_i. Equation (23) only applies to the experiments with at least one autoregressive component.

The authors employed MultiBUGS 2.0 software [57] (available at https://www.multibugs.org/, accessed on 15 August 2023). Each experiment contained five chains, 1000 samples as the burn-in, and one in every ten values for the next 5000 obtained for the estimates.

$$y_i = f(ET_{0_i}) \qquad i = \{1, 2, \ldots, N\} \tag{17}$$

$$y_i \sim \mathcal{N}(\mu_i, \tau) \qquad i = \{1, 2, \ldots, N\} \tag{18}$$

$$\mu_i = \beta_0 + \sum_{j=1}^{15} \beta_j X_{ji} + \varepsilon \qquad i = \{1, 2, \ldots, N\} \tag{19a}$$

$$\mu_i = \beta_0 + \sum_{j=1}^{15} \beta_j X_{ji} + \sum_{k=1}^{w} \beta_k y_{(i-k)} + \varepsilon \qquad i = \{1, 2, \ldots, N\}, \quad w \in \{1, 2, 3\} \tag{19b}$$

$$\tau \sim \Gamma(1, 1) \tag{20}$$

$$\beta_0 \sim \mathcal{N}(4, 0.01) \tag{21}$$

$$\beta_j \sim \mathcal{N}(0, 1) \qquad j = \{1, 2, \ldots, 15\} \tag{22}$$

$$\beta_k \sim \mathcal{N}(0, 1) \qquad k = \{1, \ldots, w\} \tag{23}$$

3.5. Structural Learning

To build a BN that presents the best likelihood to the climatological and evapotranspiration data, the authors employed both the standard $K2$ greedy algorithm [58] and a bio-inspired score-and-search algorithm [59].

The relevant nodes and the amount of required data were consonant with the results of the set of sixty experiments described in Section 3.3, resulting in eight climatological variables, alongside evapotranspiration as the target component and a dataset with 1000 samples.

The parameters for both methods were the default ones, with the Bayesian Dirichlet equivalence with the uniform prior metric (BDeu) [60] as a scoring function and k-fold (k = 5) cross-validation. Both methods resulted in the same graphical structure.

4. Results and Discussion

This section presents the results and relevant discussions about the proposed experiments. Appendix A presents the complete tables for the numeric results of all the experiments, and presents the graphical comparisons of the evapotranspiration metric, considering the inference task of each climatological variable.

4.1. Selection of Climatological Variables

The root mean square error (RMSE) metric for each of the sixty experiments against the expected time series for the Penman–Monteith equation over the original dataset ranges from 0.086 mm $\cdot$ day^{-1} to 3.502 mm $\cdot$ day^{-1}. These results range from the experiment with logarithmic transformation, one autoregressive element, and 1000 samples, to the test with the exponential function, no autoregressive terms, and 5473 samples, respectively. Table A1 presents the RMSE results for the sixty experiments proposed in Section 3.4, alongside their mean absolute error (MAE) metrics; the closer the value of each error metric to zero, the better adherence between the two series.

The total time required to reach convergence for each experiment on MultiBUGS depends mainly on the amount of data provided to the software. Experiments with 1000 samples took 12.2 min, 2500 samples required 36.3 min, and the complete dataset needed 67.3 min, on average. These times were achieved on a desktop with an Intel® Core™ i5-6600 CPU at 3.30 GHz, 8 GB of RAM, and operating a Windows® 10 Pro 64-bit.

The first finding is that the RMSE metric exhibits minor fluctuations within the same transformation but higher fluctuations across functions. For instance, the experiments within the logarithm function present RMSE from 0.086 mm $\cdot$ day^{-1} to 0.100 mm $\cdot$ day^{-1}, while experiments within the Box-Cox transform—with each λ parameter automatically obtained by the transformation algorithm for each series—achieve RMSEs from 0.097 mm $\cdot$ day^{-1} to 0.107 mm $\cdot$ day^{-1}. These results suggest that the mathematical function choice is a relevant first step to minimize the estimation error based on this approach.

Within the same mathematical transformation, when comparing the error metric according to the number of samples of each experiment, the error metric consistently presents the lowest value for experiments with 2500 samples. Although this improvement is marginal compared to the errors from those with 1000 samples, this result suggests the existence of an optimal data volume for the inference task on these variables, and a large volume of data may impair the inference task by introducing more noise into the model.

Finally, when observing the error performance based on the amount of past information provided to the model, there is a marginal improvement across all sixty experiments for each past value added. This result is consistent with the observed meteorological dynamics, as the amount of water evaporated and transpired from the previous day plays a role in the maintenance of relative humidity, dew point temperature, and surface temperature for the current day, which are factors that directly impact the evapotranspiration phenomenon.

The exponential transformation is the one that consistently presents inference results with the highest error metric, according to Table A1. The results and discussions from this point onward do not consider the experiments with the exponential function, given their errors within one or two orders of magnitude greater than those obtained by the other four transformations.

The output of the MultiBUGS software includes the interval at 95% credibility for each beta regression parameter and the most probable value within this interval. As such, a climatological element is relevant if the credible interval for its beta coefficient does not encompass zero, as any value for the climatological variable would represent no impact in the regression.

Table A2 contains a complete overview of the occurrence of each climatological variable deemed relevant by the regression model, grouped by transformation. Across all forty-eight experiments—excluding the twelve from the exponential function—precipitation and all three measurements for atmospheric pressure do not contribute to the regression model. The variables for maximum dew point temperature and minimum temperature and dew point temperature contributed to a few experiments.

Eight variables consistently supported the regression analysis in more than half of the twelve experiments for each function. Wind speed and net solar radiation are relevant across all forty-eight experiments; mean temperature and dew point temperature, maximum temperature and relative humidity, and minimum relative humidity are significant for several experiments. Mean relative humidity is the last of the relevant climatological variables, although this component did not contribute to any regression with the square root function.

The high relevance observed from the wind speed and net solar radiation throughout all experiments is consistent with the observed meteorological phenomena. Both components directly contribute to the regional climate, represented by temperature, relative humidity, and dew point temperature. The irrelevance of the precipitation is also within expectation, given the assumption that elements within a much wider region contribute to this component with their dynamics not considered here. The atmospheric pressure components directly impact the dew point temperature, and the absence of this variable from the set of relevant variables is likely due to a discrepancy in the magnitude of the pressure measurements compared to the other elements

While Table A1 shows that the natural logarithm is the function that provides a regression with the lowest errors, Table A2 presents this function as the most assertive to which variables are most relevant to the stochastic regression for evapotranspiration. Further, the error improvements within this transformation are negligible after providing more data and autoregressive components, given the trade-off of obtaining these. Based on these considerations, the authors establish that the stochastic model with a natural logarithm transformation and no autoregressive elements, trained with 1000 samples, shows the best adherence to the evapotranspiration metric.

Table 3 presents the numerical values achieved for the beta coefficients from the experiment utilizing natural logarithm transformation on the evapotranspiration data, with 1000 samples for the parameter learning and no autoregressive component. The coefficients and their respective climatological variables—whose 95% credible intervals do not contain zero and, as such, characterize a relevant variable in this context—are highlighted in blue. The independent coefficient β_0 for this experiment is $\beta_0 \approx 0.6373$.

Based on the maximum a posteriori (MAP) values of each beta parameter and the relationships detailed in Equations (17) and (19a), Equation (24) presents the approximation for the hierarchical Bayesian regression for the evapotranspiration metric; this is based on the experiment identified as the best among the sixty conducted. Although this representation for Equation (24) is similar to linear regression, each beta coefficient corresponds to a density distribution function, and its best point estimator value lies within the non-symmetrical 95% credible interval.

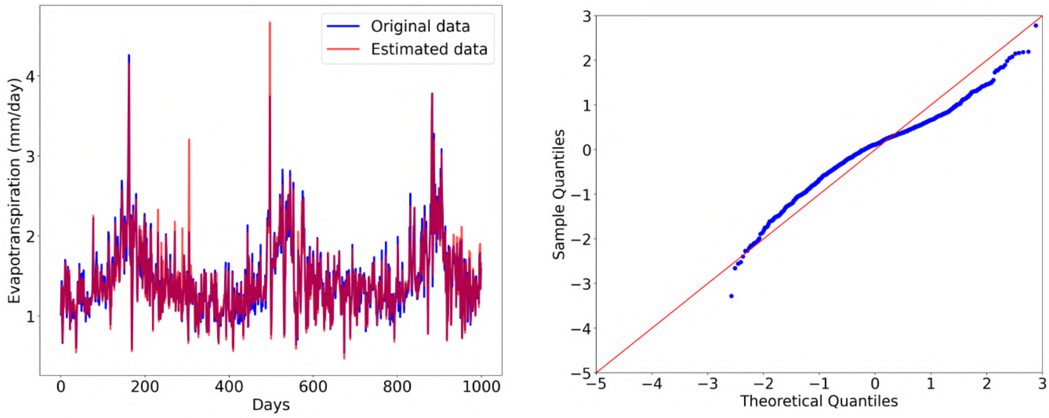
$$\begin{aligned} ln(ET_0) \approx 0.6373 - 0.024 \cdot T_{mean} + 0.022 \cdot T_{max} \\ - 0.006 \cdot RH_{mean} - 0.005 \cdot RH_{max} - 0.010 \cdot RH_{min} \\ + 0.028 \cdot DP_{mean} + 0.498 \cdot W_{speed} + 0.113 \cdot Rad \end{aligned} \tag{24}$$

Table 3. Credible interval at 95% for the beta regression coefficients obtained from the experiment with natural logarithm transformation, using a data partition of 1000 samples, and excluding the autoregressive window feature. The relevant beta coefficients (i.e., those that do not include zero on the interval or whose value is close to zero) are highlighted in blue.

Variable	Coefficient	2.5%	97.5%	MAP Value
Tmean	β_1	−0.035	−0.010	−0.024
Tmax	β_2	0.014	0.028	0.022
Tmin	β_3	−0.008	0.003	0
RHmean	β_4	−0.009	−0.004	−0.006
RHmax	β_5	−0.005	−0.004	−0.005
RHmin	β_6	−0.011	−0.008	−0.010
DPmean	β_7	0.014	0.039	0.028
DPmax	β_8	−0.005	0.006	0
DPmin	β_9	−0.008	0.004	0
Pmean	β_{10}	−0.001	0.002	0
Pmax	β_{11}	−0.003	0.001	0
Pmin	β_{12}	−0.003	0.001	0
Wspeed	β_{13}	0.484	0.513	0.498
Rad	β_{14}	0.093	0.133	0.113
Prec.	β_{15}	−0.001	0.000	0

Figure 4a contains a graphical comparison of the evapotranspiration time series based on the Penman–Monteith equation (blue continuous line) against the estimated time series from Equation (24) (red line) throughout 1000 samples. This result suggests that the proposed method and subsequent equation properly conciliated the series' non-stationarity.

It is relevant to evaluate if the error component follows a Gaussian distribution for the estimated time series against the expected one, as the proposed Bayesian regression in Equation (24) is similar to the model in Equation (19a). To this end, Figure 4b presents a quantile–quantile plot for the distribution of the residual errors. As the blue dots lie close to the red 45-degree line, the distribution for the error is close to a Gaussian distribution, as required by the theoretical model.



(**a**) Comparison. (**b**) Quantile–quantile plot.

Figure 4. Graphical representation of the results obtained from the experiment with natural logarithm transformation, using a data partition of 1000 samples, and excluding the autoregressive window feature; (**a**) presents the original time series as the blue line and the estimated time series as the red line; (**b**) presents the quantile–quantile plot of the residuals for the estimated time series.

One relevant comparison is to ascertain the proposed method's adherence to the expected evapotranspiration according to the gold-standard FAO56 method against the evapotranspiration estimated by other classical and deterministic approaches. As such, Figure 5 presents a graphical comparison between the natural logarithm of the evapotranspiration from the FAO56 method (blue line), the Benavides and Lopez method (orange line), and the Hargreaves–Samani method (green line); all three are based on the same dataset, alongside the estimate from the proposed method (red line). This result confirms the ability of the proposed approach to closely model the complex behavior of the FAO56 estimate, with an adherence better than the ones presented by other classical equations.

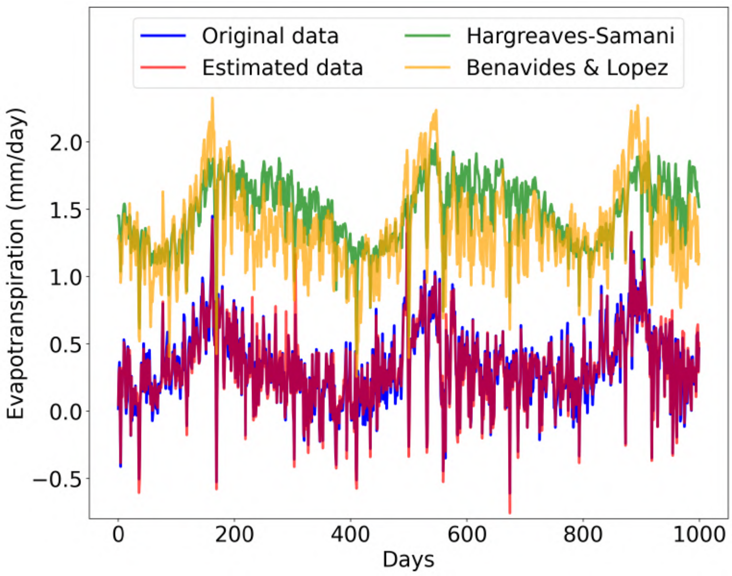


Figure 5. Graphical comparison of the results obtained from the experiment with natural logarithm transformation, using a data partition of 1000 samples, and excluding the autoregressive window feature (as the red line). The blue line represents the natural logarithm of the reference evapotranspiration according to the FAO-56 equation; the orange and green lines represent the natural logarithm for the reference evapotranspiration according to the Benavides and Lopez and the Hargreaves–Samani equations, respectively. The latter two methods represent some of the analytical approximations commonly employed under the unavailability of the climatological components required by the FAO-56 equation.

4.2. Stochastic Climatological Water Balance Comparison with In Situ Measurements

A proxy measurement for the water balance dynamics is necessary, mainly due to the characteristics of the evapotranspiration metric. Given the measurement hardship, this paper employs an estimate for the available soil water content on a coffee plantation in the proximity of Patrocínio/MG based on tensiometer readings and the van Genuchten equation.

Although the authors cannot divulge the precise location of this plantation due to confidentiality concerns, Equation (25) presents the estimated soil water content θ as a function of the matric potential ψ (kPa), according to the original Equation (7). For this particular soil, $\theta_s \approx 59\%$ and $\theta_r \approx 31\%$, $\alpha \approx 0.25\ \text{cm}^{-1}$, $n \approx 1.33$ and $m \approx 0.25$, based on the experiments described in Section 3.3. Tensiometer readings were provided by a PalmaFlex digital tensiometer, model SL-PF-TNSE.

$$\theta = vG(\psi) \approx 0.31 + \frac{0.28}{\left(1 + (0.25|\psi|)^{1.33}\right)^{0.25}} \tag{25}$$

Figure 6 presents the relationships between rainfall, estimated soil water content, and estimated evapotranspiration on the top panel, based on in situ measurements from a tensiometer monitoring the substrate from the surface to a depth of 20 cm, and from a pluviometer, with both sampling data from the 9th to the 25th January 2021. In the highlighted period (from the 9th to the 15th), the observed rainfall is enough to keep the soil water content close to its maximum capacity (≈59%). From the 16th onward, when there is no observed rainfall, the curves for soil water content and evapotranspiration have a symmetrical relationship: the higher the estimated evapotranspiration, the faster the reduction in soil water content. This characteristic against the in situ measurements reinforces the Penman–Monteith method as adequate to model the consumptive use of water due to evaporation and transpiration.

The middle panel of Figure 6 presents the influence on the estimated evapotranspiration by the two most relevant climatological variables, according to the MAP values in Equation (24). This panel exhibits the correlations for evapotranspiration trends with wind speed ($\beta_{13} \approx 0.113$) and net incident solar radiation ($\beta_{14} \approx 0.498$). These relationships of the target variable with the two most relevant climatological factors, according to the Bayesian regression model, combined with the adherence observed in the first panel against the in situ measurements, suggest the adequacy of the proposed method to represent the hydrological phenomenon, even under scenarios where the traditional method cannot be employed.

Lastly, the third panel compares the analytical time series for evapotranspiration according to the Penman–Monteith metric and the results from Equation (24) over the same dataset. These three panels suggest that this method models the expected analytical time series and the underlying real-world dynamics.

These results support the adequacy of the proposed method in tracking the analytical FAO-56 equation and functioning as a framework to model dynamic and complex phenomena based on non-linear, non-stationary time series and their underlying factors.

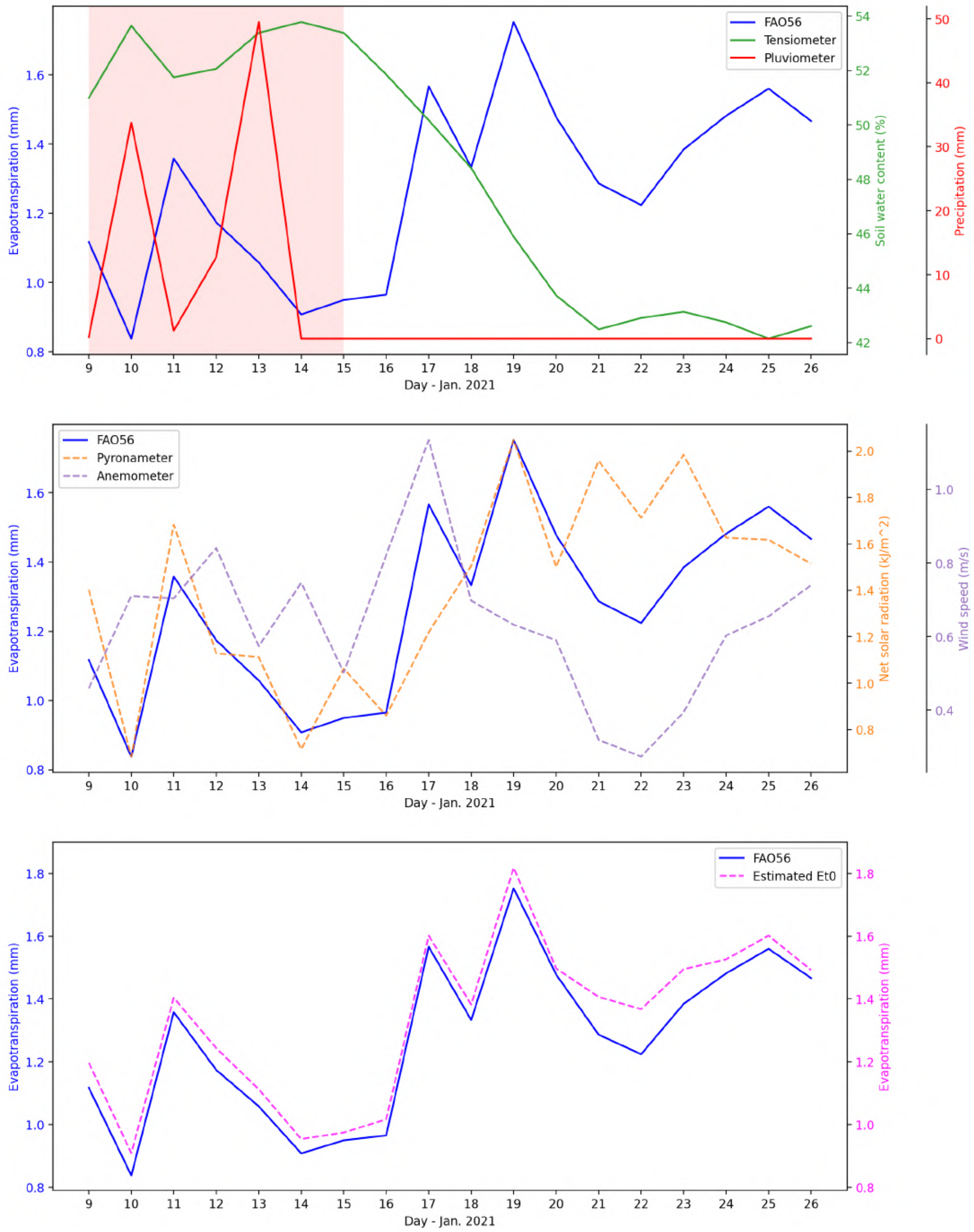


Figure 6. Graphical representation of the climatological water balance compared to in situ samplings from the 9th to the 26th of January 2021. The top panel contains a time series for the two principal factors of the climatological water balance—rainfall and evapotranspiration—compared to the dynamics of the soil water percentage. The area highlighted in red represents the days when the water intake to the plantation due to rainfall was sufficient to counteract the amount of water evaporated and transpired. The middle panel presents the interaction between the estimated evapotranspiration, wind speed, and incident net solar radiation, the latter two being the most relevant factors for evapotranspiration estimation based on the results presented in Section 4.1. Lastly, the bottom panel presents the estimated evapotranspiration based on the FAO-56 method compared to the values obtained by Equation (24), applied to the same dataset during the established timeframe.

4.3. Stochastic Evapotranspiration Inference with Complete or Incomplete Data

To evaluate the model's adequacy in estimating evapotranspiration over a dataset with new information, Figure 7 presents a graphical comparison of the natural logarithm of the evapotranspiration calculated with the Penman–Monteith equation against the time series obtained from Equation (24). This new dataset, comprising 334 samples not used in the learning process, contains complete information on all the climatological variables presented in Section 3.1.

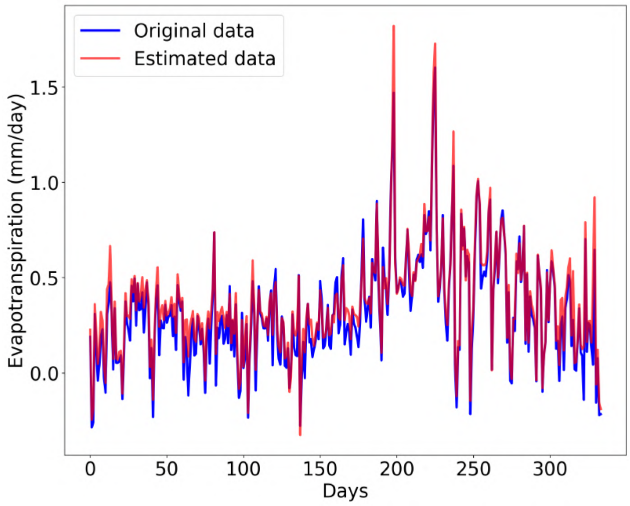
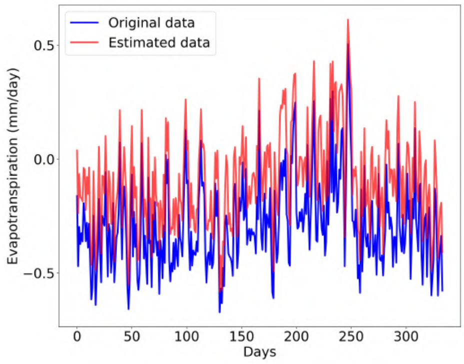


Figure 7. Graphical comparison of the reference evapotranspiration from the Penman–Monteith equation over the complimentary dataset against that obtained via the regression presented in Equation (24). The RMSE metric between the natural logarithm of both series is 0.087 mm $\cdot$ day^{-1}, consistent with the results in Table A1.

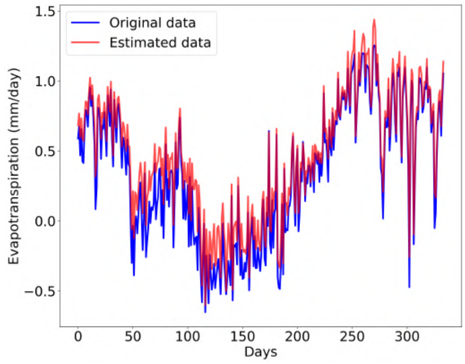
Another way to check the model's performance is to compare the expected time series for the Penman–Monteith equation against the result from Equation (24) when dealing with climatological data from regions with different climate profiles. Figure 8a–d provide a graphic comparison similar to that in Figure 7, but this time it is applied to the cities described in Table 1, with RMSE metrics and distances from Patrocínio presented in Table 4. As expected, the RMSE error for each region increases as the distance from the original city increases, reinforcing the notion that Penman–Monteith equation relies on local climatological factors for an accurate estimate.

Table 4. Error metrics for the natural logarithm of the reference evapotranspiration derived from the Penman–Monteith equation versus the values from Equation (24), based on the climatological data from different cities in Brazil with different climate profiles compared to Patrocínio/MG.

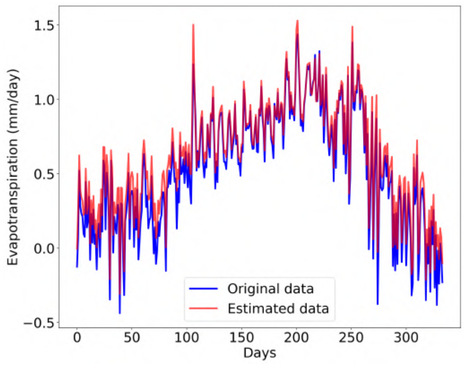
City	RMSE (mm $\cdot$ day^{-1})	Distance to Patrocínio/MG (km)
Rio Preto da Eva/AM	0.202	2278
Cabaceiras/PB	0.140	1723
Porangatu/GO	0.102	659
Santo Augusto/RS	0.185	1202



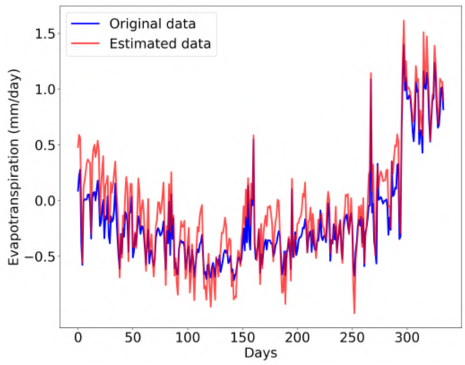
(**a**) Rio Preto da Eva/AM



(**b**) Cabaceiras/PB



(**c**) Porangatu/GO



(**d**) Santo Augusto/RS

Figure 8. Graphical comparison of the reference evapotranspiration from the Penman–Monteith equation over the complimentary dataset for each different city presented in Table 1, against the value obtained via the regression presented in Equation (24) over that city's climatological time series; (**a**) presents the comparison for Rio Preto da Eva/AM, which achieved RMSE = 0.202 mm $\cdot$ day^{-1}; (**b**) presents the comparison for Cabaceiras/PB, which achieved RMSE = 0.140 mm $\cdot$ day^{-1}; (**c**) presents the comparison for Porangatu/GO, which achieved RMSE = 0.102 mm $\cdot$ day^{-1}; lastly, (**d**) presents the comparison for Santo Augusto/RS, which achieved RMSE = 0.185 mm $\cdot$ day^{-1}.

One shortcoming of every analytical approach to estimating evapotranspiration is their prerequisite of having all the required climatological variables available. Under the Bayesian framework, the BN provides a probabilistic approach to supply the most probable value for missing data based on complete or incomplete evidence from the system.

Figure 9 presents a graphical representation of the BN's DAG, which provides the best posterior likelihood estimates based on the original dataset containing 1000 samples. The nodes represent only the random variables relevant to the Bayesian regression from Table 3 instead of the complete set of fifteen variables. This approach reduces the graph search space for the learning algorithm and employs only the relevant climatological variables, which decreases the number of required components and sensors. The evapotranspiration node is highlighted in blue; the omission of the conditional probability tables from the BN is intentional due to the high dimensionality of each one.

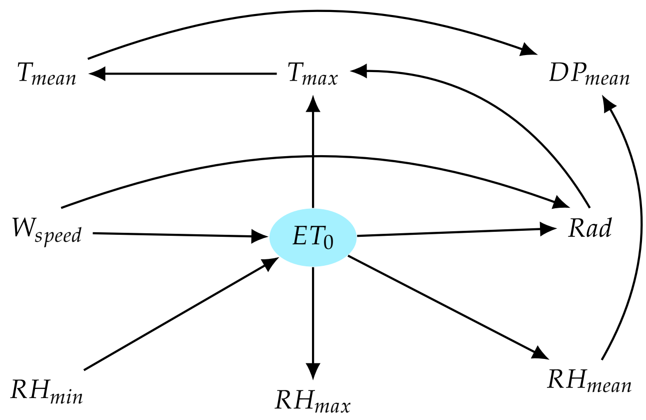


Figure 9. Bayesian network for estimating the natural logarithm of reference evapotranspiration from the climatological variables most frequently present as relevant, according to the set of experiments with natural logarithm transformation.

After learning the conditional probability tables based on the maximum likelihood of the model over the training dataset, the BN framework represented in Figure 9 allows for queries about the most probable state of a climatological variable, providing the Bayesian regression analysis with a new level of flexibility to deal with missing data. For instance, if a meteorological data collector experiences a failure in its anemometer, an estimate of the most probable value for wind speed may allow for the use of the underlying methods that require this variable, albeit within an error margin. Figure A1 presents the natural logarithm of the expected evapotranspiration time series against the values obtained from Equation (24) after replacing each climatological variable from the original dataset with values inferred from the BN.

As expected, wind speed estimates based on the BN provide a time series with the lowest adherence to the baseline time series. This result occurs because the wind speed is one of the independent variables in the BN, resulting in the inference task for this variable becoming a query for the most probable value that would induce the observed values for the other variables in the system. This type of inverted query increases the error margin for the estimated value and, alongside the relatively high MAP value for wind speed's beta coefficient, culminates in the amplification of the inference error on the final estimate for evapotranspiration when wind speed data are lacking.

Table 5 presents the RMSE metric for the evapotranspiration estimates according to Equation (24), considering the removal and subsequent inference of each one of the climatological variables from the BN in Figure 9 and the expected evapotranspiration time series. The variables presenting the highest error values are wind speed and minimum relative humidity, which is reasonable within the BN structure as both components are independent elements of the BN. The lowest error values are from the maximum relative humidity component, likely due to its position on the graphical structure, where only one random variable influences it, and it does not influence any others, and due to the lowest absolute MAP value of the beta parameter for maximum relative humidity in Equation (24).

The results in Table 5 suggest that the BN approach provides efficient estimates for the required variable under scenarios lacking a single component and allows for the maintenance and usage of methods that may require these missing factors.

Table 5. Error metric for the natural logarithm of the reference evapotranspiration metric from the Penman–Monteith equation against the value from Equation (24), after the removal of each relevant climatological variable and the inference of its value from the BN in Figure 9.

Removed and Inferred Variable	Evapotranspiration RMSE ($\mathrm{mm \cdot day^{-1}}$)
T_{mean}	0.087
T_{max}	0.092
RH_{mean}	0.087
RH_{max}	0.074
RH_{min}	0.156
DP_{mean}	0.085
W_{speed}	0.286
Rad	0.096

5. Conclusions

In this paper, the authors provided and discussed a complete stochastic and Bayesian framework to model a real-world application based on non-deterministic and non-stationary data.

The first result shows how the Bayesian regression analysis is affected by the employed mathematical pre-processing operation, the amount of provided data, and the number of available autoregressive components for this application. Based on the RMSE metric, the natural logarithm function is the best candidate to model the evapotranspiration time series.

The established baseline Bayesian regression equation provides a close approximation for the evapotranspiration metric modeled by the Penman–Monteith equation, and for the dynamics of soil water moisture measured by tensiometers. As such, the results from the entirely data-driven approach are coherent with the traditional analytical method and the real-world phenomenon it proposes to model.

The Bayesian regression analysis also highlights the subset of random variables relevant to the stochastic model. We leverage this newfound insight to build another Bayesian structure in the form of a BN, as the knowledge of relevant variables allows for a reduction in the graph‘s search space.

Lastly, we show the adequacy of probabilistic queries over the BN structure to estimate climatological variables required for the analytical methods. Our results show that the BN toolset provides synthetic data with a high degree of adherence to real-world data and, in turn, allows for the application of subsequent methods. As expected from the theoretical basis, this inference task presents the worst performance when estimating a value for independent nodes on the graphical structure. The flexibility the BN provides for analytical methods may even offset the time and computational power required to learn an adequate configuration for the query process.

Author Contributions: Conceptualisation, C.C.M.J. and C.D.M.; data curation, V.P.R. and L.D.N.; formal analysis, V.P.R., P.A.A.M., J.A.A. and C.D.M.; investigation, V.P.R., A.W.C.J. and C.C.M.J.; methodology, V.P.R., L.D.N., P.A.A.M., J.A.A., C.D.M. and J.A.P.B.; project administration, C.D.M. and J.A.P.B.; software, V.P.R. and L.D.N.; supervision, C.D.M. and J.A.P.B.; validation, P.A.A.M., J.A.A. and C.D.M.; writing—original draft, V.P.R., A.M.J., C.D.M. and J.A.P.B.; writing—review and editing, L.D.N., P.A.A.M., J.A.A., A.W.C.J., C.D.M. and J.A.P.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the National Council for Scientific and Technological Development (CNPq), grant nos. 142390/2019-4 (V.R.), 406774/2022-6 (P.M.), 301923/2019-1 (J.A.), 465755/2014-3 (C.M.), and 301853/2018-5 (J.B.), and by the São Paulo Research Foundation (FAPESP), grant nos. 2014/50851-0 (C.M.), USP-FAPESP-IBM 2019/07665-4 (L.N. and C.M.), and BPE-FAPESP 2018/19150-6 (C.M.). The APC was funded by Espectro Ltda., under grant no. 001/2019.

Data Availability Statement: Publicly available datasets were analyzed in this study concerning climatological variables. These datasets can be found at https://tempo.inmet.gov.br/TabelaEstacoes/xxxx (accessed on 1 July 2023) (site in Brazilian Portuguese), where xxxx is the required station code. Please note that there are restrictions on the availability of data from tensiometer readings. Data were obtained from a proprietary coffee plantation site under the condition of anonymity and cannot be made publicly available.

Acknowledgments: The authors thank the Foundation for UNESP's Development (FUNDUNESP), UNESP's Innovation Agency (AUIN), and the Centre for Artificial Intelligence (C4AI-USP).

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A

Appendix A.1. Complete Table for the Error Metrics

Table A1. The RMSEs and MAEs of all sixty experiments, combining each proposed transformation, original dataset partition, and autoregressive window. From a scoring standpoint, every experiment achieved error metrics within the same order, except for those with the exponential transformation.

Transformation	N° of Samples	Window	RMSE ($mm \cdot day^{-1}$)	MAE ($mm \cdot day^{-1}$)
Linear	1000	0	0.115	0.074
		1	0.112	0.073
		2	0.112	0.073
		3	0.112	0.073
	2500	0	0.118	0.076
		1	0.116	0.075
		2	0.116	0.075
		3	0.116	0.075
	5473	0	0.139	0.090
		1	0.134	0.089
		2	0.134	0.088
		3	0.133	0.088
Box-Cox	1000	0	0.107	0.054
		1	0.105	0.054
		2	0.105	0.054
		3	0.105	0.054
	2500	0	0.097	0.059
		1	0.095	0.058
		2	0.095	0.058
		3	0.095	0.058
	5473	0	0.101	0.070
		1	0.098	0.069
		2	0.097	0.068
		3	0.097	0.068

Table A1. *Cont.*

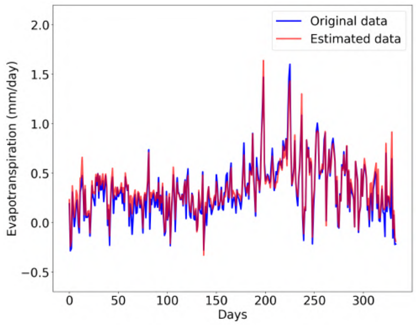
Transformation	N° of Samples	Window	RMSE ($mm \cdot day^{-1}$)	MAE ($mm \cdot day^{-1}$)
Square root	1000	0	0.094	0.061
		1	0.092	0.061
		2	0.092	0.061
		3	0.092	0.061
	2500	0	0.093	0.062
		1	0.092	0.062
		2	0.092	0.061
		3	0.092	0.061
	5473	0	0.108	0.074
		1	0.105	0.072
		2	0.104	0.072
		3	0.104	0.072
Natural logarithm	1000	0	0.088	0.054
		1	0.086	0.053
		2	0.086	0.054
		3	0.086	0.053
	2500	0	0.088	0.058
		1	0.087	0.057
		2	0.086	0.057
		3	0.086	0.057
	5473	0	0.100	0.070
		1	0.098	0.069
		2	0.097	0.068
		3	0.097	0.068
Exponential	1000	0	2.408	1.010
		1	2.296	0.979
		2	2.283	0.988
		3	2.282	0.993
	2500	0	2.190	1.029
		1	2.129	1.007
		2	2.130	1.010
		3	2.128	1.009
	5473	0	3.502	1.609
		1	3.458	1.560
		2	3.430	1.561
		3	3.429	1.560

Appendix A.2. Complete Table for the Relevant Variables

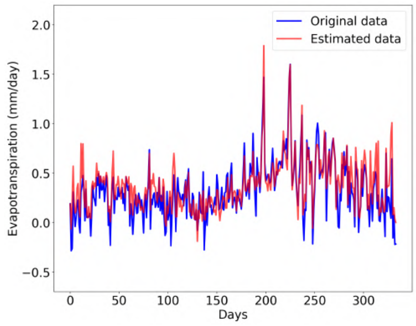
Table A2. The frequency of each climatological variable found relevant for each set of twelve experiments, according to the transformation applied to the evapotranspiration metric. The variables most frequently present are highlighted in blue.

Variable	Linear	Box-Cox	Square Root	Natural Logarithm
T_{mean}	6	12	8	12
T_{max}	9	12	12	12
T_{min}	3	4	1	4
RH_{mean}	8	10	0	11
RH_{max}	11	8	4	12
RH_{min}	11	12	12	12
DP_{mean}	6	12	8	12
DP_{max}	2	1	0	0
DP_{min}	8	1	1	2
P_{mean}	0	0	0	0
P_{max}	0	0	0	0
P_{min}	0	0	0	0
W_{speed}	12	12	12	12
Rad	12	12	12	12
$Prec.$	0	0	0	0
$ET_0[-1]$	9/9	9/9	9/9	9/9
$ET_0[-2]$	4/6	4/6	4/6	4/6
$ET_0[-3]$	1/3	1/3	1/3	2/3

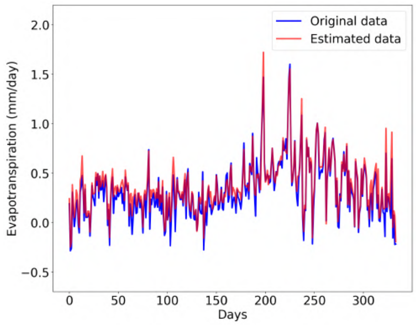
Appendix A.3. Graphical Comparison between the Expected Time Series and Inferred Data



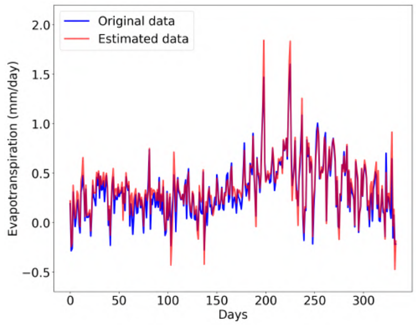
(**a**) Maximum relative humidity



(**b**) Minimum relative humidity

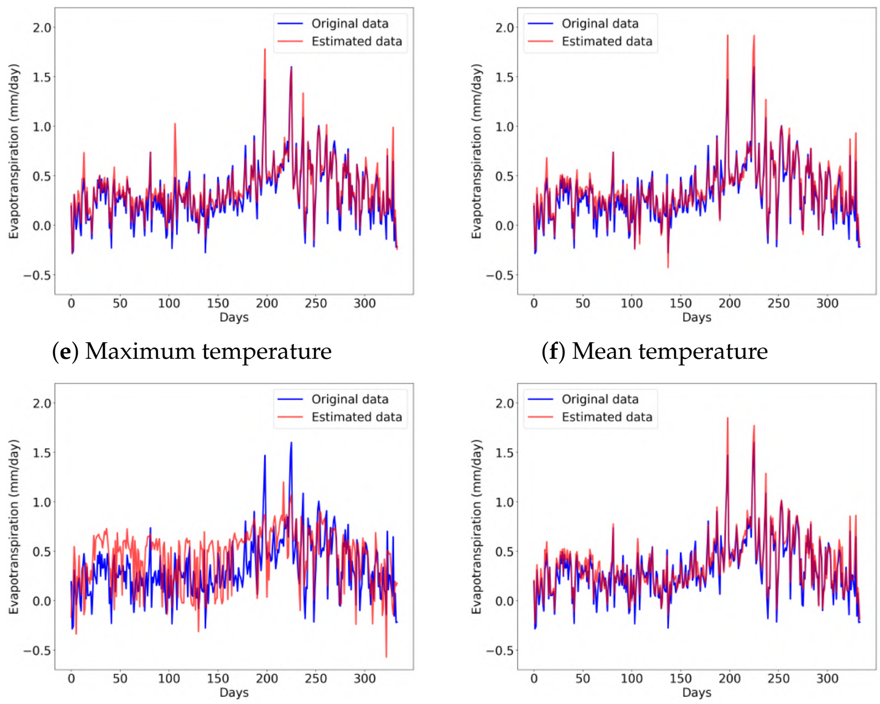


(**c**) Mean relative humidity



(**d**) Mean dew point temperature

Figure A1. *Cont.*



(**e**) Maximum temperature (**f**) Mean temperature

(**g**) Wind speed (**h**) Net solar radiation

Figure A1. A graphical comparison between the expected time series and the results obtained after each climatological variable was removed from the dataset; subsequent inference of the components based on the Bayesian Network structure.

References

1. Lu, Y.; Yuan, J.; Lu, X.; Su, C.; Zhang, Y.; Wang, C.; Cao, X.; Li, Q.; Su, J.; Ittekkot, V.; et al. Major threats of pollution and climate change to global coastal ecosystems and enhanced management for sustainability. *Environ. Pollut.* **2018**, *239*, 670–680. [CrossRef]
2. Chen, B.; Han, M.; Peng, K.; Zhou, S.; Shao, L.; Wu, X.; Wei, W.; Liu, S.; Li, Z.; Li, J.; et al. Global land-water nexus: Agricultural land and freshwater use embodied in worldwide supply chains. *Sci. Total Environ.* **2018**, *613–614*, 931–943. [CrossRef]
3. Jägerbrand, A.K.; Brutemark, A.; Barthel Svedén, J.; Gren, I.M. A review on the environmental impacts of shipping on aquatic and nearshore ecosystems. *Sci. Total Environ.* **2019**, *695*, 133637. [CrossRef]
4. Hecht, J.S.; Lacombe, G.; Arias, M.E.; Dang, T.D.; Piman, T. Hydropower dams of the Mekong River basin: A review of their hydrological impacts. *J. Hydrol.* **2019**, *568*, 285–300. [CrossRef]
5. Koumparou, D. The Right of Thirst: Water as a Human Right and as a Commons. *Glob. Nest J.* **2018**, *20*, 637–645. [CrossRef]
6. United Nations Department for Economic and Social Affairs. *Sustainable Development Goals Report 2020*; United Nations: New York, NY, USA, 2020.
7. Empresa de Pesquisa Energética. Anuário Estatístico de Energia Elétrica 2022. 2022. Available online: https://www.epe.gov.br/sites-pt/publicacoes-dados-abertos/publicacoes/PublicacoesArquivos/publicacao-160/topico-168/Fact%20Sheet%20-%20Anu%C3%A1rio%20Estat%C3%ADstico%20de%20Energia%20El%C3%A9trica%202022.pdf (accessed on 14 July 2023).
8. FAO. FAOSTAT: Crops and Livestock Products. 2021. Available online: https://www.fao.org/faostat/en/#data/QCL (accessed on 2 February 2023).
9. Rasera, J.B.; Silva, R.F.D.; Piedade, S.; Mourão Filho, F.D.A.A.; Delbem, A.C.B.; Saraiva, A.M.; Sentelhas, P.C.; Marques, P.A.A. Do Gridded Weather Datasets Provide High-Quality Data for Agroclimatic Research in Citrus Production in Brazil? *AgriEngineering* **2023**, *5*, 924–940. [CrossRef]
10. Bwambale, E.; Abagale, F.K.; Anornu, G.K. Data-driven model predictive control for precision irrigation management. *Smart Agric. Technol.* **2023**, *3*, 100074. [CrossRef]
11. Porter, J.; Xie, L.; Challinor, A.; Cochrane, K.; Howden, S.; Iqbal, M.; Lobell, D.; Travasso, M. Chapter 7: Food Security and Food Production Systems. In *Food Security and Food Production Systems. In: Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Chan*; Cambridge University Press: Cambridge, UK, 2014; pp. 485–533.
12. Neupane, D.; Adhikari, P.; Bhattarai, D.; Rana, B.; Ahmed, Z.; Sharma, U.; Adhikari, D. Does Climate Change Affect the Yield of the Top Three Cereals and Food Security in the World? *Earth* **2022**, *3*, 45–71. [CrossRef]
13. Zhang, C.; Xie, Z.; Wang, Q.; Tang, M.; Feng, S.; Cai, H. AquaCrop modeling to explore optimal irrigation of winter wheat for improving grain yield and water productivity. *Agric. Water Manag.* **2022**, *266*, 107580. [CrossRef]

14. Silva Fuzzo, D.F.; Carlson, T.N.; Kourgialas, N.N.; Petropoulos, G.P. Coupling remote sensing with a water balance model for soybean yield predictions over large areas. *Earth Sci. Inform.* **2020**, *13*, 345–359. [CrossRef]
15. Allen, R.; Pereira, L.; Raes, D.; Smith, M. *Crop Evapotranspiration: Guidelines for Computing Crop Water Requirements*; FAO Irrigation and Drainage Paper 56; Food and Agriculture Organization of the United Nations: Rome, Italy, 1998.
16. Salam, R.; Islam, A.R.M.T. Potential of RT, bagging and RS ensemble learning algorithms for reference evapotranspiration prediction using climatic data-limited humid region in Bangladesh. *J. Hydrol.* **2020**, *590*, 125241. [CrossRef]
17. Ghiat, I.; Mackey, H.R.; Al-Ansari, T. A Review of Evapotranspiration Measurement Models, Techniques and Methods for Open and Closed Agricultural Field Applications. *Water* **2021**, *13*, 2523. [CrossRef]
18. Goyal, P.; Kumar, S.; Sharda, R. A review of the Artificial Intelligence (AI) based techniques for estimating reference evapotranspiration: Current trends and future perspectives. *Comput. Electron. Agric.* **2023**, *209*, 107836. [CrossRef]
19. Brekel, J.; Thorp, K.R.; DeJonge, K.C.; Trout, T.J. Version 1.1.0—pyfao56: FAO-56 evapotranspiration in Python. *SoftwareX* **2023**, *22*, 101336. [CrossRef]
20. Aghelpour, P.; Norooz-Valashedi, R. Predicting daily reference evapotranspiration rates in a humid region, comparison of seven various data-based predictor models. *Stoch. Environ. Res. Risk Assess.* **2022**, *36*, 4133–4155. [CrossRef]
21. Chia, M.Y.; Huang, Y.F.; Koo, C.H.; Fung, K.F. Recent Advances in Evapotranspiration Estimation Using Artificial Intelligence Approaches with a Focus on Hybridization Techniques—A Review. *Agronomy* **2020**, *10*, 101. [CrossRef]
22. Jing, W.; Yaseen, Z.M.; Shahid, S.; Saggi, M.K.; Tao, H.; Kisi, O.; Salih, S.Q.; Al-Ansari, N.; Chau, K.W. Implementation of evolutionary computing models for reference evapotranspiration modeling: Short review, assessment and possible future research directions. *Eng. Appl. Comput. Fluid Mech.* **2019**, *13*, 811–823. [CrossRef]
23. Ma, N.; Szilagyi, J.; Zhang, Y. Calibration-Free Complementary Relationship Estimates Terrestrial Evapotranspiration Globally. *Water Resour. Res.* **2021**, *57*, e2021WR029691. [CrossRef]
24. McMahon, T.A.; Peel, M.C.; Lowe, L.; Srikanthan, R.; McVicar, T.R. Estimating actual, potential, reference crop and pan evaporation using standard meteorological data: A pragmatic synthesis. *Hydrol. Earth Syst. Sci.* **2013**, *17*, 1331–1363. [CrossRef]
25. Liu, J.; Hertel, T.W.; Lammers, R.B.; Prusevich, A.; Baldos, U.L.C.; Grogan, D.S.; Frolking, S. Achieving sustainable irrigation water withdrawals: Global impacts on food security and land use. *Environ. Res. Lett.* **2017**, *12*, 104009. [CrossRef]
26. Paek, J.; Pollanen, M.; Abdella, K. A Stochastic Weather Model for Drought Derivatives in Arid Regions: A Case Study in Qatar. *Mathematics* **2023**, *11*, 1628. [CrossRef]
27. Chidzalo, P.; Ngare, P.O.; Mung'atu, J.K. Pricing weather derivatives under a tri-variate stochastic model. *Sci. Afr.* **2023**, *21*, e01768. [CrossRef]
28. Oliveira, R.P.D.; Achcar, J.A.; Chen, C.; Rodrigues, E.R. Non-homogeneous Poisson and linear regression models as approaches to study time series with change-points. *Commun. Stat. Case Stud. Data Anal. Appl.* **2022**, *8*, 331–353. [CrossRef]
29. Elmer, J.; Coppler, P.J.; Jones, B.L.; Nagin, D.S.; Callaway, C.W.; on behalf of University of Pittsburgh Post-Cardiac Arrest Service. Bayesian Outcome Prediction After Resuscitation From Cardiac Arrest. *Neurology* **2022**, *99*, e1113–e1121. [CrossRef] [PubMed]
30. Shedbalkar, K.H.; More, D.S. Bayesian Regression for Solar Power Forecasting. In Proceedings of the 2022 2nd International Conference on Artificial Intelligence and Signal Processing (AISP), Vijayawada, India, 12–14 Feburary 2022; pp. 1–4. [CrossRef]
31. Koller, D.; Friedman, N. *Probabilistic Graphical Models: Principles and Techniques*; MIT Press: Cambridge, MA, USA, 2009.
32. Hui, E.; Stafford, R.; Matthews, I.M.; Smith, V.A. Bayesian networks as a novel tool to enhance interpretability and predictive power of ecological models. *Ecol. Inform.* **2022**, *68*, 101539. [CrossRef]
33. Ojha, R.; Ghadge, A.; Tiwari, M.K.; Bititci, U.S. Bayesian network modelling for supply chain risk propagation. *Int. J. Prod. Res.* **2018**, *56*, 5795–5819. [CrossRef]
34. Li, H.; Guedes Soares, C.; Huang, H.Z. Reliability analysis of a floating offshore wind turbine using Bayesian Networks. *Ocean. Eng.* **2020**, *217*, 107827. [CrossRef]
35. Neapolitan, R. *Learning Bayesian Networks*; Pearson: Upper Saddle River, NJ, USA, 2003. [CrossRef]
36. Hargreaves, G.H.; Samani, Z.A. Reference crop evapotranspiration from temperature. *Appl. Eng. Agric.* **1985**, *1*, 96–99. [CrossRef]
37. Linacre, E.T. A simple formula for estimating evaporation rates in various climates, using temperature data alone. *Agric. Meteorol.* **1977**, *18*, 409–424. [CrossRef]
38. Turc, L. Estimation of irrigation water requirements, potential evapotranspiration: A simple climatic formula evolved up to date. *Ann. Agron.* **1961**, *12*, 13–49.
39. Benavides, J.G.; Lopez, D. Formula para el calculo de la evapotranspiracion potencial adaptada al tropico (15° N–15° S). *Agron. Trop.* **1970**, *20*, 335–345.
40. Cruz, T.A.C.D.; Marques, P.A.A. Low-cost irrigation management system: Improving data confidence through artificial intelligence. *Eng. AgríCola* **2023**, *43*, e20210164. [CrossRef]
41. Silva-Júnior, R.O.; Souza-Filho, P.W.M.; Salomão, G.N.; Tavares, A.L.; Santos, J.F.; Santos, D.C.; Dias, L.C.; Silva, M.S.; Melo, A.M.Q.; Costa, C.E.A.d.S.; et al. Response of Water Balance Components to Changes in Soil Use and Vegetation Cover Over Three Decades in the Eastern Amazon. *Front. Water* **2021**, *3, 749507*. [CrossRef]
42. Raghunath, H.M. *Hydrology: Principles, Analysis and Design*, 2nd ed.; New Age International (P) Ltd.: New Delhi, India, 2006.
43. Mutti, P.R.; Dubreuil, V.; Bezerra, B.G.; Arvor, D.; Funatsu, B.M.; Santos E Silva, C.M. Long-term meteorological drought characterization in the São Francisco watershed, Brazil: A climatic water balance approach. *Int. J. Climatol.* **2022**, *42*, 8162–8183. [CrossRef]

44. Saha, A.; Sekharan, S.; Manna, U. Evaluation of Capacitance Sensor for Suction Measurement in Silty Clay Loam. *Geotech. Geol. Eng.* **2020**, *38*, 4319–4331. [CrossRef]
45. Van Genuchten, M. A Closed-form Equation for Predicting the Hydraulic Conductivity of Unsaturated Soils1. *Soil Sci. Soc. Am. J.* **1980**, *44*, 892–898. [CrossRef]
46. McElreath, R. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2020.
47. Draper, N.R.; Smith, H. *Applied Regression Analysis*, 2nd ed.; Wiley Series in Probability and Mathematical Statistics; Wiley: New York, NY, USA, 1981.
48. Gelman, A.; Hill, J.; Vehtari, A. *Regression and Other Stories*; Cambridge University Press: Cambridge, UK, 2020.
49. Doucet, A.; Wang, X. Monte Carlo methods for signal processing: A review in the statistical signal processing context. *IEEE Signal Process. Mag.* **2005**, *22*, 152–170. [CrossRef]
50. Cox, L.A., Jr. Modernizing the Bradford Hill criteria for assessing causal relationships in observational data. *Crit. Rev. Toxicol.* **2018**, *48*, 682–712. [CrossRef] [PubMed]
51. Pearl, J.; Mackenzie, D. *The Book of Why: The New Science of Cause and Effect*, 1st ed.; Basic Books: New York, NY, USA, 2020.
52. Gross, T.J.; Bessani, M.; Darwin Junior, W.; Araújo, R.B.; Vale, F.A.C.; Maciel, C.D. An analytical threshold for combining Bayesian Networks. *Knowl.-Based Syst.* **2019**, *175*, 36–49. [CrossRef]
53. Moura, A.D.; Lucas, E.W.M.; Rodrigues, J.E.; de Rezende, J.M. Nota Técnica No. 001/2011/SEGER/LAIME/CSC/INMET. 2019. Available online: http://www.cemtec.ms.gov.br/wp-content/uploads/2019/02/Nota_Tecnica-Rede_estacoes_INMET.pdf (accessed on 31 July 2023).
54. Santos, H.G.d.; Jacomine, P.K.T.; Anjos, L.H.C.d.; Oliveira, V.A.d.; Lumbreras, J.F.; Coelho, M.R.; Almeida, J.A.d.; Araujo Filho, J.C.d.; Oliveira, J.B.d.; Cunha, T.J.F. *Brazilian Soil Classification System*; Embrapa: Brasília, Brazil, 2018.
55. IUSS Working Group WRB *World Reference Base for Soil Resources. International Soil Classification System for Naming Soils and Creating Legends for Soil Maps*, 4th ed.; International Union of Soil Sciences (IUSS): Vienna, Austria, 2022.
56. Box, G.E.P.; Cox, D.R. An Analysis of Transformations. *J. R. Stat. Soc. Ser. B (Methodol.)* **1964**, *26*, 211–243. [CrossRef]
57. Goudie, R.J.B.; Turner, R.M.; De Angelis, D.; Thomas, A. MultiBUGS: A Parallel Implementation of the BUGS Modeling Framework for Faster Bayesian Inference. *J. Stat. Softw.* **2020**, *95*, 1–20. [CrossRef] [PubMed]
58. Cooper, G.F.; Herskovits, E. A Bayesian method for the induction of probabilistic networks from data. *Mach. Learn.* **1992**, *9*, 309–347. [CrossRef]
59. Ribeiro, V.P.; Pereira, B.R., Jr.; Maciel, C.D.; Balestieri, J.A.P. An Improved Bayesian Network Super-Structure Evaluation using Physarum polycephalum Bio-inspiration. In Proceedings of the XXIV Brazilian Congress on Automation, Fortaleza, Brazil, 16–19 October 2022.
60. Buntine, W. Theory Refinement on Bayesian Networks. In Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence, Los Angeles, CA, USA, 13–15 July 1991; UAI'91, pp. 52–60.