MSc. Thesis Civil Engineering

Rapid assessment tool to quantify the spatial influence of surfaces on heat stress

Name student:

P.J.P. (Paul) Koot

Student ID number:

4210727

TU Delft

24th June 2021

Estimating the Physical Equivalent Temperature in urban regions using dependence modelling

by

P.J.P. Koot

Student number:	4210727	
Thesis committee:	Dr. ir. O. Morales Nápoles,	TU Delft, supervisor
	Dr. ir. J. A. A. Antolínez,	TU Delft
	Ir. G. A. Torres Alves,	TU Delft
	Dr. ir. E. Aparicio Medrano,	Nelen & Schuurmans
	I. Lokhorst MSc,	Nelen & Schuurmans

Preface

This report is the product of my MSc. thesis on the application of Non-Parametric Bayesian Networks to assess heat stress. This thesis has been written as the final product for completing my Master of Science in Civil Engineering at Delft University of Technology. The research has been conducted in collaboration with Nelen & Schuurmans. The subject of my research comes from a interest in two sectors, those of Civil Engineering and Econometrics. With this thesis I aimed to apply my knowledge on statistics and data analysis on a topic related to Civil Engineering.

I would like to thank the members of my graduation committee. First of all, Oswaldo, who guided me through the process of this research as chairman of my committee. Being an expert on NPBNs, he was always willing to answer questions, to make suggestions and to listen to my own ideas and plans. Next, Evelyn, who helped me with feedback on a regular base, in order to keep the outcomes of the research useful for both the academic world and the industry. She could also give me good advice regarding the physical aspects of heat stress. I would also like to thank Ivar, José and Gina. Ivar could give me advice based on his own experience with data visualisation, while José helped me with his knowledge on Nature Based Solutions. Gina advised me regarding the model setup, with her experience on NPBNs. They all managed to give good guidance and feedback without restricting me in following my own curiosity.

Besides, I am grateful to the other people at Nelen & Schuurmans who helped me out with all kind of questions during my time there.

Finally, I would like to thank my family for supporting me during this research. Especially I would like to mention Dorine, who supported me during the process of writing this thesis and who was always available to give feedback on my work, despite working on a PhD research herself.

Contents

List of Symbols

μ	Mean
Φ	Relative humidity
ϕ	Solar elevation angle
σ	Population standard deviation
σ_b	Stefan Boltzmann constant
τ_a	Atmospheric transmissivity
B_b	Bowen ratio
$daily\ rate$	Fraction indicating how strong the effect of the UHI is at a certain time of the day
$F_{veg,UHI}$	Derived from vegetation dummy F_{veg}, used in UHI effect
F_{veg}	Vegetation dummy, indicating either vegetation or not
H	Height
$land\ use$	Land use dummy, indicating either green space or not
$NDVI$	Normalised Difference Vegetation Index
NIR	Near-infrared bandwidth
PET_{shade}	Physical Equivalent Temperature in the shadow and at night
PET_{sun}	Physical Equivalent Temperature of a surface exposed to direct sunlight
Q_d	Diffuse radiation
Q_{gl}	Global radiation
R	Red bandwidth
S	Average hourly global radiation at the reference meteorological station
$S_{vf,UHI}$	Adapted Sky view factor used in UHI effect
S_{vf}	Sky view factor
T_a	Actual temperature, including the UHI effect

T_w	Wet bulb temperature
T_{max}	Maximum temperature measured on a day at the reference meteorological station
T_{min}	Minimum temperature measured on a day at the reference meteorological station
$T_{station}$	Temperature at the reference meteorological station
U	Average wind speed at 10 m height on a day at the reference meteorological station
$u_{1,2}$	Wind speed at 1.2 m height
UHI_{max}	Maximum urban heat island effect
W	Width

List of Acronyms

BBN	Bayesian belief network
BN	Bayesian network
BwN	Building with nature
CBA	Cost-benefit analysis
CDF	Cumulative distribution function
CLC	CORINE land cover
CvM	Cramer-von Mises
DAG	Directed acyclic graph
DEM	Digital elevation model
GDP	Gross domestic product
GFCF	Gross fixed capital formation
GIS	Geographic information system
GVA	Gross value added
HBN	Hybrid Bayesian network
IDF	Intensity-duration-frequency
MAE	Mean Absolute Error
MCMC	Markov chain Monte Carlo
MN	Markov network
MTE	Mixtures of truncated exponentials
NBS	Nature-based solution
NPBN	Non-Parametric Bayesian Network
NUTS	Nomenclature of Territorial Units for Statistics
PET	Physiological Equivalent Temperature

PIM	Perpetual inventory method
RCM	Regional climate model
RIVM	Rijksinstituut voor Volksgezondheid en Milieu
SRTM	Shuttle Radar Topography Mission
TNO	Nederlandse Organisatie voor toegepast-natuurwetenschappelijk onderzoek
UHI	Urban heat island
WB	World Bank

List of Figures

List of Tables

Abstract

Climate change causes cities to deal with increased temperatures and more frequent weather extremes. Heat waves will occur more often, becoming a more prevalent issue in especially urban areas. The quantification of heat stress is a first step to define mitigation measures. For that purpose, a standardised method to assess the spatial influence of surfaces on the Physiological Equivalent Temperature (PET) was developed. This study aims to reshape this model into a statistical dependence model which is more flexible regarding missing data. To this end, we used a Non-Parametric Bayesian Network (NBPN). We created a model driven by both data and expert knowledge, that is capable of dealing with input data layers with a grid resolution up to 20 m. Results show that training the model with only 20 sample points did not affect the performance considerably, compared to using 2,000 data points. Inclusion of a layer with sky view factor mainly improves the estimation of observations in the tails of the distribution. The model predicts the PET with a Mean Absolute Error (MAE) of 1 to 2 °C, dealing adequately with missing data layers. With this limited amount of necessary input, the NPBN in our study helps in standardising the assessment of heat stress outside the borders of the Netherlands. Also, our model offers a framework to make a first assessment regarding the effect of NBSs on heat stress.

Summary

Climate change causes cities to deal with increased temperatures and more frequent weather extremes. Heat waves will occur more often, becoming a more prevalent issue in especially urban areas. In these areas the increasing temperatures combined with the Urban Heat Island effect will cause heat stress more frequently. Quantification of heat stress is a first step to come to mitigation measures. De Nijs et al. (2019) developed a standardised method to assess the spatial influence of surfaces on the Physiological Equivalent Temperature (PET). Nelen & Schuurmans (2020) built a data driven tool around this standardisation, that uses multiple 1 m resolution input data layers to estimate the PET both during sunlight exposure and shadow conditions.

We used a statistical dependence model for this same purpose, but our model is flexible for missing data layers and is capable of dealing with coarser data input (up to 20 m resolution). This model, a Non-Parametric Bayesian Network (NPBN), therefore allows the use of non-commercially available data sources to estimate heat stress in urban areas. Using a NPBN we created a model driven by both data and expert knowledge. Also, not only a PET estimate but also the uncertainty of the model can be estimated. The model was built with data from surfaces in Wageningen and tested on surfaces in Rotterdam, both cities in the Netherlands. The heat stress was quantified using data of an one-in-thousand summer day in the Netherlands. To this end, we developed an open-source Python package to build NPBNs, called BANSHEE-y.

With this study, we show two main findings. Firstly, we show that a representative training sample is key in obtaining a NPBN with good theoretical and prediction performance. The model performed with a Mean Absolute Error (MAE) for PET_{sun} and PET_{shade} of respectively 1.6 °C and 1.2 °C. We found that the number of observations used in this training sample was of minor importance. Using only 20 sample points resulted in the same performance as using 2,000 sample points. Secondly, we found that our NPBN can handle missing data layers up to a point in which only $NDVI$ and $AHN2$ layers are needed during prediction of expected heat stress in a new city. Inclusion of a layer with sky view factor mainly improves the estimation of observations in the tails of the distribution. The network performs better at coarser grid resolutions (up to 20 m), at the cost of the level of detail in the output. With this limited amount of necessary input, the NPBN in our study can help in standardising the assessment of heat stress outside the borders of the Netherlands, since a standardised method to assess heat stress is currently not globally available. Also, with this model a first assessment can be made regarding the effect of NBSs, which influence the vegetation index $NDVI$, on urban heat stress.

1 Introduction

As a result of climate change society will have to cope with both increased temperatures and more frequent weather extremes (IPCC, 2013). Heat waves will occur more often with changing of the climate. This leads to more heat stress in both rural and urban areas. The Nederlandse Organisatie voor toegepast-natuurwetenschappelijk onderzoek (TNO) describes heat stress as temperatures above a certain threshold, that can lead to a lower thermal comfort, sleep deprivation, behavioural changes (more aggression), and less productivity. However, serious heat-related diseases can occur as well. For example skin rashes, cramp, exhaustion, strokes, kidney failure and troubles with breathing. Heat stress can even lead to death (TNO, 2021). The development of heat stress expresses itself to an even greater extent in urban regions, due to the Urban heat island (UHI) effect. We elaborate on this in Section 2.

As the potential consequences of heat stress are much larger in urban areas due to a larger population density, both industries and the academic world are focusing on this phenomenon and exploring solutions to this increasing heat stress. Hintz et al. (2017) conduct a systematic literature review exploring the transferability of solutions to cope with urban heat stress. They find that most solutions focus on 'green and blue infrastructure'. These solutions in the form of green and blue infrastructure are also called Nature-based solutions (NBSs). To combat heat stress in urban regions using NBSs, one should be able to quantify their impact. Quantification is relevant in order to investigate the impact of green and blue solutions compared to that of grey solutions. NBSs, like wadis, green roofs, green corridors and permeable pavements, have in common the geographical scale to which they are applied. Although cities differ in their layout globally, one property they share is the lack of space. Especially in city centres, only small areas can be used for mitigation purposes. Small NBSs can be effective too, but it is important to choose the right locations to apply them. To get enough support to implement NBSs in these locations where ground is often a scarce commodity, it is key to quantify the risk reduction that could come with such a solution, not in the last place to assess the financial viability. To come to a quantification of the effects of mitigation measures, a standardisation of assessing heat stress is necessary.

De Nijs et al. (2019) developed such a standardisation and Nelen & Schuurmans (2020) built a data driven tool around this standardisation. This tool aims to indicate the urban areas within a city that experience the most heat stress. However, compared to the classical physical models, an obstacle arises for data driven models, in the form of data availability. In the mentioned heat stress tool detailed data layers at a 1 m resolution are used. Although this kind of data is available in the Netherlands, a tool that could be applied globally is currently not openly available (Appendix A). One of the main reasons is the absence of high res-

olution data or even the complete absence of certain data layers which are essential for the model. Another difficulty encountered in predictive modelling, is the risk of creating a 'black box' model. As data replaces physics as the starting point of the model, one should be careful when interpreting results, especially for new locations that inherit different mechanisms from a physics point of view. This calls for a model that can deal properly with lower-resolution input layers and complete missing data layers, while still inheriting the physical aspects of the phenomenon.

Our study aims to fill this knowledge gap. The heat stress tool currently used by Nelen & Schuurmans is used as a starting point. This tool uses data and regression models on this data. We investigate a method to adapt this model to deal with data of coarser resolution and to deal with incomplete data, in particular missing data layers. The method that is used to accomplish these goals, is found in the research field of Non-Parametric Bayesian Networks (NPBNs). Both the development of and applications of NPBNs are treated in Section 2. This leads to the main research question:

To what extent can a NPBN adequately represent the joint distribution of environmental variables to assess heat stress?

This research question involves the following sub questions:

1. To what extent is the theoretical framework of NPBNs suitable to estimate heat stress?
2. How does a NPBN perform estimating heat stress in a new city?
3. How does input resolution influence the performance of a NPBN estimating heat stress?
4. How does a missing data layer influence the performance of a NPBN estimating heat stress?
5. How does the availability (number) of training data points influence the performance of a NPBN estimating heat stress?
6. To what extent can we deduce the NDVI value at a location, given its current PET value?

2 Literature

As mentioned in Section 1, climate change causes society to have to cope with both increased temperatures and more frequent weather extremes (IPCC, 2013). On top of these more frequent weather extremes, a second effect causes even more heat stress. This phenomenon is called a UHI (Koopmans, Ronda et al., 2018 and Oke, 1982) and specifically occurs within cities. These heat islands develop due to several reasons, amongst which a reduced evaporation and dark surfaces of urban building materials which absorb more solar radiation (Bornstein, 1968). This larger heat storage in urban regions during daytime causes the city to cool down less quickly after sunset, causing the heat island to persist during the night. The UHI is measured in temperature difference between urban and rural area and can obtain values of around 6 °C in Europe (Steeneveld et al., 2011). Important variables that have influence on this phenomenon can be found in Section 3.1. Urban regions will have to deal with this effect more frequently. In order to mitigate problems that may arise with increasing heat stress in the coming years, a is quantification of heat stress was necessary.

Several studies introduced ways to quantify heat stress in urban regions. Urban heat islands were described already by Oke (1982). Since then, multiple researchers have studied different approaches to assess this UHI effect. Especially the last decade multiple models were introduced to quantify this. H. Li, Zhou, X. Li et al. (2018) quantified surface urban heat island intensity in Berlin using the relationship between land surface temperature and impervious surface areas. X. Li et al. (2017) investigated "the relationship between surface UHI and urban area size in the climate and ecological context, and its spatial and temporal variations, based on a panel analysis of about 5000 urban areas of 10 km^2 or larger, in the conterminous U.S." H. Li, Zhou, Wang et al. (2019) combined the Weather Research and Forecasting Model with the Urban Canopy Module to assess the urban heat island effect. H. Li, Zhou, Jia et al. (2021) assessed the effect of urbanisation on the urban heat island effect. To this end, they used an annual temperate cycle model and applied it on an annual surface UHI series from 2003 to 2018 in the Jing-Jin-Ji region in China. Besides these different models to assess the urban heat island effect, different studies have investigated the consequences of heat waves. Mayrhuber et al. (2018) conducted a scoping review on the vulnerability to heatwaves and what implications this has for public health interventions. Jänicke et al. (2018) were one of the first to combine air temperature with vulnerability and risk, to come to a comprehensive heat-stress impact assessment. They therefore quantified heat-stress hazard (air temperature), vulnerability (heat vulnerability index and age score), and risk (heat-related mortality) on the district scale in Seoul, Korea. Arifwidodo and Chandrasiri (2020) also aimed to link heat stress directly to health and well-being outcomes. They therefore aimed "to (1) understand the determinants of heat stress, especially the roles of the urban environment in exacerbating the heat stress, and (2) to explore the effects of heat stress to human health using self-reported health

assessment." A cross-sectional study was done using a survey questionnaire from 505 respondents living in the urban area of Bangkok, Thailand. Two main observations can be drawn from these studies. First of all, in the last few years multiple methods to quantify heat stress and to assess the implications on health have been developed globally. Secondly, a lot of studies quantify heat stress by focusing on the urban heat island effect, while a second group of studies file multiple variables under the term 'heat stress', sometimes including socioeconomic characteristics as well (Arifwidodo and Chandrasiri, 2020).

Within the Netherlands, in 2018 it became clear that attempts to quantify heat stress led to a large diversity of heat stress maps. This diversity became prevalent in the different units of measure to quantify heat stress and different spatial scales across the heat stress maps, leading to confusion among stakeholders (Koopmans, Droste et al., 2020). De Nijs et al. (2019) therefore brought together a group of experts to come to a standardised way of measuring heat stress, using the Physiological Equivalent Temperature (Physiological Equivalent Temperature (PET)). The PET is based on the human energy system, making it a physical unit of measure and to a much lesser extent an empirical measure (Koopmans, Droste et al., 2020). Also they decided that the most relevant time slot to evaluate this unit of measure is between 12:00 and 18:00 local time.

Based on these recommendations, the Rijksinstituut voor Volksgezondheid en Milieu (RIVM) came up with a methodology to quantify heat stress in the Netherlands (De Nijs et al., 2019). The specific purpose of this methodology was to create a standard test for mapping vulnerability for heat in the Netherlands. In this method the apparent temperature is represented as the Physiological Equivalent Temperature (PET). Pogačar et al. (2018) showed an indirect link exists between the PET and human health, which gives reason to map this PET. The PET represents the heat stress experienced by a 35-year-old, 1.75-meter tall male, weighing 75 kilo. Also a set clothing factor is incorporated, as well as a standardised level of effort which is equivalent to walking 4 km/h (De Nijs et al., 2019). The methodology takes into account air temperature, wind speed, direct and indirect radiation and air moisture. Next to these meteorological conditions, the method takes account of geographically dependent conditions. Land-use, vegetation and the amount of shadow is incorporated in the model. Furthermore, the model takes into account combinations of both of these types of variables, e.g. by incorporating the UHI effect. This model is elaborated on in Section 3.2

Besides the work done to map heat stress, efforts have been made to also counteract and mitigate the effects of this phenomenon. In this respect multiple mitigation measures fall within the field of nature-based solutions (NBSs). An elaborated definition of NBSs is given by the European Commission (2020): "Nature-based solutions to societal challenges are solutions that are inspired and supported by nature, which are cost-effective, simultaneously provide environmental, social and economic benefits and help build resilience.

Such solutions bring more, and more diverse, nature and natural features and processes into cities, landscapes and seascapes, through locally adapted, resource-efficient and systemic interventions. Nature-based solutions must benefit biodiversity and support the delivery of a range of ecosystem services."

Although the term NBS appeared in 2008, Ruangpan, Vojinovic, Sabatino et al. (2020) identified eight different terms specifically used in the field of hydro-meteorological risk reduction, of which some have been used since 1977. The idea of NBSs is to use the natural forces present and let it grow (Koningsveld and Slinger, 2015). Besides waves, tidal movements, river discharges and rain, this also includes dune-vegetation interaction and complex interactions between marine organisms and sand or silt. The key is not to start from a certain design concept focusing on the primary function, but to start from the natural system instead. This means that besides forcing, one should also consider the secondary functions and services of nature. From an ecological perspective the primary goals of this concept, also called Building with nature (BwN), is conservation and restoration. This means that there is more focus on the external conditions that can be influenced, in order for the conditions to be optimal for the development of healthy, biodiverse and resilient ecosystems (Wesenbeeck and Slinger, 2015).

In the context of civil engineering, the primary function of a NBS is to prevent drought, heat stress, or flood. The latter can be split up into different types of flood, namely fluvial, pluvial, coastal, flash and groundwater flood (Alves, Gersonius, Sanchez et al., 2018). Sometimes multiple primary functions can be addressed by one NBS, such as green roofs, which counteract both heat stress and pluvial flood. Several secondary functions, in literature also often called co-benefits, are water savings (which influences groundwater recharge), water quality, energy savings, air quality improvement, carbon sequestrations (Alves, Gersonius, Kapelan et al., 2019). Other co-benefits, like aesthetics and recreational uses are receiving more attention as well, especially when considering NBSs in urban environments.

NBSs are receiving more attention as an alternative for the common grey solutions. As the common grey solutions are well known to be very effective in fulfilling primary functions in a very cost efficient way, research is done on how to monetise the co-benefits of NBSs. Alves, Gersonius, Kapelan et al. (2019) presented a method to include some of these co-benefits into a cost-benefits analysis of flood risk mitigation measures. They first used 1D-1D models to estimate the expected annual damage regarding the primary function of the measure and then added to that an annual monetary value for each co-benefit. They applied this framework in a case study in the Dutch side of Sint Maarten Island, to assess the co-benefits of NBSs with the primary function of counteracting damage done by storm water. In the case study they calculated co-benefits of green roofs, consisting of energy savings due to insulation, air quality, CO_2 sequestration and an increasing roof

longevity. Also the co-benefits of applying rain barrels were monetised: water savings, energy savings and freedom from water restrictions during drought. A final calculation was done for pervious pavements, with heat stress reduction (which boils down to energy savings) as main co-benefit. Alves, Gersonius, Kapelan et al. (2019) showed that including co-benefits can lead to changes in decision making. The research field considering NBSs mainly focuses on small-scale NBSs and the discussion of large-scale NBSs, which often offers more possibilities for habitat creation, is limited. Actually, the European Commission stated that "the review of literature to date confirms a large gap between the research efforts concerning small- and large-scale NBSs with small-scale NBSs receiving far greater attention" (European Commission, 2020). They also mention several reasons for this. One is that small-scale NBSs are more attractive for storm water management and regeneration of urban areas. Besides, they are less complex and the benefits can be observed quickly after implementation. Installation of lab tests is done more easily and finally the costs of pilot implementation, operation and maintenance is more affordable. To this list we could add the fact that small-scale NBSs are more flexible in application in urban areas with often limited space. Larger scale NBSs would imply sacrifices in terms of land use, while small-scale NBSs can often be realised by minor adaptions to current buildings or infrastructure. Also, the fact that large-scale NBSs come with co-benefits that are less easy to monetise is not attractive from an economic point of view, which might also partly explain the current research focus on small-scale NBSs.

The European Commission is not the only one that has done research on knowledge gaps concerning NBSs. Ruangpan, Vojinovic, Di Sabatino et al. (2020) endorsed the fact that natural hazards such as floods and droughts become worse as a fact of "projected changes in climate, degradation of ecosystems, population growth and urbanisation". They provided an extensive literature review of the research area concerning nature-based solutions for hydro-meteorological risk reduction. This research showed that green roofs are often studied in the current literature. Based on this literature review, several knowledge gaps and recommendations considering future research were described. A few relevant knowledge gaps were identified with regard to our research. Firstly, it was found that a framework for cost-benefit analysis of NBSs does not yet exist. Especially in the research field of small-scale NBSs like green roofs, none of the contributions that were included in the study incorporated a Cost-benefit analysis (CBA). However, CBAs are at the core of tools that support decision-making processes, as they give insight into the feasibility of a NBS. Also by creating an accurate and systematic CBA the NBS solution can be compared more easily to grey solutions. An extension of this first knowledge gap is that there is no assessment of multiple benefits of NBSs. Quantification of co-benefits has not been done until now and a framework is necessary, in which methods and tools are available to evaluate both intangible and tangible benefits. The last knowledge gap was found in the application of tools. Considering the application of new technologies and concepts, a

trade-off needs to be made between high-resolution numerical models and the accuracy of results. Also, tools to assess the multiple benefits for both small and large scale NBSs and their hybrid combinations are needed.

As this research brings together the field of heat stress and that of Bayesian Networks (BNs), the second part of this section treats the development of this statistical dependence model. The concept of BNs was firstly discussed by Pearl (1988). Bayesian Networks (BNs), or equivalently Bayesian belief networks (BBN), were developed in the context of knowledge-based systems (Van Der Gaag, 1996). The term knowledge-based systems in general covers computer systems which are used to represent (human) expert knowledge. They are applied over a wide ranges of disciplines, varying from medical applications (Kahn Jr et al., 1997) to financial decision making (C. Shenoy and P. P. Shenoy, 2000). However, as the number of applications grew, the need for more flexible models grew as well. An example of this was the growing demand for models that could cope with the uncertain nature of processes. Given a set of evidence, the probability of each hypothesis out of a set of hypotheses should be assessed. Using Bayes' Theorem (see Appendix B) as a starting point, first efforts were made in the field of the naive Bayesian approach (Van Der Gaag, 1996). However, this type of methods needed a lot of conditional probabilities to be computed ($2^N - 1$ for N hypotheses). Several assumptions were made to decrease this number of probabilities. Although the models worked well under these restrictions, they did not function well for larger or more complex domains (Van Der Gaag, 1996). Especially in the field of medical diagnostic systems a second branch of models arose. These so-called rule-based expert systems aim to capture "human expertise in terms of rules of the form if condition-then action" (Oniésko, Lucas and Druzdzel, 2001). Although Oniésko, Lucas and Druzdzel (2001) argued that both types of models have their advantages and disadvantages, there was a lot of debate especially on the mathematical correctness of these rule-based systems (Cheeseman, 1988).

To understand the development from BNs to other models, a conceptual description of this former follows. A Bayesian Network is a graphical representation of a set of random variables and their conditional dependencies. The nodes in such a network represent the random variables and the directed edges represent the conditional dependencies, the direction of the edge indicates the direction of causality. This property implies that a BN is a Directed acyclic graph (DAG). Pearl (1988) introduced BNs as being advantageous over Markov networks (MN). A Markov network is a graphical model such as the BN that describes joint probability distribution using probability tables. Nodes represent random variables as well. However, the edges in a MN are undirected, causing this kind of network to have difficulties with capturing a 'common effect' pattern (see Section C). In a MN, this can only be solved by a direct edge between two possible causes of a common effect, while intuitively speaking, dependence between both causes would only be induced whenever something is known about their common effect. The first applications focused on using discrete BNs. In

these BNs the nodes represent discrete random variables. An advantage of using BNs to model dependency, is the fact that there is no longer the need to construct the full joint density. Incorporating N binary random variables, a full joint distribution would already have $2^N - 1$ entries. This number of conditional probabilities to be estimated from the data can be significantly reduced using BNs. This is illustrated in Appendix B in more detail.

Although applications of this kind of BNs have been proven to be successful, discrete BNs come with several disadvantages. First of all, the number of conditional probabilities needed to completely specify the BN can still be very large. Every node has its own conditional probability table and such a table grows exponentially with the number of parent nodes. Therefore, a BN with N nodes, including nodes with up to k parents, lets the number of parameters grow rapidly: $N * 2^k$. Although this is an upper limit, the number of entries can become quite large easily for small networks with a small number of edges, especially when one wants to e.g. discretise a continuous distribution in several bins. To fill all those entries one either needs a lot of expert domain knowledge or a lot of input data. Besides, the restriction of only having discrete variables on its own is not suitable for every application.

Therefore, a next step in the framework of BNs was made by introducing Hybrid Bayesian Networks (HBN). Langseth et al. (2009) gave an overview of different HBNs. They focused on comparing approaches to do inference in HBNs using discretisation, mixtures of truncated exponentials (MTEs), variational methods and Markov chain Monte Carlo (MCMC) methods. This research mainly discussed inference, without purposely discussing the quantification of the model. However, Hanea, Morales-Nápoles and Ababei (2015) made a fair point by stating that striving towards an analytically exact method to do inference becomes only important after specifying the BN itself in a transparent, reliable and defensible quantification method (Hanea, Morales-Nápoles and Ababei, 2015). When looking at these HBNs, the specification of the model is done by using marginal distributions for parent nodes and conditional distributions for child nodes. For continuous nodes these methods either use discretisation or fully parametrised joint distribution functions. The problem with these kind of methods is that they still encounter the problem mentioned earlier; to quantify the model either a lot of expert knowledge is necessary or a lot of data, to fill all entries in the conditional distributions. As one of the advantages of using BNs is that in general not a lot of data is required to come to a model, this data requirement in such HBNs would lead to less interesting models from a practical point of view.

Hanea, Morales-Nápoles and Ababei (2015) therefore shifted the focus towards Non-Parametric Bayesian Networks, which were introduced by Kurowicka and Cooke (2005). NPBNs have been applied in several fields and a review of its applications was made by Hanea, Morales-Nápoles and Ababei (2015). They clas-

sified applications in the fields of "risk analysis, reliability of structures, properties of materials" (Hanea, Morales-Nápoles and Ababei, 2015). In Appendix D we elaborate on this review. They underlined the broad spectrum of disciplines in which NPBNs can lead to useful results. If we converge towards applications in the field of urban resilience against climate change, the topics of flooding and loss estimation in urban areas are subject of two studies using NPBNs. Therefore, in the final part of this section we elicit more elaborately on these studies, touching upon both the use cases and data sources used. Both of these practical aspects are important for our research.

Paprotny and Morales-Nápoles (2017) used a NPBN to estimate extreme river discharges in Europe. They assembled discharge data of 1841 river gauge stations. The measured daily discharges in these time series were converted to annual maximum discharges (Q_{AMAX}). The NPBN uses the data sources that are listed in Table 1 to estimate Q_{AMAX} (Paprotny and Morales-Nápoles, 2017).

Table 1: Datasets used by Paprotny and Morales-Nápoles (2017)

Name Dataset	Information	Coverage	Resolution	Variables derived
CCM River and Catchment Database v2.1	River network and catchments, derived from DEM	Pan Europe	100 m	Catchment area, river network density and catchment circularity
COSMO_4.8_clm17	Regional climate model (RCM)	Pan Europe	12 km	Total precipitation, snow melt, near-surface temperature and total runoff (daily)
EU-DEM complemented with SRTM data	Digital elevation model (DEM) and Shuttle Radar Topography Mission (SRTM) data	Pan Europe	25 m - 100 m	Average river slope. average catchment slope and elevation
CORINE land cover (CL), complemented with Global Land Cover 2000	Land cover	Pan Europe	100 m - 1 km	Land cover
European Soil Database v2.0	Soil data	Pan Europe	1 km	Occurrence of peat, unconsolidated and aeolian deposits, average water content and soil texture
Revised Soil Sealing 2006	Soil sealing	Pan Europe	100 m - 1 km	Soil sealing
SoilGrids1km	Grain-size structure of the soil (gravel, sand, silt, or clay)	Pan Europe	1 km	Grain-size structure of the soil

Using several performance measures (see Appendix D to compare the performance of NPBNs with other models from the literature, this NPBN outperformed all of them, while not having geographical restrictions in applicability. However, it should be noted that although this holds theoretically, in practice we encounter the problem of restricted data availability outside of Europe.

The research mentioned above used a NPBN to estimate extreme river discharges, which makes a contribution towards estimating flood hazards. Also, efforts have been made to use NPBNs in estimating potential flood losses in both rural and urban areas. Paprotny, Kreibich, Morales-Nápoles, Terefenko et al. (2020) estimated the exposure of residential assets to natural hazards in Europe. In Table 2 the data sources, that were used for this NPBN, can be found.

Table 2: Datasets used by Paprotny, Kreibich, Morales-Nápoles, Terefenko et al. (2020)

Name Dataset	Information	Coverage	Resolution	Variables derived
Building Height 2012 (Copernicus Land Monitoring Service)	Building heights	30 capitals in Europe	Building scale	Building height
GEOSTAT 2011 (Eurostat)	Population density	Pan Europe	1 km	Population density
HANZE [(1)]	Population density	Pan Europe	100 m	Population density
Urban Clusters 2011 (Eurostat)	Information on urban clusters	Pan Europe	1 km	Population in an urban cluster and distance from centre of an urban cluster
Imperviousness 2012 (Copernicus Land Monitoring Service)	Imperviousness	Pan Europe	100 m	Soil sealing
European Settlement Map 2012 (Copernicus Land Monitoring Service)	Settlements	Pan Europe	100 m	Built-up surfaces
OpenStreetMap	Building footprints	Global	Building scale	Building footprint area

[(1)] Paprotny, Morales-Nápoles and Jonkman (2018)

It should be noted that in the datasets presented above, only OpenStreetMap contains data outside Europe. In a different study Paprotny, Kreibich, Morales-Nápoles, Castellarin et al. (2020) also vulnerability was investigated, next to flood exposure (see Appendix D). What stood out from all of these applications, is the variety of data sources needed to come to a model with a good accuracy. All studies that we have elaborated on, are applied in pan-Europe. Although they have the potential to be extended to a global scope, a problem arises in the form of data availability. Most of the datasets used, are restricted to the pan-European borders. On the other hand, if datasets are available outside the pan-European borders, they often inherit a much coarser grid size. Digital Elevation Models (DEMs) are commercially available in the SPOT datasets (ESA) with a 20 m resolution and also the freely available Sentinel-1 and 2 data (ESA) have a resolution of 15 m. A third global data source, LANDSAT-9 (NASA), has multiple resolutions of 15, 30 and 100 m, depending on spectral band.

3 Methodology

In Figure 1 the workflow of this study can be found. This section starts with the data collection. Next, the four topics regarding model development are treated. We pose the current PET method used by Nelen & Schuurmans and explain how we combine the current PET method implementation with a NPBN. In doing this, the assumptions made and the limitations our proposed model has, are discussed. Then we go into more detail by showing both the qualification and the quantification of the NPBN. The implementation of the NPBN is done in BANSHEE-y, a Python version of BANSHEE (Paprotny, Morales-Nápoles et al., 2020). We introduce an overview of our simulations to answer each of the sub questions. After this, we discuss the validation of our model and we finish discussing more extensively how to different runs have different extensions to the base model to answer the sub questions.

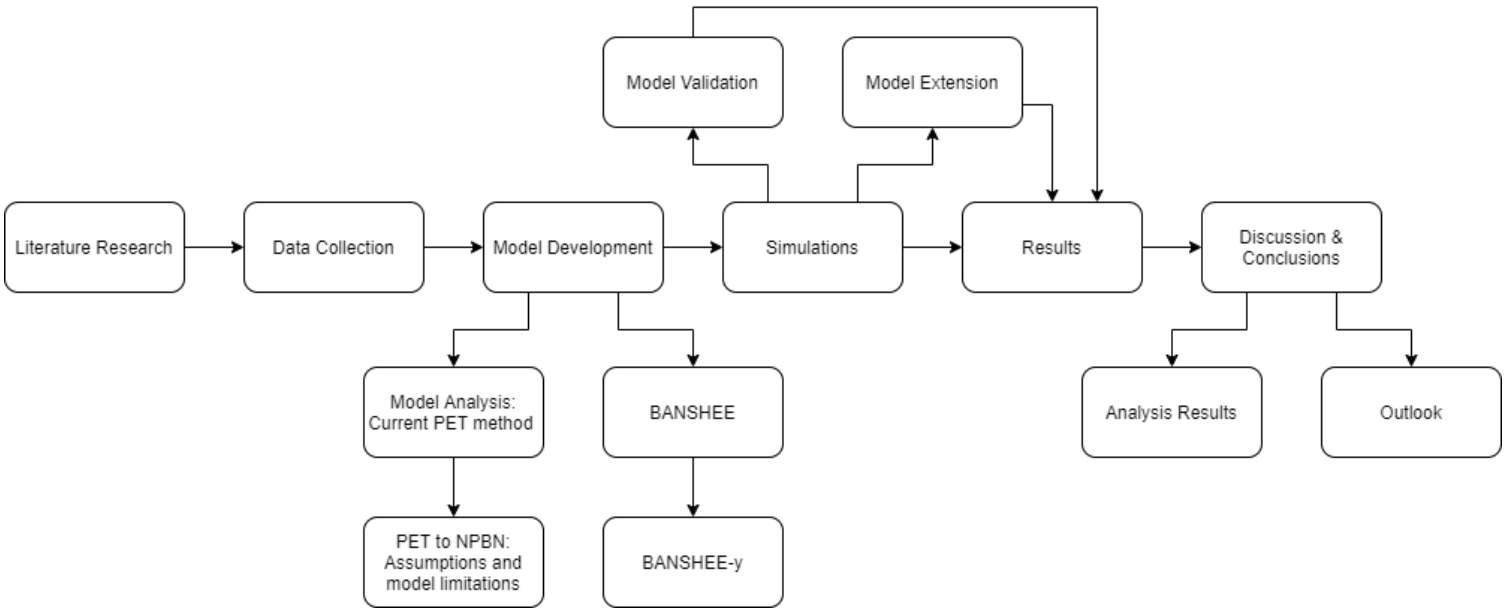


Figure 1: Workflow of our study. The different topics are treated in this thesis the same order they appear in this schematic overview

3.1 Data collection and preparation

In Table 3 the raster datasets of the variables used in our study, are summarised. Besides these raw datasets, we used the following data layers from the current PET model of Nelen & Schuurmans to train our model: T_a, F_{veg}, $F_{veg,UHI}$, $S_{vf,UHI}$ and both PET_{sun} and PET_{shade}.

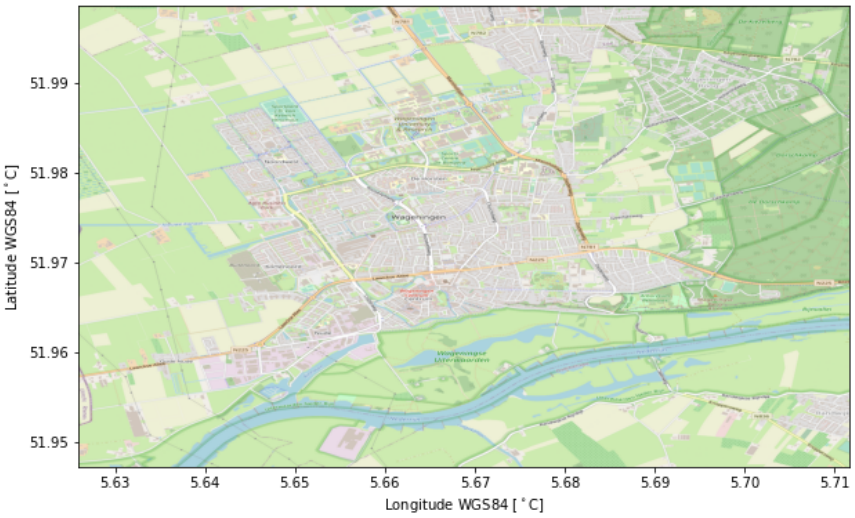
Table 3: Datasets used in this study

Name Dataset	**Information**	**Coverage**	**Resolution**	**Variables derived**
Aerial photograph (Waterschapshuis, 2018)	RED and NIR	Netherlands	1 m	$NDVI$
AHN2 (Georegister, 2012)	Raster with heights of objects	Netherlands	0.5 m	Sky view factor (S_{vf} and $S_{vf,UHI}$)
Sky View Factor of the Netherlands (KNMI, 2020)	Sky view factor	Netherlands	1 m	Sky view factor (S_{vf} and $S_{vf,UHI}$)
Land use Netherlands (Schuurmans, 2019)	Land use classification	Netherlands	0.25 m	*land use*
Basisregistratie Adressen en Gebouwen[(1)] (Kadaster, 2019a)	Buildings & Public spaces	Netherlands	0.25 m	*land use*
Basisregistratie Grootschalige Topografie[(1)] (Kadaster, 2019b)	Topography	Netherlands	0.5 m	*land use*
Basisregistratie Personen[(1)] (Rijksoverheid, 2019)	Buildings classification	Netherlands	0.25 m	*land use*
Nationaal Wegen Bestand[(1)] (Rijkswaterstaat, 2019)	Roads	Netherlands	0.25 m	*land use*
Topografie Nederland (TOP10NL)[(1)] (Kadaster, 2018)	Topography	Netherlands	2 m	*land use*

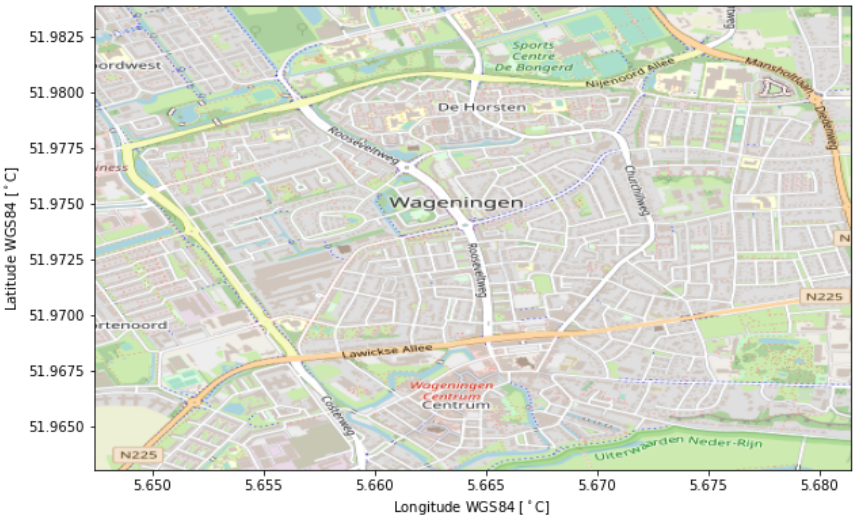
[(1)]These datasets are indirectly used as they are input to the Land use Netherlands layer, which we use in this research.

To train the model we used two areas. We started our analysis in Wageningen. The first area of interest includes both Wageningen city and the rural area around Wageningen. Next, we narrowed down our area of research to Wageningen city only. Besides training the model, in-sample validation within these study areas can be done by using the same 2,000 sample points to infer predictions on. However, to assess the applicability of our model in new Dutch cities (out-sample), we also looked into a third study area, Rotterdam. Validation was done on part of Rotterdam city. Maps of all these three areas can be found in Figure 2.

(a) Rotterdam city (out-sample)



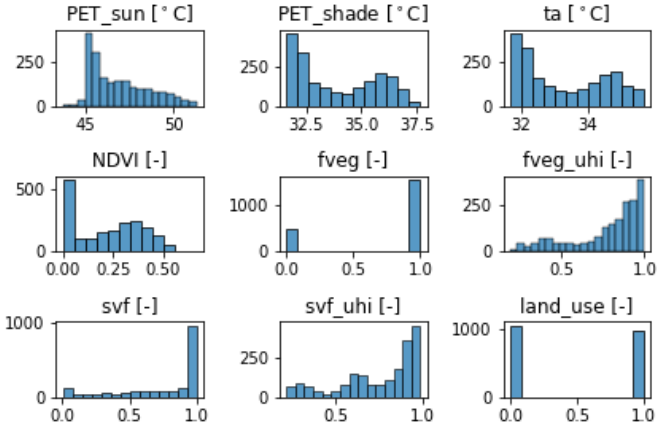
(b) Wageningen (in-sample)



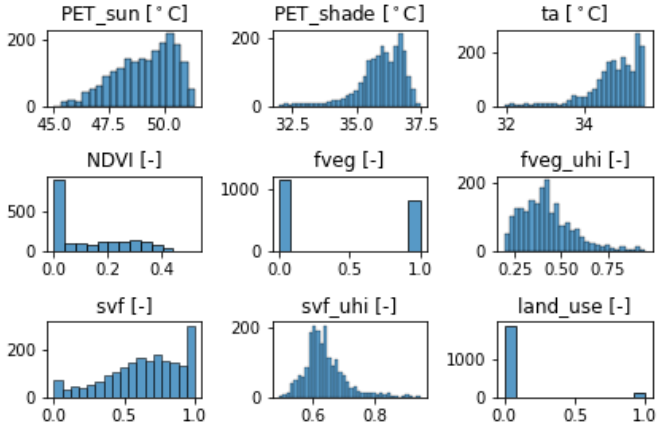
(c) Wageningen city (in-sample)

Figure 2: Areas used in the study

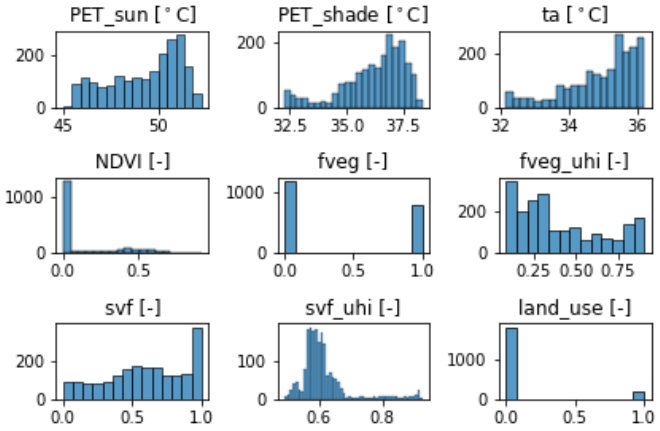
From these data training and validation is done on 2,000 sample points. The margins for all variables of all three areas can be found in Figure 3.



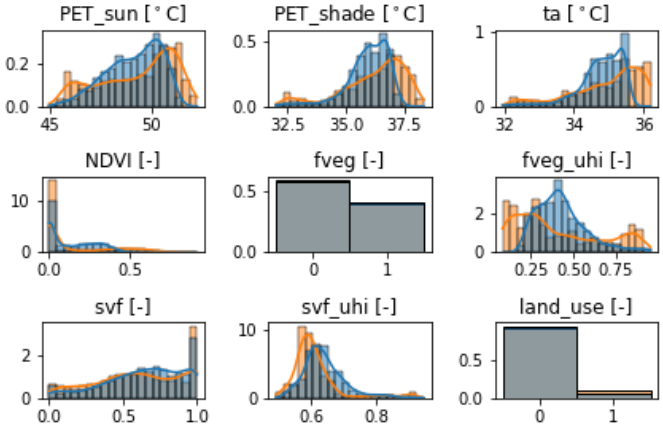
(a) Wageningen (rural and city)



(b) Wageningen city



(c) Rotterdam city



(d) Comparison distributions Wageningen city (blue) and Rotterdam city (orange)

Figure 3: Marginal distributions of all variables used in the base model

A main observation that can be made from the NPBN visualisation is the bimodal distributions in the nodes $NDVI$, $F_{veg,UHI}$, T_a and PET_{shade} in Wageningen (rural and city) (Figure 3a). Also $S_{vf,UHI}$ has a trimodal distribution. When comparing Wageningen city to Rotterdam city (Figure 3d), one can conclude that although the domain of both variables are the same, Rotterdam city has some more extreme values in the right tail for PET_{sun}, PET_{shade} and T_a. Also the distribution of $F_{veg,UHI}$ is much closer towards a normal distribution for Wageningen city.

In the study three resolution scales were used: grid cells of 1x1 m, 10x10 m and 20x20 m. In Section 3.6 the choice for these scales is elaborated on. To come to representative performance metrics, the sample was created by randomly choosing 2,000 sample points within the area of interest. Initially this was done for each area at the 20x20 m grid scale. When upsampling the grid, we chose from the set of sample points that fall within the same locations to get a fair comparison in test statistics. Starting at a 20x20 m grid, upsampling was done twice to a maximal resolution of 1x1 m, which coincides with the grid of the input layers. When upsampling from 20x20 m to 10x10 m, one of the four new grid cells which fall within the original 20x20 m area was used in the finer sample. Moving from 10x10 m to 1x1 m, again one of the four grid cells in the centre of the new 100 cells was used in the finer sample. Exceptions occurred for chosen cells that do not contain any data. When encountering this, a new cell of these four cells was randomly chosen. If all four centre cells did not contain any data, the algorithm randomly chose a cell from the 100 grid cells when upsampling from 10x10 m to 1x1 m. As we only had the four 'middle cells' to choose from when upsampling from 20x20 m to 10x10 m, the search grid was enlarged to all sides with one grid cell if none of these four cells contained data. This enlargening process was done at most five times and finally resulted in a value for all 2,000 sample point on every scale. In Table 4 can be found for each sample how many draws had to be done from a subset larger than the four centre grid cells.

Table 4: Number of grid cells (out of N=2,000) that were not selected from the middle four cells when upsampling from 20x20 m to 10x10 m and from 10x10 m to 1x1 m (first and second row respectively)

Upsampling Resolution [m]	**Wageningen**	**Wageningen city**	**Rotterdam city**
10	16	3	13
1	61	33	121

3.2 Current PET Method

In Section 2 we mentioned De Nijs et al. (2019) came with a standardised method to assess heat stress in the Netherlands. We start explaining this method, which is the foundation of the current PET model of Nelen & Schuurmans. Then, we finish this section by introducing the current PET method as implemented by them.

Using local measurements in the city of Rotterdam done by Heusinkveld et al. (2010), De Nijs et al. (2019) has come up with two regression equations to calculate the PET. The variable to be explained (the PET) is obtained by using a human energy balance model Rayman (Matzarakis, Rutz and Mayer, 2010). This Rayman model is able to model radiation in a complex environment based on a 3D street. Next, it uses the

influence of air temperature, global radiation (both diffuse and direct), thermal radiation and wind velocity as input to calculate the PET. The empirical regression equations are a simplification of the original Rayman model. This simplification is necessary as it is not easy to calculate a human energy balance for a whole city on a square meter scale. For example, radiation on itself is a variable that is influenced by specific location parameters, resulting in a lot of variability at one location with changing direction. In the Rayman model the PET was modelled for nine reference streets (De Nijs et al., 2019), based on detailed input variables. Next the regression equations below were fitted with a minimum number of variables to come to an efficient way of estimating the PET in streets on larger maps. For shadow and nocturnal conditions the following equation is obtained:

$$PET_{shade} = -12.14 + 1.25T_a - 1.47\ln(u_{1.2}) + 0.060T_w + 0.015S_{vf}Q_d + 0.0060(1 - S_{vf})\sigma_b(T_a + 273.15)^4. \quad (1)$$

For locations exposed to the sun, the following equation holds:

$$\begin{aligned} PET_{sun} = &-13.26 + 1.25T_a + 0.011Q_{gl} - 3.37\ln(u_{1.2}) + 0.078T_w + 0.0055Q_{gl}\ln(u_{1.2}) + 5.56\sin(\phi) \\ &- 0.0103Q_{gl}\ln(u_{1.2})\sin(\phi) + 0.546B_b + 1.94S_{vf}. \end{aligned} \quad (2)$$

In these equations T_a is the air temperature at 2 m (°C), $u_{1.2}$ the wind velocity at 1.2 m height (m/s), σ_b the Stefan Boltzmann constant, T_w the wet-bulb temperature, Q_d the diffuse radiation (W/m^2), S_{vf} the sky view factor, Q_{gl} the global radiation (W/m^2), ϕ the solar elevation angle and B_b the Bowen ratio. The sky view factor (Dirksen et al., 2019) represents the fraction of visible sky and ranges from 0 (closed space) to 1 (an open field). It is calculated using the relative height H of objects with respect to the point of interest and the distance W between those objects:

$$S_{vf} = \cos\left(\arctan\left[\frac{H}{0.5W}\right]\right). \quad (3)$$

The Bowen ratio is a categorical variable, which is calculated based on the vegetation layer F_{veg}:

$$B_b = \begin{cases} 3 & \text{for } F_{veg} = 0, \\ 0.4 & \text{for } F_{veg} = 1. \end{cases} \quad (4)$$

This vegetation layer F_{veg} is a ordinal variable indicating whether vegetation is present in a grid cell or not. This is determined by:

$$F_{veg} = \begin{cases} 1 & \text{for } land\ use = water, \\ 1 & \text{for } land\ use = agri, \\ 1 & \text{for } NDVI \geq 0.16, \\ 0 & \text{otherwise .} \end{cases} \tag{5}$$

where *land use* is a categorical variable indicating the main use of the land within a grid cell and $NDVI$ is the Normalised Difference Vegetation Index. $NDVI$ is a graphical indicator to express the amount of living green in an area. It is calculated by combining the red (R) and near-infrared (NIR) bands obtained from satellite imagery or aerial photography, as

$$NDVI = \frac{NIR - R}{NIR + R}. \tag{6}$$

The hourly actual temperature $T_a[h]$ is also calculated from several variables:

$$T_a[h] = T_{station} + UHI_{max} * daily\ rate[h], \tag{7}$$

in which $T_{station}$ is the temperature at the reference meteorological station, the daily rate is a fraction indicating how strong the effect of the UHI is at a certain time of the day. This daily rate is initially based on **oke** and adapted by De Nijs et al. (2019). This fraction corrects for the maximum UHI effect, which is present in the time interval of 4 hours after sunset (Landsberg, 1981 and Unger et al., 2001). This maximum UHI effect is defined as:

$$UHI_{max} = (2 - S_{vf,UHI} - F_{veg,UHI}) \sqrt[4]{\frac{S(T_{max} - T_{min})^3}{U}}. \tag{8}$$

For both the sky view factor $S_{vf,UHI}$ and the vegetation fraction $F_{veg,UHI}$ an average value is taken in the surrounding area of the grid cell of interest. If wind velocities are higher than 1.5 m/s, this area stretches 1000 m against the direction of the wind, 100 m along with the direction of the wind and 250 m to both sides orthogonal to the wind direction, leading to a total area of 1100 x 500 m. For lower wind velocities a square of 700 x 700 m around the grid cell of interest is taken. S is the average hourly global radiation in Wm^{-2}, T_{max} and T_{min} are maximum and minimum temperatures measured that day (8UTC - 7UTC next day) at the reference meteorological station and the same is done for U, the average wind velocity at 10 m height. The global radiation Q_{gl} is also measured at a reference meteorological station and the diffuse radiation Q_d is derived as:

$$Q_d = \begin{cases} Q_{gl} & \text{for } \tau_a < 0.3, \\ (1.6 - 2\tau_a)Q_{gl} & \text{for } 0.3 < \tau_a < 0.7,, \\ 0.2Q_{gl} & \text{for } \tau_a > 0.7, \end{cases} \tag{9}$$

where τ_a represents the atmospheric transmissivity, which is approximated by the formula

$$\tau_a = \frac{Q_{gl}}{1367 \sin(\phi)}. \tag{10}$$

In this equation ϕ again represents the solar elevation angle. The last value that needs to be calculated to get all variables in Equations 1 and 2, is T_w. This value is calculated according to Stull (1988):

$$\begin{aligned} T_w = & T_a \arctan\left(0.151977(\Phi + 8.313659)^{0.5}\right) + \arctan(T_a + \Phi) - \arctan(\Phi - 1.676331) \\ & + 0.00391838\Phi^{\frac{3}{2}} \arctan(0.023101\Phi) - 4.686035, \end{aligned} \tag{11}$$

in which Φ is the relative humidity in %.

3.2.1 PET implementation Nelen & Schuurmans

To implement the PET model in a tool, some assumptions and simplifications were made regarding several variables. In this section the assumptions and limitations of the current approach are touched upon.

The model of Nelen & Schuurmans (2020) focuses on spatial variation of heat stress within cities. Therefore it uses data at one point in time, 1 July 2015. Measurements at the meteorological measurement station of the KNMI in Herwijnen showed that the weather conditions that day coincides with a one in thousand summer day. This means the conditions on that day have a probability of 0.1% to occur on any day within the meteorological summer (1 June to 31 August). However, on that day it was relatively cool in the provinces of Zeeland and Friesland. To come to a first heat map that covers the Netherlands completely under the same one in thousand circumstances, the meteorological variables S, T_{max}, T_{min}, T_{NL}, $Q_{gl,NL}$ and Φ_{NL} were taken from measurements on that day at the meteorological measurement station of the KNMI in the Bilt, 30 km from Herwijnen. Also the hourly UHI rates between 10CET and 16CET were averaged to come to a daily rate. These hourly rates do not have spatial variation across the Netherlands. The average sun height angle ϕ_{NL} at 1 July 2015 was derived from taking the average value of the sun height angles between 12CET and 18CET. σ_b, the Stefan Boltzmann constant, also does not have a spatial variation.

Another assumption is no wind is present during a one in thousand summer day. There were multiple reasons for this simplification. First of all, within cities the largest PET values occur naturally at locations where the wind is reduced to a minimum. Theoretically the wind speed does not have to be zero in the whole city. However, due to the structures present in a city, calculating a detailed wind field at street level is difficult, while the theoretical formulas from the previous section, that include wind, are quite sensitive to this input

parameter. Incorporating a little bit of wind would lead to unrealistic scenarios, as was seen e.g. near the edge of forests in rural areas during model development. When adding wind to the equation, the model estimated relatively large PET values close to the edge of a forest, as a result from the combination of two effects; wind reduction close to the edge of the forest compared to the adjacent rural area and no shadow from the forest. According to both Wageningen University & Research (WUR) researchers and consultants at Nelen & Schuurmans the model without including wind effects gave much more realistic results in these areas and also in cities, where most heat stress is experienced at moments of complete absence of wind. However, one should be careful when using this model at locations where wind is more prevalent. A more windy coastal climate could cause too pessimistic PET values in coastal cities when completely neglecting wind. Nevertheless, overall this model gave good results for the Netherlands.

A third simplification was made in the way $F_{veg,UHI}$ and $S_{vf,UHI}$ were calculated. Instead of a spatially rolling mean as was described in the previous section, a less computationally expensive smoothing method was used, Gaussian blur:

$$G(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right), \tag{12}$$

where $\sigma = \frac{700m}{3}$. So the extent of the smoothing was divided by a factor 3 to come to a value of σ. As could be seen in the equations from the previous section, some variables occurred in multiple equations. In Figure 4 all variables and the routes via which they affect the PET values are visualised.

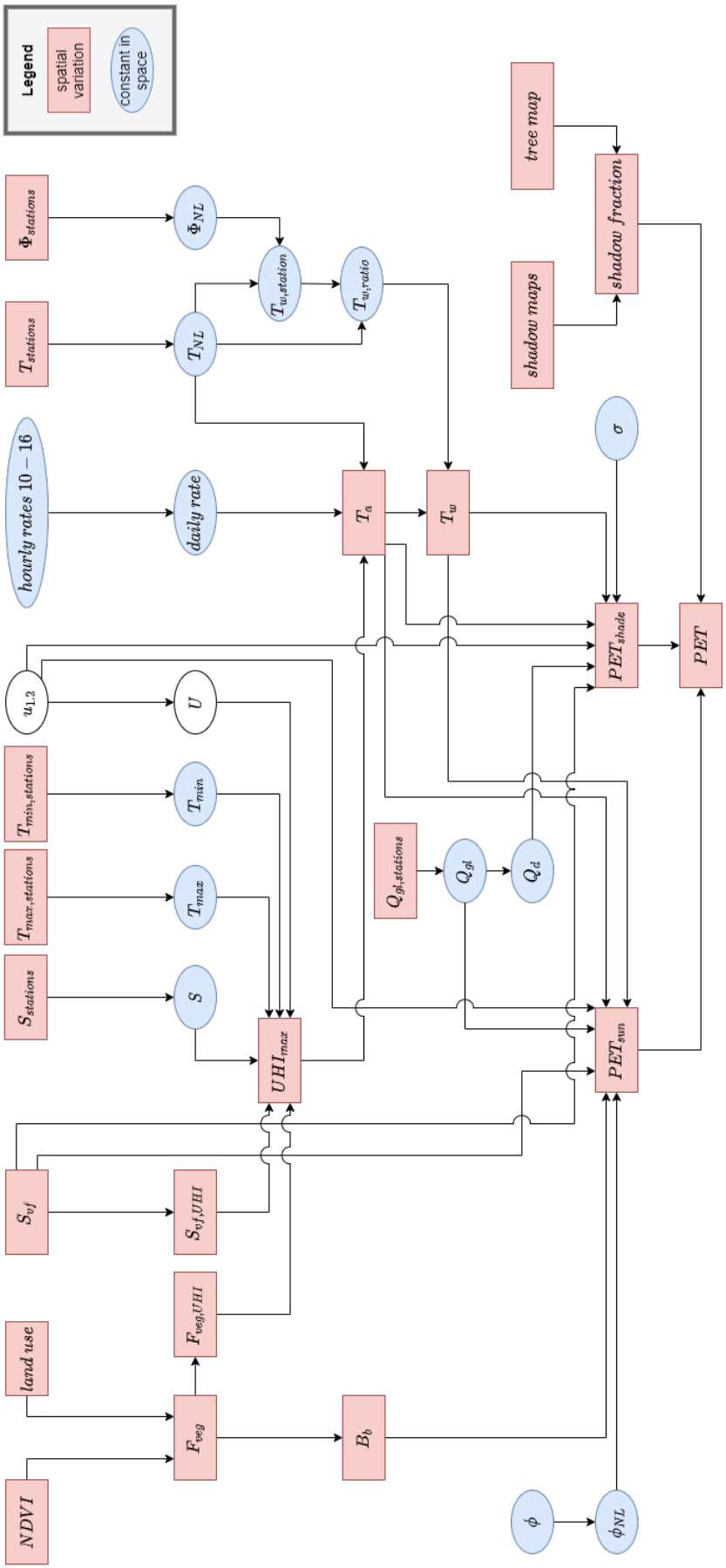


Figure 4: PET implementation Nelen & Schuurmans (2020). A distinction is made between variables with spatial variation (red squares) and variables without spatial variation (blue ellipses). The latter group was deduced from either averaging variables or by choosing values at KNMI station the Bilt.

3.3 NPBN to estimate PET

The current PET regression uses high quality input maps with a high resolution to come to PET values at a street level. However, these high resolution maps are not always available. Therefore, another approach needed to be developed to come to PET values in data scarcer regions. In Figure 5 the current method used in the Netherlands is presented and two configurations, based on using a NPBN, are illustrated below.

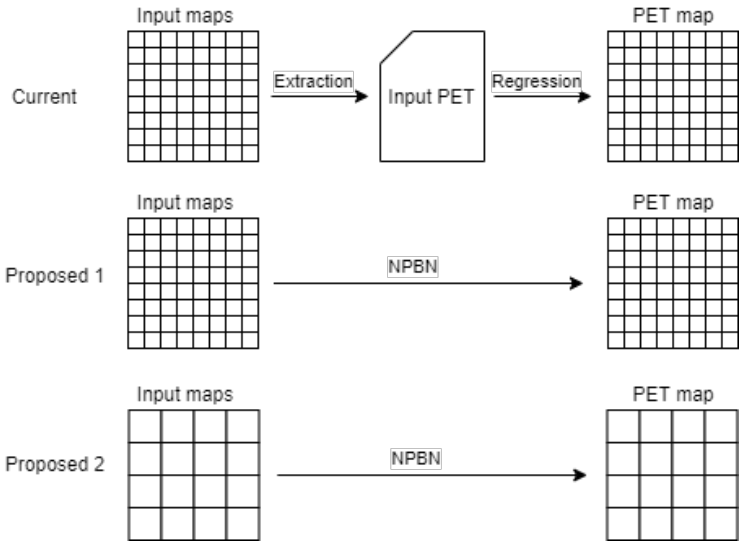


Figure 5: Schematic overview of current method and proposed research method. The NPBN fulfils the role of dependence model, directly mapping all input layers into a PET estimate. The NPBN was examined using the same input as the current model (Proposed 1), but also using less information as input (Proposed 2)

In the proposed model we started using the input and output scales corresponding to the current model. After this we experimented with using a smaller amount of input information in three ways. Coarser grids were tested to see to which point the output is accurate enough and also experiments were conducted omitting complete input data layers. The third way of reducing input data was obtained by reducing the number of sample points to train the model on gradually from 2,000 to 20. In Section 3.6 we elaborate on the experiments.

Building a NPBN consists of two steps: qualification of the network and quantification of the parameters. In the following subsections, the assumptions and limitations of the model are explained. Then, for both steps, the design choices for the case study are discussed.

3.3.1 Assumptions and limitations

To come to a model to estimate PET values, a base model was created with the data layers as being used in the current model of Nelen & Schuurmans (2020). With this model as a base case, some assumptions and limitations were also inherited. These were mentioned in Section 3.2.1. First of all, regarding the meteorological variables, our NPBN model also uses values observed on 1 July 2015 as a one in thousand summer day event for the same set of variables. Secondly, for these variables the values at the Bilt were used, leading to values which are constant in space. Thirdly, we also assumed the absence of wind. These latter two assumptions led to variables that are constant for every grid cell. As the NPBN uses rank correlation between variables to determine the dependence between them, this implies that these spatially constant variables do not add any information to the model. Therefore, we removed all variables that are spatially constant (blue in Figure 4). A limitation different from the original model is the fact that we wanted to estimate both PET_{sun} and PET_{shadow} separately. We did not combine these two values, as was done in the Nelen & Schuurmans model, to estimate the average PET value in a location. This has two reasons. First of all, we were mainly interested in the ability of our model to point out the warmest areas within a city. Secondly, these shadow layers use a detailed input layer with locations of all trees in the Netherlands. It would have been hard to find such a layer abroad and also on a more coarse resolution it does not add a lot to point out the warmest places in a city. That is, the shadow of trees does not cause warmest places to shift elsewhere. However, in the final implementation of a solution at a certain area on a more detailed level it may be important to determine the places that are most exposed to direct sunlight. Using the model from Figure 4 and by implementing these assumptions the resulting model is shown in Figure 6. Also three special nodes are indicated with yellow diamonds, which caused trouble in the model (see Section 3.3.2).

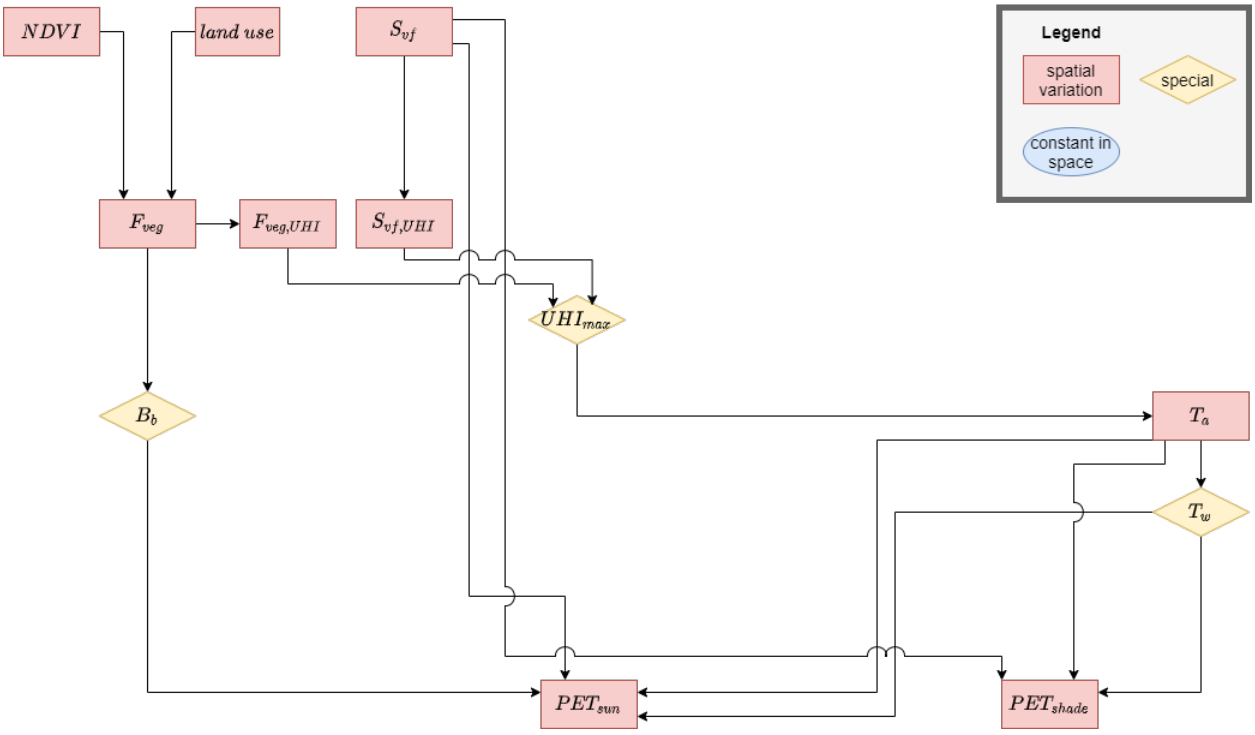


Figure 6: Proposed PET implementation according to assumptions. A distinction is made between variables with spatial variation (red squares) and variables without spatial variation (blue ellipses). The latter group was removed from the model

3.3.2 Perfect rank correlation

Using this model, problems arose with three nodes: the wet bulb temperature (T_w), maximum urban heat island effect (UHI_{max}) and Bowen ratio (B_b). As mentioned in Section 3.4, the Bayesian Network was quantified using the recursive partial correlation formula on ranked, standard normalised variables, according to Equation 15. Equation 11 implies that T_w is monotonically increasing with T_a, given a non changing relative humidity Φ. As the current method of Nelen & Schuurmans (2020) includes the relative humidity as a constant, this is one of the variables omitted in our proposed model (Figure 6). However, as a consequence, variability in T_w is only caused by a changing value of T_a. As this relationship is monotonic, this resulted in a perfect rank correlation of $\rho_{T_a,T_w} = 1$. In the recursive solution of the whole NPBN, chances were this variable was occurring in one of the square-roots in the denominator, leading to a partial rank correlation that could not be solved as the denominator equals 0. Also from a physical point of view it did not make sense to include both T_a and T_w when T_w is only dependent on T_a. When looking at rank correlations, adding the node T_w when T_a is already in the network does not add any information to the network. The same holds for UHI_{max} and its relationship with T_a, for a spatially constant $T_{station}$ (Equation 7). Finally, the B_b and F_{veg} were also perfectly correlated by definition (Equation 4). Therefore both T_w, UHI_{max} and

B_b were removed from the model, leading to the base model which is explained in more detail in the next section.

3.3.3 Qualification of the final base model

In Figure 7 the Directed Acyclic Graph (DAG), with the current approach to estimate the PET, is summarised by connecting all used variables. Whether variables were included as nodes, has been discussed at length in the previous sections. In this section we discuss the design choices regarding the edges between the nodes.

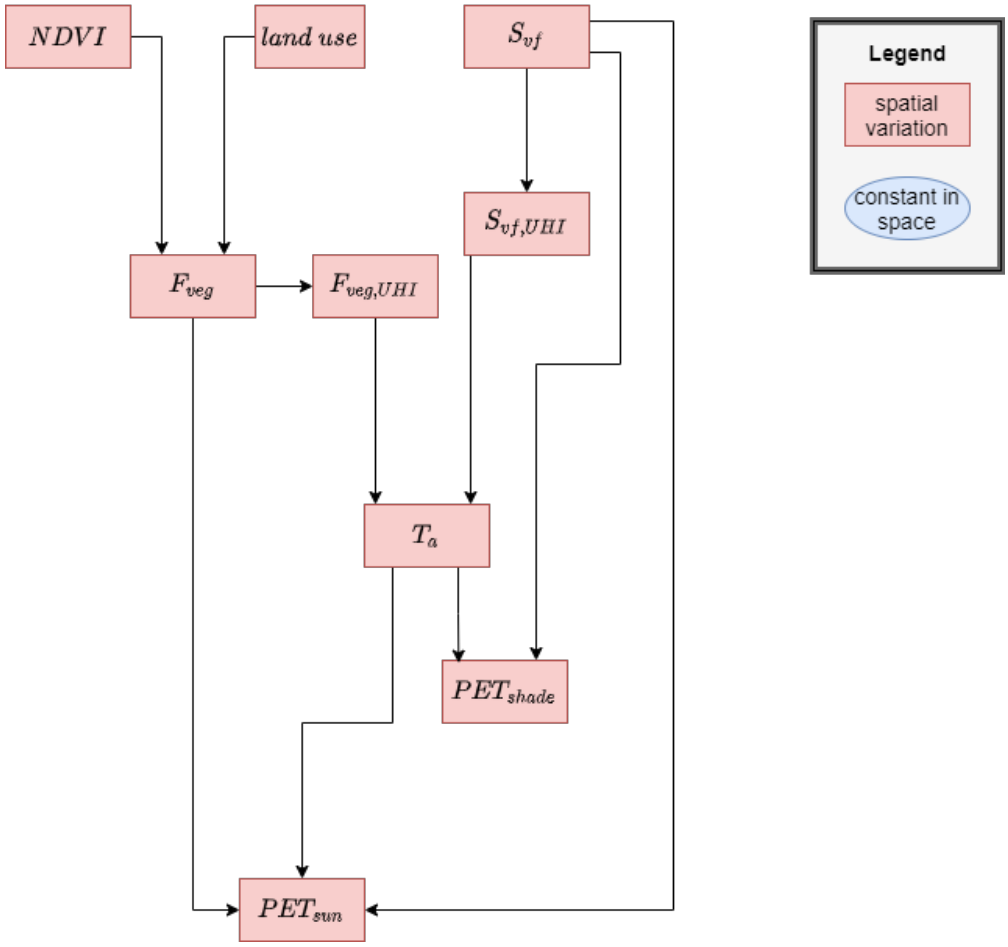


Figure 7: Proposed PET implementation according to assumptions and limitations of the model

First of all, this plot shows the ways one variable has influence on the final PET estimates through the connections between the nodes, the edges. Each edge represents a direct dependence between two variables. Specifying the edges between a DAG can be done in multiple ways, which would result in different dependence models. To come to this base model, we chose to stay as close to the current approach as possible. The current approach used by Nelen & Schuurmans (2020) is composed of multiple steps in which finally a PET

value is calculated, according the the formulas in Section 3.2. In these steps variables are linked to each other by using physical underlying processes. For this reason, it made sense to add edges in the same way the connections are shown in Figure 7. $NDVI$ and *land use* determine the vegetation dummy F_{veg}. Due to the cooling evaporation gives, this F_{veg} is directly connected to PET_{sun} as well. S_{vf} influences PET_{shade} directly and through the urban heat island effect. This latter connection was investigated by Dirksen et al., 2019. The sky view factor affects the surface energy balance, local air circulation, and outdoor thermal comfort (He et al., 2014). The surface energy balance involves evaporation and therefore a link to PET_{sun} is made. The other factors influence both PET_{shade} and PET_{sun}. The actual temperature resembles both air temperature and the urban heat island effect, explaining the connections to this node.

Considering the marginal distributions of the variables in the nodes, it is important to note that there are several non-continuous data layers present in the network. First of all, *land use* is a categorical variable. We changed this variable to either zeros for land use classes that do not act as vegetation and ones for land use classes that do act as vegetation (agricultural land and water bodies). This way, we reformed this variable to an ordinal variable, making it applicable for the model. F_{veg} is also an ordinal discrete variable. Hanea and Harrington (2009) made clear that up to a certain part of the NPBN may consist of ordinal discrete random variables, so we experimented with adding and removing these variables to study the performance of the model.

3.3.4 Quantification

Non-Parametric Bayesian Networks differ in the method of quantification compared to Bayesian Networks. Although the name of this former type of models suggests no parameters are involved at all, some nuance should be added. Although there still are some parameters involved in quantifying a NPBN, there are a lot less of them. In NPBNs the random variables in each node can be defined with a marginal distribution. For this, the empirical distribution from data can be used or one can specify a parametric distribution. Then, bivariate copulas are used to establish the dependency between pairs of nodes. The copulas that are often used, contain one parameter and are parametrised by Spearman's rank correlations. Thus, in total we now have as many parameters as we have arcs in the BN, plus any parameters needed to specify a marginal distribution (when) choosing to represent such a marginal distribution with a parametric distribution). We do not need to fill conditional probability tables and also we do not need to specify parameters for fully parametric joint distributions, including conditional distributions and correlations. So, although there are still some parameters left to be specified, the amount is greatly reduced. Also, deducing them by using expert knowledge is less complex (Morales, Kurowicka and Roelen, 2008), as these lower number of parameters are easier to interpret as well. Finally, Hanea, Kurowicka and Cooke (2007) showed that it is possible to

add discrete variables to a NPBN as well, under the restriction that these variables are ordinal. However, they also stated that one should consider switching to a regular BN if ordinal discrete random variables are dominating the network.

As briefly mentioned above, the dependence between nodes in a NPBN is represented by bivariate copulas. A copula of two continuous random variables X and Y is the joint distribution of the cumulative distribution functions (CDFs) $F_X(X)$ and $F_Y(Y)$. The copula is a distribution on the two-dimensional unit square area $I^2 = [0,1]^2$, with uniform marginal distributions. This latter condition can be met by transforming our marginal (empirical) distributions of our variables to uniform marginal distributions through their ranks. Because of this step, the copula of two variables can also be interpreted as the dependence structure of the ranks of these two variables. Copulas with only one parameter have a 1-to-1 relationship between the copula parameter and Spearman's rank correlation, so any copula realising all correlations in [-1,1] can be used. The Gaussian copula is used when the dependence is symmetric. However, if the dependence becomes much more pronounced in the tail(s) of a distribution, one can consider either the Gumbel copula for upper tail dependence or the Clayton copula for lower tail dependence, or any other type of copula depending on the type of asymmetry.

During the quantification of the model, only the copula parameters were estimated. In the quantification of a NPBN therefore the only model assumption was done with respect to the copula choice. The Cramer-von Mises statistic M could be used to validate this choice. This statistic can be used in general to assess the degree of agreement of a CDF with the empirical CDF. In our specific case the we could compare the CDF of the copula parametrised during the quantification with an empirical copula. This empirical copula can be calculated as the empirical CDF of the rank transformed data. A lower value for M indicates a better fit, so without a formal test we used this statistic to compare the Gaussian copula against two other copulas, the Gumbel and the Clayton copula.

Quantification of the NPBN was done using the recursive partial rank correlations. This was done the same way as described by Paprotny, Morales-Nápoles et al. (2020). For each pair of variables of which the partial rank correlation had to be calculated, the data of both nodes was transformed into standard normalised data. In this way, a Gaussian copula parametrised by the Spearman's rank correlation could be used to define the correlation between two connected nodes. This process was done recursively using Equation 15. Each edge represents a bivariate copula, which indicates the partial rank correlation between the variables connected by this edge. With 'partial rank correlation' we mean the correlation between two ranked variables, given the rank information of the other variables.

3.4 BANSHEE

Despite the growing literature on NPBNs and the growing evidence of useful applications, this type of models are not widely applied within industries or educational environments. As Paprotny, Morales-Nápoles et al. (2020) noted "an important limiting factor in making new analyses with NPBNs is software availability." They therefore introduced BANSHEE, a MATLAB toolbox with open-access code. This enables researchers to provide transparent, reproducible results with NPBNs and gives a framework in which new ideas or theories can be tested. We quickly walk through the capabilities of this toolbox, as we use these in our research.

The first step consists of defining the NPBN. The user needs to select the data he/she wants to use in the NPBN and the structure (edges between nodes) needs to be defined. This qualification needs to be done manually, as expert knowledge regarding the specific use case is involved in this step, which prevents the ability to automate this step. Secondly, quantification of the NPBN is done. The code was written in such a way that it can quantify a NPBN based on Gaussian copulas. This assumption can be checked later on. One of the reasons to start with implementing the option of using Gaussian copulas, is the fact that inference of a Gaussian NPBN is most practical regarding computational speed. During quantification of the NPBN, we can apply fast calculations to come to the Gaussian copulas because a multivariate normal distribution can be built from and split into separate univariate normal distributions easily. Also the transformation from Pearson's correlation coefficient ρ to Spearman's correlation coefficient r is easily done for a bivariate normal distribution (Moran, 1948):

$$r = \frac{6}{\pi} \arcsin\left(\frac{\rho}{2}\right) \tag{13}$$

and the inverse formula:

$$\rho = 2 \sin\left(\frac{\pi}{6} r\right). \tag{14}$$

The following steps are taken during quantification. First the data of each node is transformed to ranked data. This data can be made uniformly distributed on $[0, 1]$ by dividing by the number of observations. These uniformly distributed rank data are the margins of the one-parameter bivariate copula C_θ. Since we use Gaussian copulas, we can transform these margins to standard normal distributed data using the inverse CDF method. Then the Pearson's correlation between both nodes describes the Gaussian copula. To come to a partial correlation matrix given the network, we calculate these Pearson correlation values using Equation 15:

$$\rho_{XY \cdot Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{ZY}}{\sqrt{1 - \rho_{XZ}^2}\sqrt{1 - \rho_{ZY}^2}}. \tag{15}$$

Using Equation 13 we could calculate the conditional rank correlations from these conditional Pearson correlations.

After qualification and quantification of the NPBN, the assumption of Gaussian copulas was checked using two validation criteria. The first criterium is the Cramer-von Mises statistic M that was introduced in Section 3.3.4. It was implemented as (Paprotny, Morales-Nápoles et al., 2020):

$$M_n(\boldsymbol{u}) = n \sum_{|\boldsymbol{u}|} \left\{ C_{\hat{\theta}_n}(\boldsymbol{u}) - B(\boldsymbol{u}) \right\}^2, \boldsymbol{u} \in [0,1]^2 \tag{16}$$

where the empirical copula was calculated as $B(\boldsymbol{u}) = \frac{1}{n} \sum_{i=1}^{n} 1(U_i \leq \boldsymbol{u})$ and the parametric copula $C_{\hat{\theta}_n}(\boldsymbol{u})$ was determined by estimating $\hat{\theta}_n$ from the sample.

Besides this test, we also used the d-calibration score (Morales-Nápoles, Hanea and Worm, 2014). During the qualification process of a BN, one can check if the chosen edges agree with the data using this score (Morales-Nápoles, Hanea and Worm, 2014). We elaborate on the concept of D-separation in Section C. The d-calibration score compares "the empirical correlation matrix (the data) with both the BN rank correlation matrix and the empirical normal rank correlation matrix (the model)" (Paprotny, Morales-Nápoles et al., 2020). Comparing with the latter one gives an indication of the applicability of the Gaussian copulas to describe the joint distribution. Comparing the empirical correlation matrix with the BN rank correlation matrix (from the quantification) measures the degree to which the chosen conditional independence statements implied by the BN agree with the data. This makes this metric useful for validation of the assumptions done during the qualification of the NPBN.

3.5 BANSHEE-y

The original BANSHEE package has been written in MATLAB. As mentioned in Section 3.4 one of the reasons to develop this toolbox, was to provide an open-access code of NPBNs. Although this is useful for researchers, the MATLAB software itself is not open source. This is disadvantageous for both companies and for students. The former would like to experiment with implementing NPBNs for their challenges, while the latter are already learning more about the open source programming language Python instead of MATLAB. Therefore we have converted BANSHEE into BANSHEE-y. In this process, we tried to minimise the number of dependencies with other packages. Compared to BANSHEE, we made an addition to the visualisation of the NPBN. Instead of only showing the rank correlations for all edges, we also added the marginal distribution of each node in the graph.

3.6 Overview simulations

To structure the results, we colour coded the results of each sub question in all performance tables (Table 5). Sub question four was answered by runs 10 to 24. As the data layer that is omitted, differs amongst these runs, we used four shades of blue to make this distinction.

Table 5: Each sub question came with one or more runs which were grouped into colour-coded categories, indicating the goal of the run. Note that sub question 4, covering missing data layers, consists of multiple categories, each focusing on a different missing data layer

Sub question	Goal
1	**Theory**
	City vs rural
2	**Out of sample**
3	**Coarseness**
4	**Replace Sky view with** $AHN2$
	$NDVI$ **importance compared to** F_{veg}, $F_{veg,UHI}$
	No *land use* **available/refine** *land use*
	T_a **predicted instead of conditionalised on**
5	**# training observations**
6	**Given** PET**, what** $NDVI$

In Table 6 we give a small textual description for each run, while Table 7 summarises the technical details for every run.

Table 6: Descriptive summary runs

Goal	Run ID	Use Case Description
Theory	**01**	Using resolution 1 m, see in-sample performance (Rural area)
	02	Using resolution 10 m, see in-sample performance (Rural area)
City vs rural	**03**	Using resolution 1m, see in-sample performance for city-sample
	04	Using resolution 10 m, see in-sample performance for city-sample
Out of sample	**05**	Using resolution 1m, see out-sample performance
Coarseness	**06**	Using resolution 10 m, see out-sample performance
	07	Using resolution 20 m, see out-sample performance
	08	Train on fine data (1m), test on coarser data (10m)
	09	Train on fine data (10m), test on coarser data (20m)
Replace Sky view with AHN2	**10**	Svf and svf_uhi used to train, but not to predict
	11	Same as 10, but with extra data layer ahn2 to back up missing svf during prediction
	12	Same as 11, but with ahn2_stdev instead of ahn2
	13	Same as 11, but with ahn2_mean_corrected instead of ahn2
	14	Completely leave out svf and svf_uhi (also during training), just use ahn2
	15	Same as 12, but with ahn2_stdev instead of ahn2
	16	Same as 12, but with ahn2_mean_corrected instead of ahn2
NDVI importance	**17**	Fveg and fveg_uhi available in training, but not in test
compared to Fveg,	**18**	Completely leave out fveg, but fveg_uhi is still used in training and depends on NDVI
Fveg_uhi	**19**	Completely leave out fveg and fveg_uhi, only use NDVI
Refine *land use*	**19a**	Same as 19, but split binary variable to increase performance. NPBN with sample filtered on *land use*==1
	19b	Same as 19, NPBN with sample filtered on *land use*==0, combine with 19a
	19c	Same as 19, split *land use* into 5 ordinal classes based on evaporation rate: buildings, unpaved, water, grass, trees
No *land use* **available**	**20**	Completely leave out fveg and fveg_uhi. Also exclude land_use
	21	Use case with only raw data input: no fveg or fveg_uhi, no land_use and also no svf or svf_uhi
Ta pred	**22**	Same as 21, but now ta is also predicted instead of being conditionalized on
	23	Same as 20, but now ta is also predicted instead of being conditionalized on
	24	Same as 05 (base case), but now ta is also predicted instead of being conditionalized on
# training observations	**25**	Same as 22, but with only 200 observations from Wageningen to quantify the NPBN
	26	Same as 22, but with only 50 observations from Wageningen to quantify the NPBN
	27	Same as 22, but with only 20 observations from Wageningen to quantify the NPBN
Given PET, what NDVI	**28**	Given a PET, what is the NDVI? This can be used to assess what NDVI we need for a certain PET
	29	Given a PET, what is the NDVI? Also leave out fveg_uhi, only vegetation related factor now is NDVI

Table 7: Technical summary runs

Goal	Run ID	Resolution (m)	# obs N	Variable set (P=prediction, C=condition)									Extra variable
				PET_sun	PET_shade	svf	ta	fveg_uhi	svf_uhi	fveg	NDVI	land use	ahn2
Theory	**01**	1	2,000	P	P	C	C	C	C	C	C	C	
	02	10	2,000	P	P	C	C	C	C	C	C	C	
City vs rural	**03**	1	2,000	P	P	C	C	C	C	C	C	C	
	04	10	2,000	P	P	C	C	C	C	C	C	C	
Out of sample	**05**	1	2,000	P	P	C	C	C	C	C	C	C	
Coarseness	**06**	10	2,000	P	P	C	C	C	C	C	C	C	
	07	20	2,000	P	P	C	C	C	C	C	C	C	
	08	train 1, test 10	2,000	P	P	C	C	C	C	C	C	C	
	09	train 10, test 20	2,000	P	P	C	C	C	C	C	C	C	
Replace svf with AHN2	**10**	1	2,000	P	P	P	C	C	P	C	C	C	
	11	1	2,000	P	P	P	C	C	P	C	C	C	C
	12	1	2,000	P	P	P	C	C	P	C	C	C	C (σ_{ahn2})
	13	1	2,000	P	P	P	C	C	P	C	C	C	C (ahn2-μ_{ahn2})
	14	1	2,000	P	P		C	C		C	C	C	C
	15	1	2,000	P	P		C	C		C	C	C	C (σ_{ahn2})
	16	1	2,000	P	P		C	C		C	C	C	C (ahn2-μ_{ahn2})
NDVI importance	**17**	1	2,000	P	P	C	C	P	C	P	C	C	
compared to Fveg,	**18**	1	2,000	P	P	C	C	P	C		C	C	
Fveg_uhi	**19**	1	2,000	P	P	C	C		C		C	C	
Refine *land use*	**19a**	1	142	P	P	C	C		C		C		
	19b	1	1,858	P	P	C	C		C		C		
	19c	1	2,000	P	P	C	C		C		C	C[(1)]	
No *land use* **available**	**20**	1	2,000	P	P	C	C		C		C		
	21	1	2,000	P	P		C				C		C (ahn2-μ_{ahn2})
Ta prediction	**22**	1	2,000	P	P		P				C		C (ahn2-μ_{ahn2})
	23	1	2,000	P	P	C	P		C		C		
	24	1	2,000	P	P	C	P	C	C	C	C	C	
# training observations	**25**	1	200	P	P		C				C		C (ahn2-μ_{ahn2})
	26	1	50	P	P		C				C		C (ahn2-μ_{ahn2})
	27	1	20	P	P		C				C		C (ahn2-μ_{ahn2})
Given PET, what NDVI	**28**	1	2,000	C	C	C	P	P	C		P		
	29	1	2,000	C	C	C	P		C		P		

(1) In this run *land use* was refined into 5 ordinal classes representing evaporation rate: buildings, unpaved, water, grass, trees

3.7 Model validation

The model was validated on several levels, which mainly considers the first two sub questions. First of all, the underlying assumptions made during the quantification of the model could be tested. Specifically, the choice of using a Gaussian copula was validated using the Cramer-von Mises statistic M (Equation 16). This statistic was calculated for the Gaussian, Gumbel and Clayton copulae. Besides this, the d-calibration score could be used to evaluate the choices made regarding the edges between nodes. Although these statistics had been calculated and reported for all runs, we focused on these results during the runs that are part of the first sub question about theoretical applicability (runs 01 to 04, see also Section 3.6).

The next level of performance was done by considering the performance of the model in terms of inference. We wanted the model to be able to estimate both PET_{sun} and PET_{shadow} for several locations, while being able to use less data as input during inference. A first performance measure used to quantify this, is the Mean Absolute Error (MAE) score. This score was constructed as:

$$MAE_j = \frac{1}{n} \sum_{i=1}^{n} \{|\hat{y}_{i,j} - y_{i,j}|\}, \tag{17}$$

where j indicates the variable to be assessed, $\hat{y}_{i,j}$ is the prediction of observation i for variable j and $y_{i,j}$ is the predicted value for the same observation. We started testing the in-sample accuracy with this statistic, but we also tested the model out-of-sample. This concerns runs 01 up to run 05 (the latter being the base model), see also Section 3.6.

3.7.1 Reporting and visualisation of the metrics

The results are presented by visualising different metrics:

- NPBN visualisation plots: in these plots the conditional rank correlations are shown for the edges that are directly connected, as specified during the construction of the DAG (e.g. Figure 12);
- Cramer-von Mises plot: plot with Cramer-von Mises scores between all variable combinations, relevant to compare between different DAGs or train datasets (e.g. Figure 13);
- Inference plots: in these plots the inference of one observation is shown, drawing K samples. In this research $K = 1,000$ (e.g. Figure 26c);
- Visualisation of the Gaussian Distance: this plot visualises both ERC and BNRC d-calibration scores compared to the 90% confidence interval of the determinant of the empirical normal distribution (e.g. Figure 16);
- D-calibration scores tables: different DAGs or train datasets result in different outcomes regarding performance measures related to the theoretical applicability. Whenever relevant, we show the D-calibration scores (e.g. Table 15);
- MAE tables: For all variables we predict in a test, we obtain in-sample and out-sample MAE (Mean Absolute Error) scores. We show these scores, or a subset for each relevant run. Also we add CI-scores. This score indicates the percentage of observations that fall within the 95% confidence interval during inference. This 95% confidence interval is determined per observation, by assuming a normal distribution in the K samples drawn and estimating the population standard deviation $\hat{\sigma}_i$ within this

set of K draws. We also estimate the population mean $\hat{\mu}_i$. For each observation i we determine whether it falls within the constructed 95% confidence interval $[\,\hat{\mu}_i - 1.96\,\hat{\sigma}_i, \hat{\mu}_i + 1.96\,\hat{\sigma}_i\,]$. A CI score of 95% means that there is a 95% probability that the constructed interval $[\,\hat{\mu}_i - 1.96\,\hat{\sigma}_i, \hat{\mu}_i + 1.96\,\hat{\sigma}_i\,]$ contains the true parameter value of interest (e.g. Table 8);

- AE plots: in these plots we split up the Mean Absolute Error score into individual Absolute Errors (AE), grouped per degree of temperature according to the observed (real) temperatures (e.g. Figure 17a);
- Error distributions: distribution of the errors, color coded per degree of temperature according to the observations (e.g. Figure 23a);
- Geo plots: either the observed temperatures, the predicted temperatures or the errors are plotted in a 2D grid (e.g. Figure 11).

3.8 Model extension: adding variables

The qualification of our model using the DAG was mainly based on the heat stress recipe from the RIVM and starts with the input data layers available at Nelen & Schuurmans (2020). Therefore the first runs explored different configurations regarding training sample, resulting into the base case in run 05. As the sub questions three, four and five concern limited data availability in the form of respectively data resolution, missing data layers and a limited amount of data points, we extended this model of run 05 in several ways. First of all, run 06 and run 07 used the base model (resolution 1 m) to explore model applicability using 10 and 20 m resolution respectively. These runs covered the range of finest resolutions available on a global scale at this moment, as was mentioned at the end of Section 2. Run 08 and 09 are slightly different to run 06 and 07 (see Table 6). The next sub question considers missing data layers. Run 10 up to 16 covers sky view as a missing data layer. We replaced this layer by adding the AHN2 map. This map is also used to calculate the sky view factor. However, calculating the sky view factor is a very heavy computational process, resulting into a very limited number of places where this layer is available. We tested what happens when not conditioning on the sky view node during inference and compared the results with and without an extra $AHN2$ node, to see if adding such a layer would lead to better PET estimates in regions where no sky view factor is available. In run 10 we modelled the situation of training the NPBN with information on S_{vf}, but testing on Rotterdam without using information about S_{vf} over there. In run 11 the AHN2 layer was added during training and testing to see if any correlation exists between $AHN2$ and S_{vf} that could compensate for the missing data on S_{vf} in the test data. Finally, run 14 completely excludes both S_{vf} and $S_{vf,UHI}$ and trains and tests the NPBN using only $AHN2$ as a sky view-related variable. The marginal distributions of this new layer for our study areas are shown in Figure 8. As can be seen in the upper graph

of Figure 8c, the marginal distributions of both training and validation area differ. Besides that, because of the way the sky view factor S_{vf} is constructed (Equation 3), the rank correlation between the raw $AHN2$ data and S_{vf} might not be high. For these two reasons we derived two statistics from $AHN2$ to compare their performance in the NPBN with the raw $AHN2$ layer. We only did this for the layers with resolution of 1 m. In the upscaling proces from a 10 m grid to a 1 m grid, we constructed these two statistics as

$$\sigma_{ahn2,i,j} = \sqrt{\left(\frac{1}{10 * 10} \sum_{k=i-5}^{10} \sum_{l=j-5}^{10} (ahn2_{k,l} - \mu_{i,j}) \right)} \tag{18}$$

and

$$ahn2_{i,j} - \mu_{i,j}, \tag{19}$$

where

$$\mu_{i,j} = \frac{1}{10 * 10} \sum_{k=i-5}^{10} \sum_{l=j-5}^{10} ahn2_{k,l}. \tag{20}$$

i and j are the indices on the 1 m resolution grid of size (X, Y). Basically for every square meter we constructed two statistics that resemble the variability in height around that square meter, which approaches the sky view factor better than the raw $AHN2$ layer. We used these statistics in run 12 and 13 and in run 15 and 16 in the same set up as run 11 and run 14 respectively (see also Tables 6 and 7).

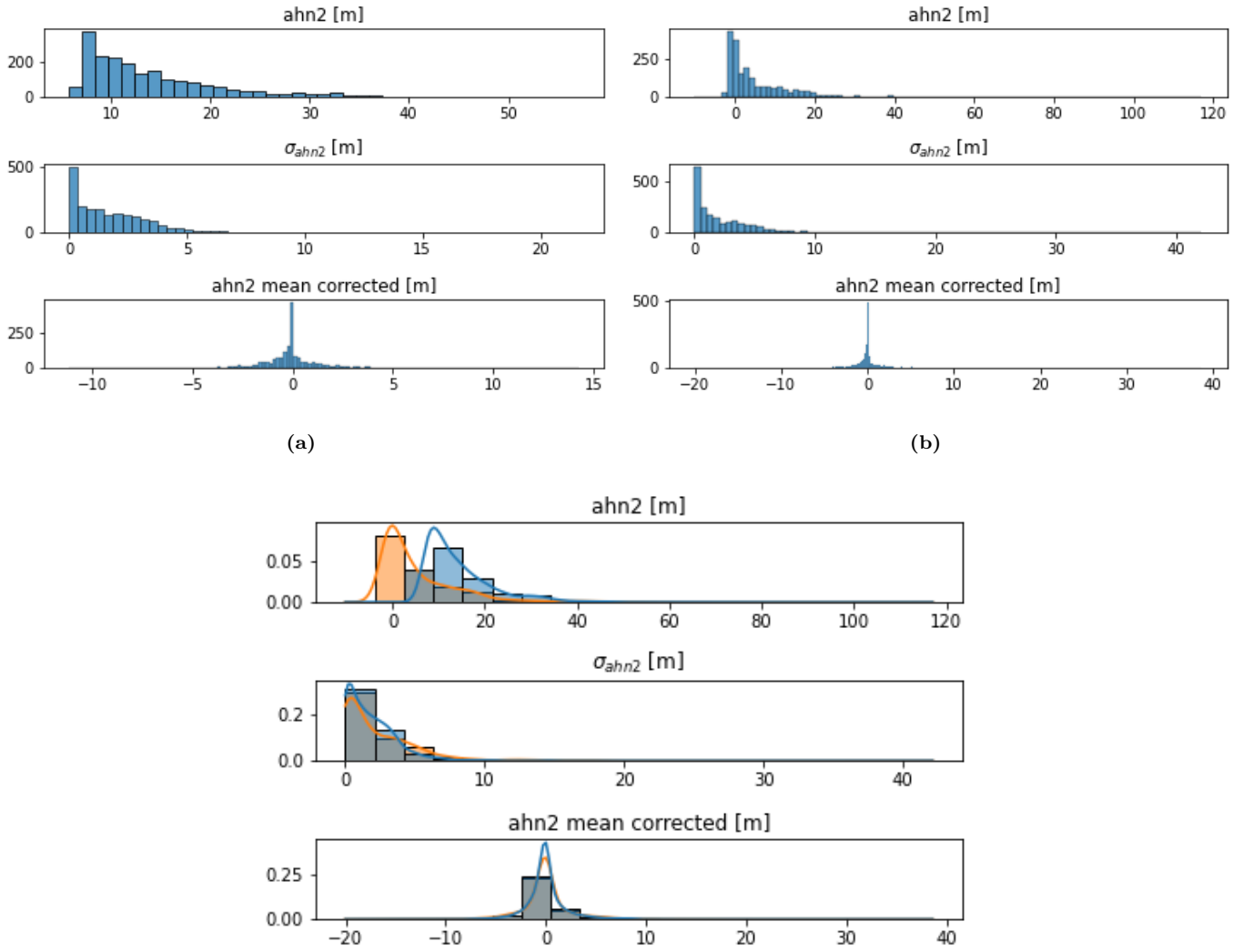


Figure 8: Marginal distributions of the AHN2 layer and two derived layers σ_{ahn2} and AHN2 mean corrected, for (a) Wageningen city (b) Rotterdam city (c) both of them on one axis

Next, with run 17 to 19 we explored ways to work around the vegetation layers. 19a, 19b and 19c were done to refine the *land use* layer, while run 20 and 21 were done without using the *land use* layer. In run 22 to 24 we added an extra missing layer to these layers, the actual temperature T_a. Finally runs 25, 26 and 27 covered the situation in which less than 2,000 data points are available to train on (respectively 200, 50 and 20 observations).

A final model extension followed naturally from the applications of the urban heat stress tool. An important application of this tool is to determine the influence of possible NBSs on the PET values. This can be

done in two ways: one can change the vegetation input for a certain location and observe the differences in PET values induced by this change. However, we investigated a second method, in which we did not predict PET_{sun} and PET_{shade}, but $NDVI$ instead, based on given PET_{sun} and PET_{shade} (runs 28 and 29). Following this method, the way of applying the results is by choosing a target PET value and deducing what $NDVI$ value would lead to such PET value.

4 Results

We first pose a general overview of the performance metrics of our complete study in Figures 9 and 10. After illustrating some key observations, we show the results for each run in the six subsequent sections. Each section covers one of the six sub questions posed in Section 1 and starts with a short summary with the most important conclusions regarding that sub question. After this first paragraph of each section, we focus on results that need extra explanation.

In Figure 9 the in-sample MAE scores and the in-sample CI scores are shown for both variables PET_{sun} and PET_{shade}. A first thing that stands out, is the fact that the variable PET_{shade} was easier to predict than PET_{sun} in all of our models. The Mean Absolute Error from inference was strictly lower and the percentage of observations that fall within the 95% confidence interval of the predictions from the NPBN, was almost always larger. For most of the runs this CI-score was close to 95%, as one would expect. From run 17 to run 27, this also held for the CI-scores for PET_{sun}. This improved score occurred in all runs in which less or no binary data layers (F_{veg} and *land use*) are being used. Run 24 is an exception: in this run the base model was tested when missing the actual temperature layer T_a. The CI for PET_{sun} decreases as in this run the binary data layers were present again.

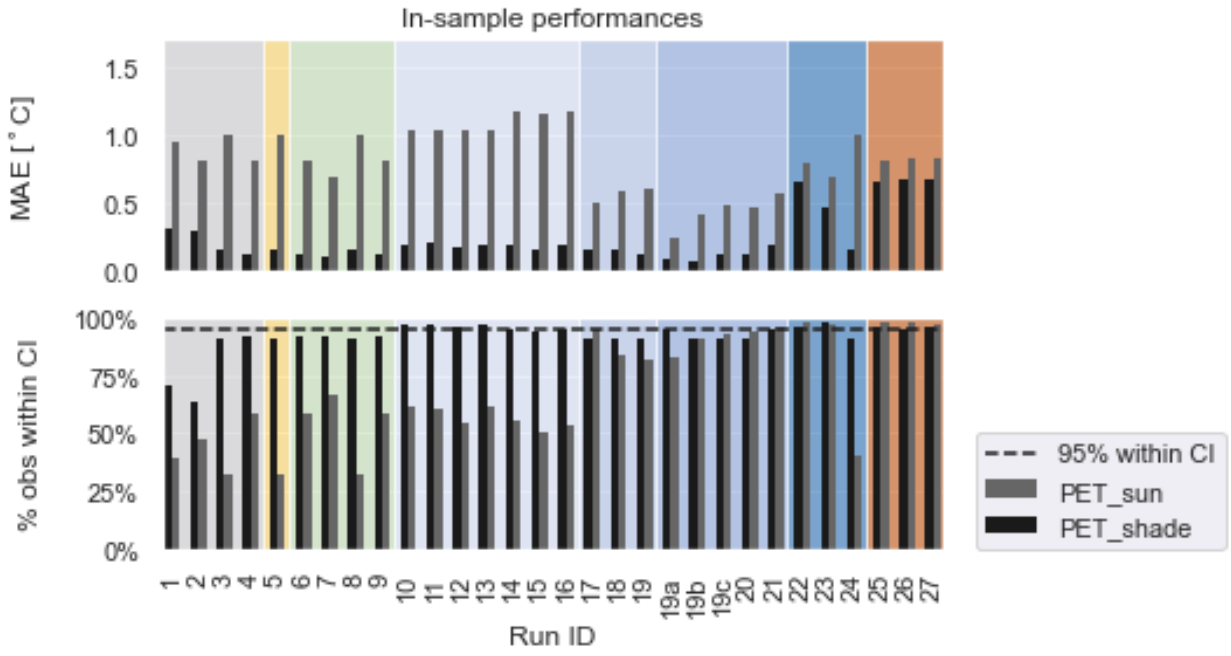


Figure 9: MAE scores (upper) and CI scores (lower) from the in-sample inference for each run

Out-sample inference was done on runs 5 to 27, using an urban area in Rotterdam. When looking at the performance, we see that the mean absolute errors were strictly equal or larger than in-sample. However,

the relative differences were larger for PET_{shade} than for PET_{sun}. We see that the mean average error was at most around 1.5 °C. A general pattern, which can be observed in runs 22 to 24 clearly, is the fact that a better MAE score implies a worse CI score. An explanation for this can be found in the fact that a lower MAE score in general means a better fit of the model to the data. This is reflected in a model that predicts with 'more certainty' and therefore computes predictions with less spread, leading to a smaller 95% confidence interval. Consequently, chances are more observations fall out of this smaller CI, leading to a lower CI score. Another observation was that never 95% of the observed values fell within the 95% confidence interval for the out-sample test. A main reason for this is the fact that during inference, we used the ECDF of PET_{sun} and PET_{shade} from Wageningen to do inference on points in Rotterdam. However, still run 22 (no F_{veg}, $Fveg, UHI$, $land\ use$, Svf and Svf, UHI layers) was able to obtain a CI-score of 67% and 79% for PET_{sun} and PET_{shade} respectively, hereby outperforming the base model (run 05) significantly against the cost of a larger MAE score. In Section 5 we pose a way to improve these results to even more satisfactory results. Using the sky view (run 23) we obtained the same performance when looking at the MAE and CI scores, but the predictions of observations in the tails of the distribution improved.

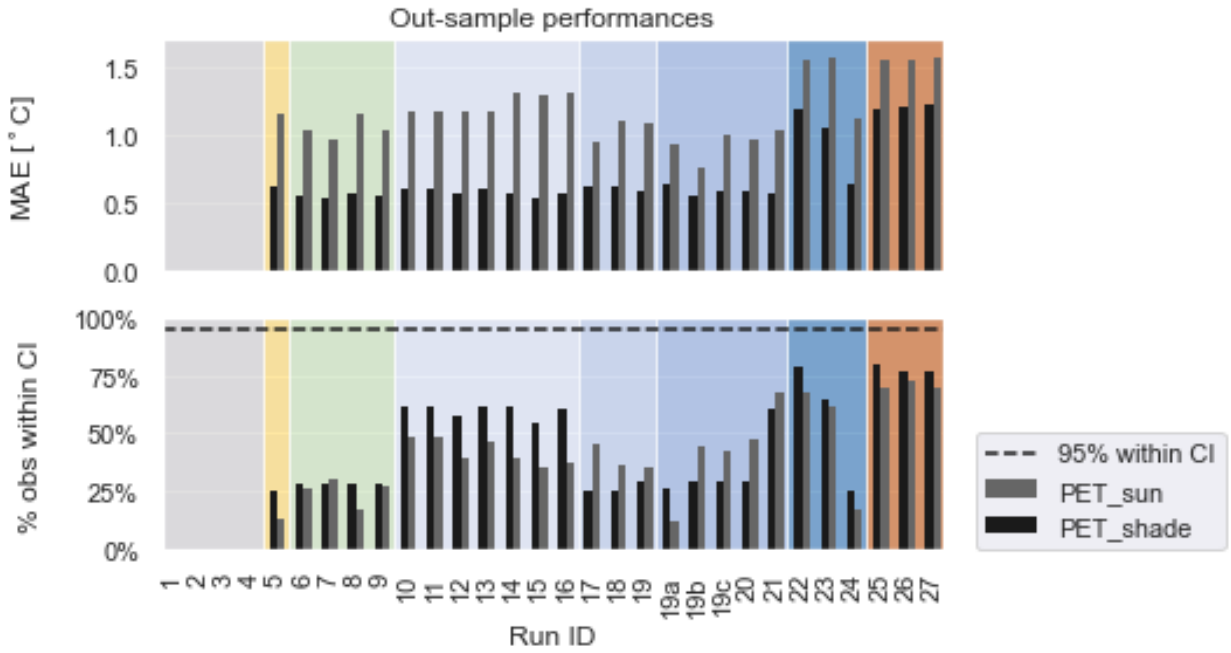


Figure 10: MAE scores (upper) and CI scores (lower) from the out-sample inference (Rotterdam city) for each run

4.1 Theoretical applicability: City vs Rural area

The first key finding is that the framework of NPBNs is applicable to assess urban heat stress, after making sure the sample data in the NPBN resembles the data in the area of interest and it does not contain too

much binary variables. When looking at the applicability of a NPBN to estimate heat stress, two main conclusion could be drawn. First of all, although the MAE scores did not differ a lot, the NPBN model trained only on Wageningen city data performed much better than the model trained on both rural and urban data points. The MAE scores did not represent this difference that much because the in-sample score for run 02 was calculated for a sample that includes both rural and urban data points. So although the models had roughly the same performance in terms of MAE scores, one should consider what the goal of the model is. In our case it was to predict urban heat stress, so training on only urban data points gave much better results. Secondly, we observed a bad performance in copula fit for all pairs of variables with at least one of them being a binary random variable. In the remainder of this section, we further elaborate on the differences between the model trained on both rural and urban data and the model trained on urban data.

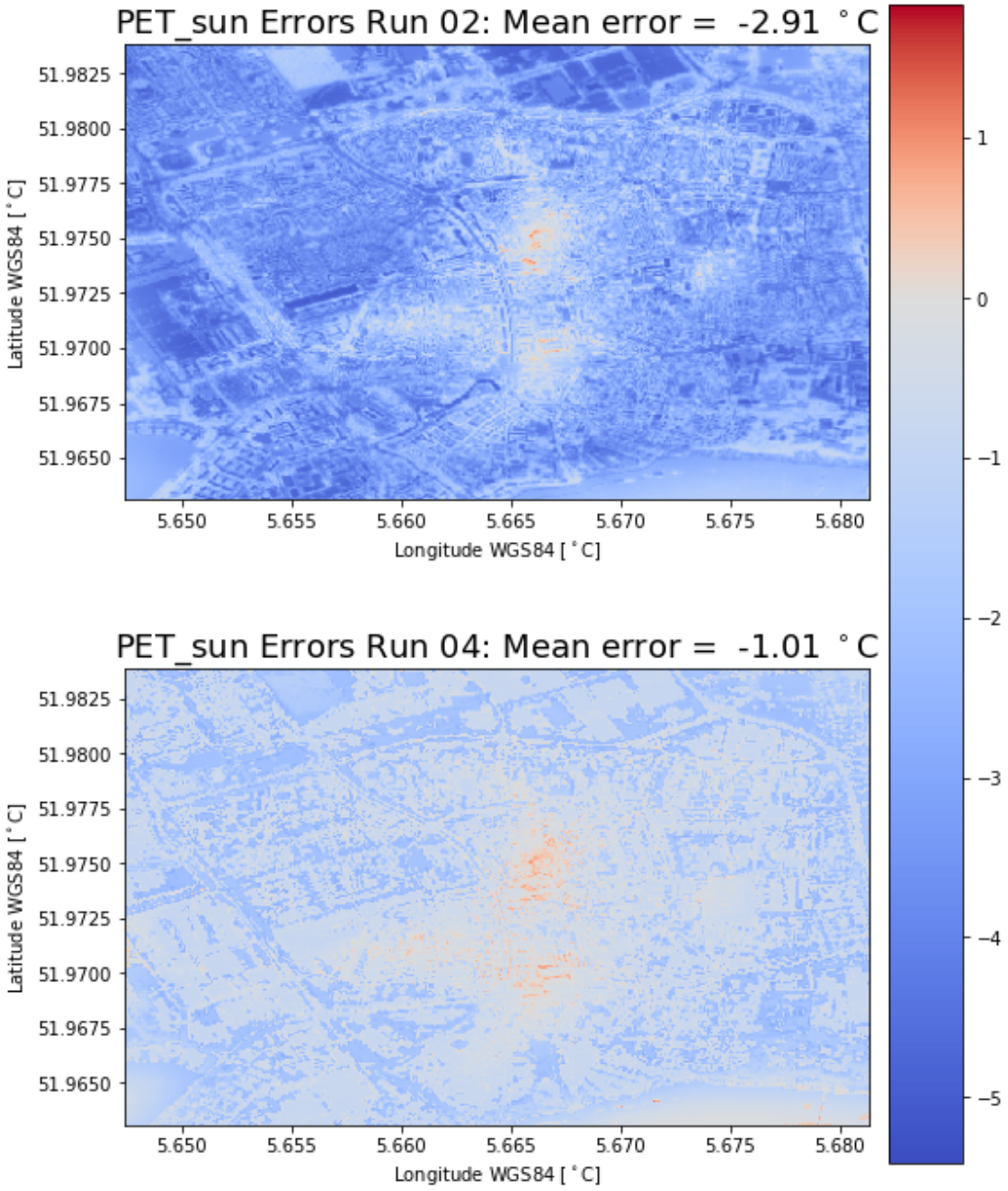


Figure 11: Wageningen city with the PET_{sun} errors from run 02 (trained on both rural and urban data) and run 04 (only trained on urban data). Resolution = 10 m. The mean error was calculated only for the area visible in the image; run 04 had a mean error that is almost three times lower than run 02 when considering only urban area

In Figures 12 and 13 we show respectively the quantified NPBN and the Cramer-von Mises plot of run 01. The DAG corresponds with the proposed base model in Figure 7. The NPBN in run 01 was quantified using the Wageningen Rural sample.

Figure 12: Visualisation NPBN run 01. In the nodes the marginal distributions of the variables were plotted

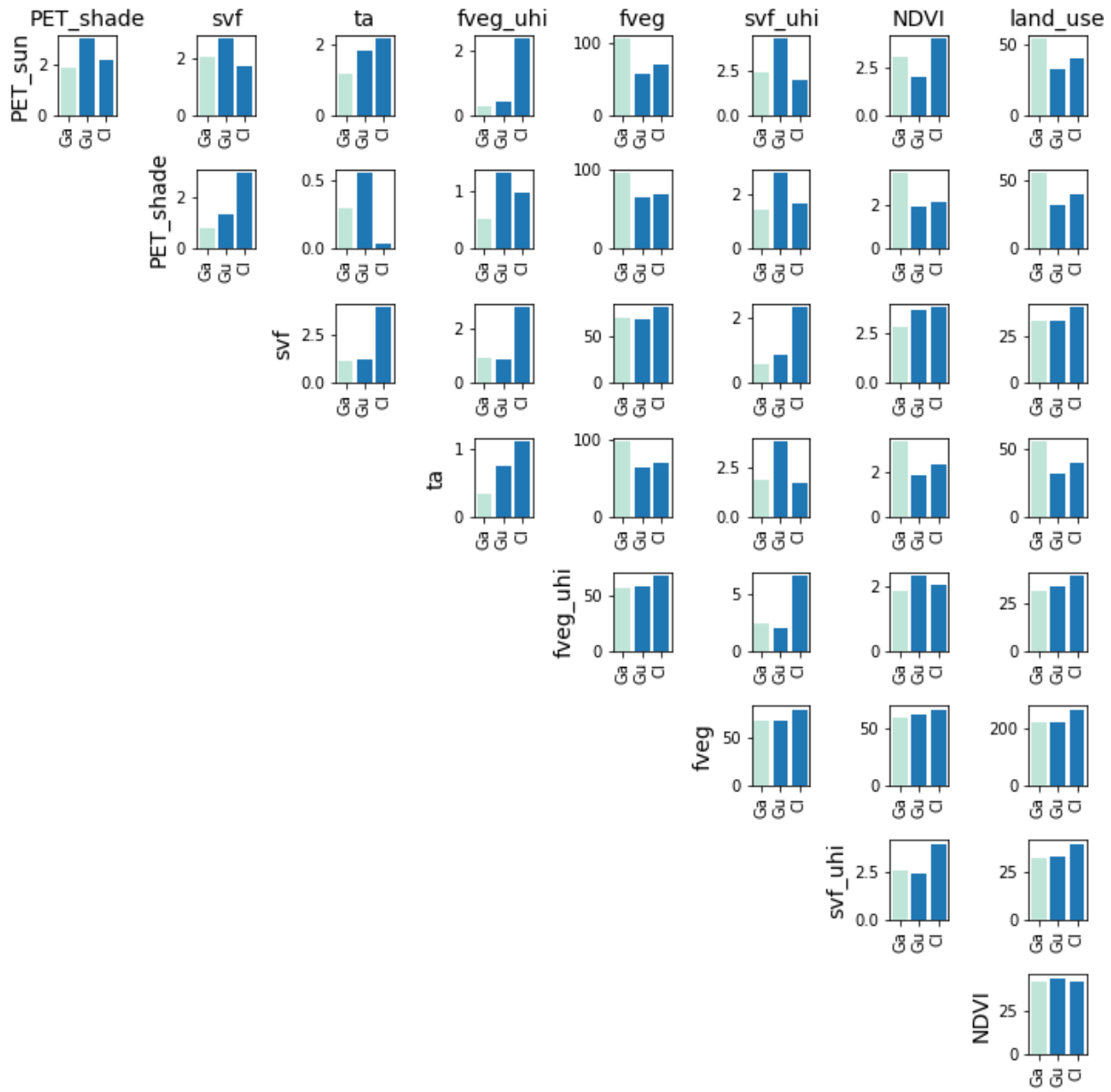


Figure 13: Visualisation CvM statistics run 01; three empirical copulas were fit to the data and compared to their theoretical equivalent: the Gaussian (Ga), Gumbel (Gu) and Clayton (Cl) copulas. For each variable combination the lowest bar indicates the best copula fit to represent that particular bivariate dependency.

This sample was compared to run 03, that was trained on the Wageningen City sample (Figures 14 and 15). Next to the observation that multiple variables transformed from bimodal to unimodal distributions, we observed that the sky view factor S_{vf} changes from a negative correlation with PET_{sun} for run 01 to a positive correlation with PET_{sun} for run 03. When comparing the Cramer-von Mises statistics, we observed that most values had decreased with a factor 2 to 4. Only the combinations containing at least one of the binary

random variables *land use* and F_{veg} had Cramer-von Mises (CvM) values that were an order of magnitude larger. These results indicate that using the dataset with only urban data points was more suitable to train the NPBN, leading to better copula fits in general. When excluding the bivariate copulas containing one or more binary random variables, for 10 of the 20 remaining copulas the Gaussian copula has the lowest CvM value. For only three copulas, all containing the variable $NDVI$, the Gaussian copula was the worst fit. However, the Clayton and Gumbel copulas also did not perform very well for these cases. This makes sense when looking at the marginal distribution of the $NDVI$ variable, which contains 774 zeros, making it hard for a rank-correlation based method to fit very well.

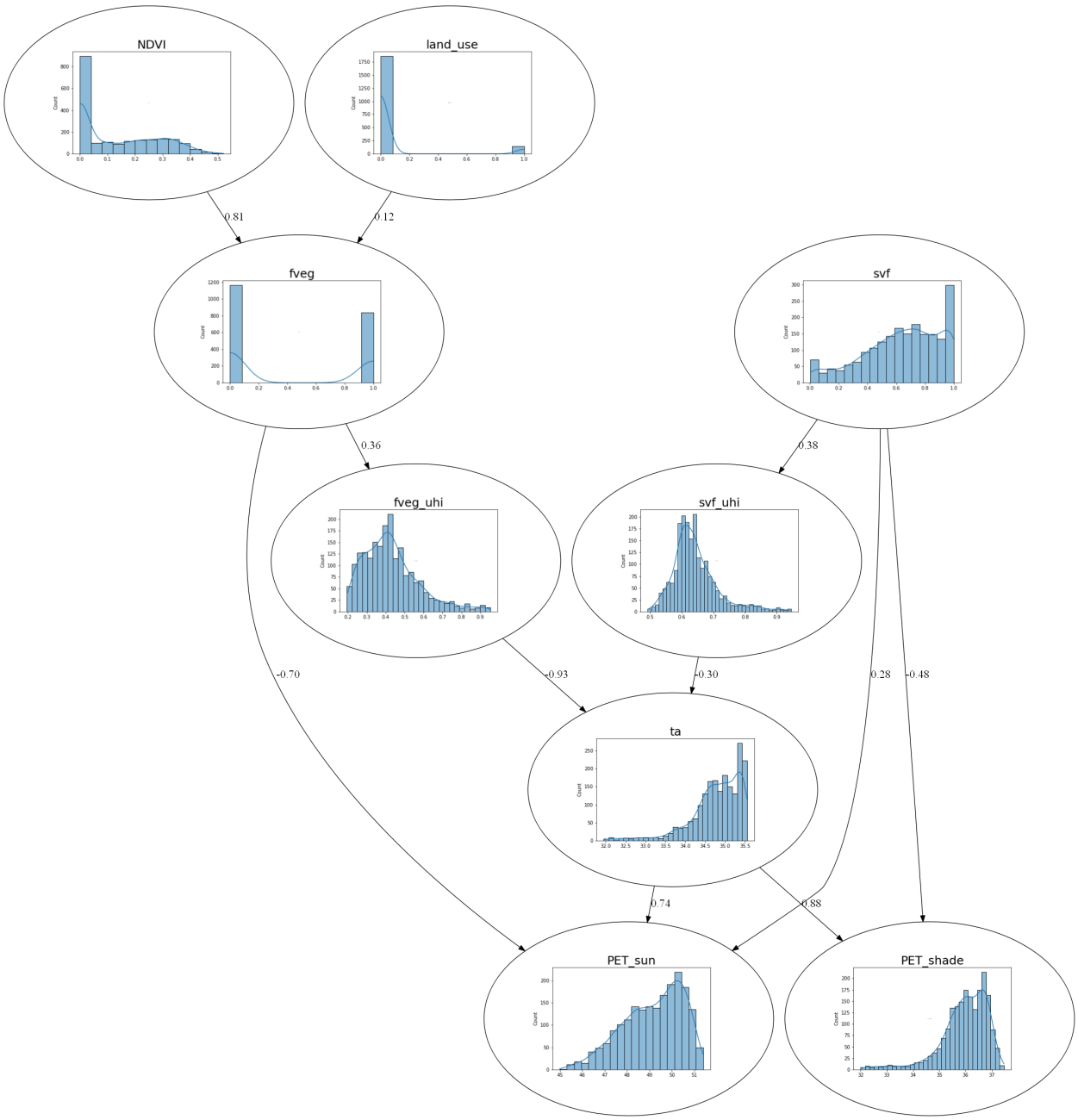


Figure 14: Visualisation NPBN run 03

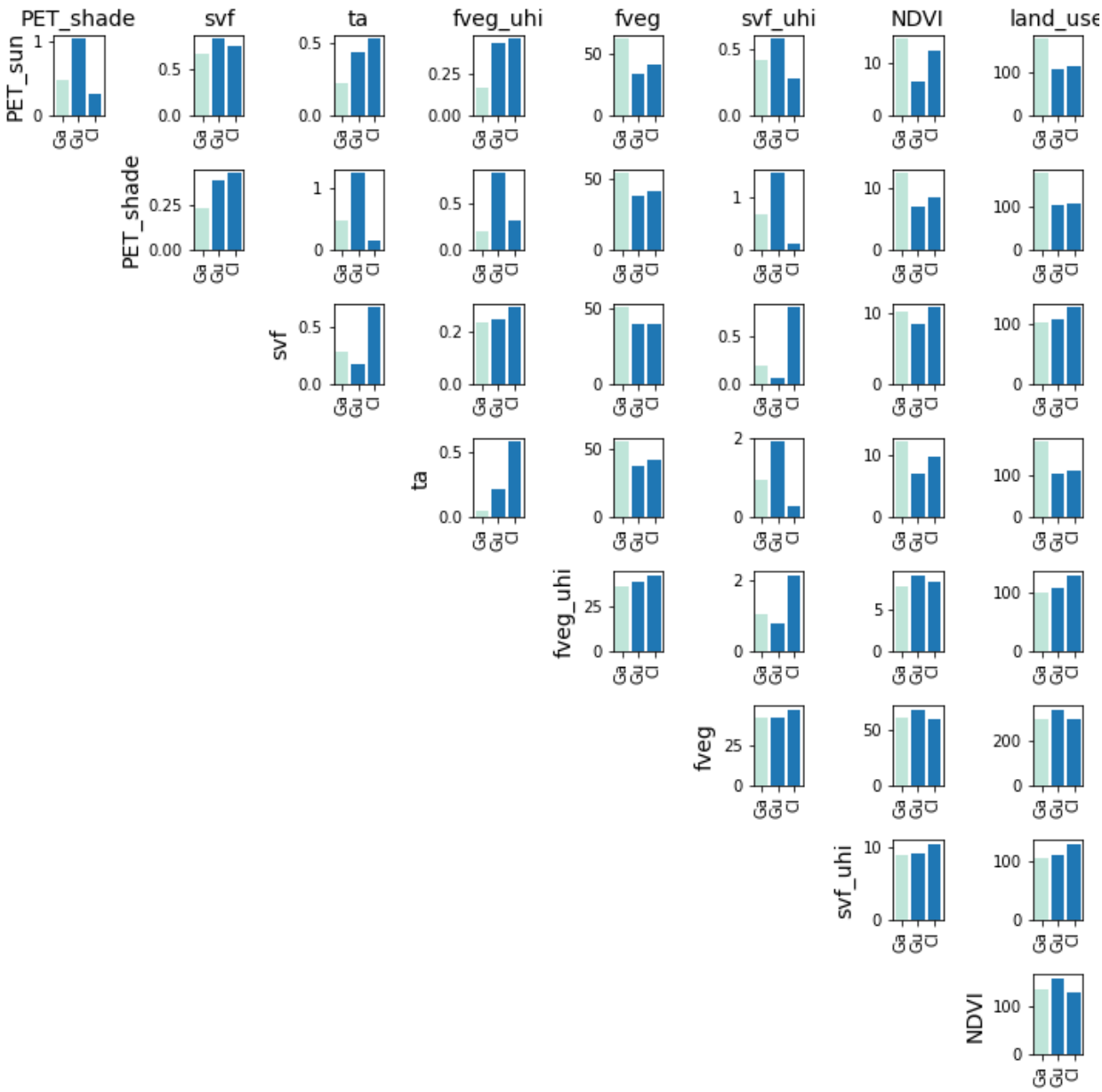


Figure 15: Visualisation CvM statistics run 03; three empirical copulas were fit to the data and compared to their theoretical equivalent: the Gaussian (Ga), Gumbel (Gu) and Clayton (Cl) copulas. For each variable combination the lowest bar indicates the best copula fit to represent that particular bivariate dependency.

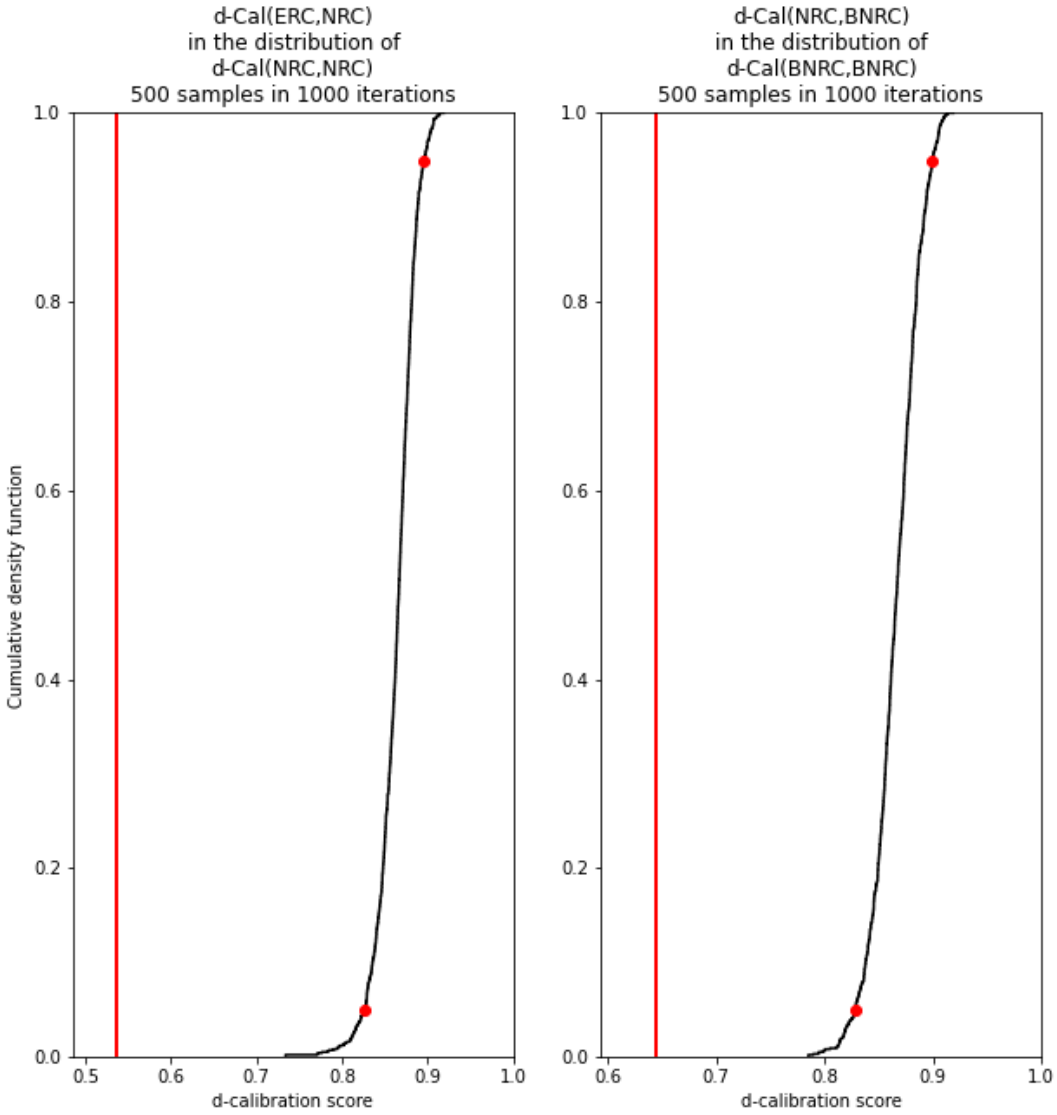


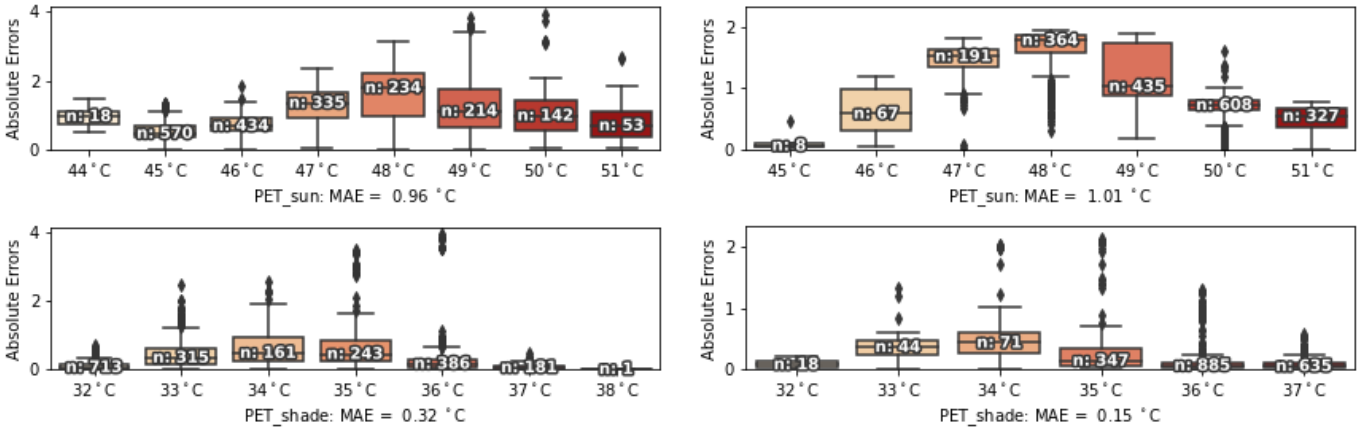
Figure 16: Visualisation Gaussian Distance (d-calibration scores) run 03, empirical rank corrrelation (ERC) on the left and Bayesian Network rank correlation (BNRC) on the right.

In Figure 16 the d-calibration scores for the empirical rank correlation (ERC) matrix and the Bayesian Network rank correlation (BNRC) matrix are shown. Both scores were outside the 90% confidence interval of the determinant of the empirical normal distribution, which is unsatisfactory. This was the case for all tests. However, the scores improved significantly for some tests in the following sections. Although for some runs particularly the ERC got close to the boundaries and also the BNRC for other runs, it was never within the 90% confidence interval for both. See Table 20 in Appendix E.

Table 8: MAE and CI values for 01) Wageningen Rural and 02) Wageningen City

	PET_{sun}				PET_{shade}			
	In-sample		Out-sample		In-sample		Out-sample	
Run ID	CI	MAE [°C]	CI	MAE [°C]	CI	MAE [°C]	CI	MAE [°C]
01	39%	0.96	NA	NA	70%	0.32	NA	NA
03	32%	1.01	NA	NA	91%	0.15	NA	NA

In Table 8 the MAE and CI results of both run 01 and run 03 can be found. In-sample they performed comparably in estimating PET_{sun}, but PET_{shade} was estimated better using the Wageningen City sample. A reason for this could be the fact that in run 03 the prevalence of shadow was more pronounced as this sample only used urban data. When comparing the absolute errors at an individual level, we saw a better performance for run 03 (Figures 17a and 17b). In Figure 26 (Appendix E) the difference in inference between both models is shown when predicting one urban sample point.



(a) Box plots absolute errors run 01

(b) Box plots absolute errors run 03

Figure 17: When comparing run 01, trained on Rural Wageningen, with run 03 (Wageningen City) we saw smaller absolute errors for the latter case. Notice the different scales on the y-axis

4.2 Out of sample: Rotterdam

Regarding the estimation of heat stress in a new city, the main finding is that this is possible at the cost of some larger absolute errors and a larger MAE value for PET_{sun} and PET_{shade} (respectively 1.15 and

0.62 °C). In Table 9 the performance of the test set in Rotterdam city from run 05 is shown. This run can be seen as the base-run, as it used the model from Figure 7 with 1 m resolution grid. Both MAE scores were worse for the out-of-sample data than for the in-sample data, as were the CI scores. However, outliers in individual absolute errors were in the same range as for the in-sample data, as can be seen when comparing Figure 18 with Figure 17b. Another thing that stood out, was the fact that both runs showed a lower median in absolute errors for the lowest and highest observations in the dataset compared to the more average observations. An import conclusion that could be drawn for this, is the fact that errors do not grow for observations in the tails of the distribution. Since the warmest places might be the most problematic, these are the most important observations from a practical point of view.

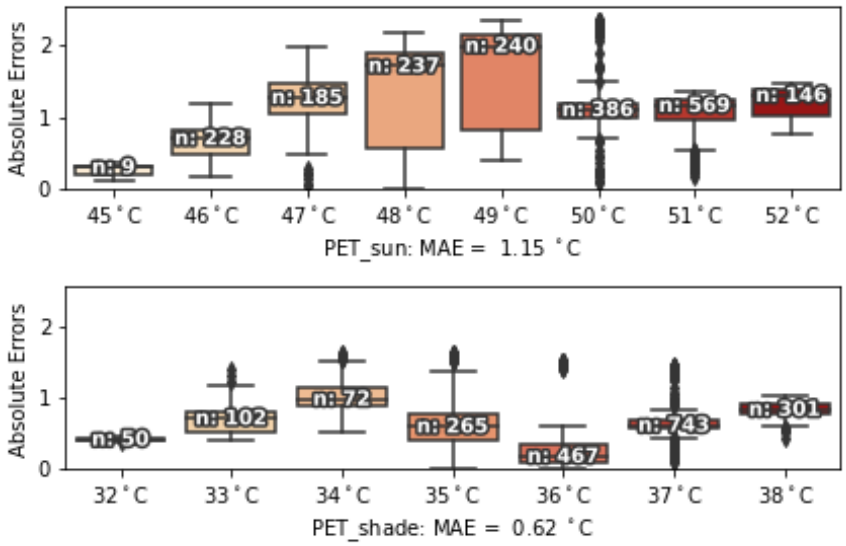


Figure 18: Box plots absolute errors run 05

Table 9: MAE and CI values for 05) base case model

	PET_{sun}				PET_{shade}			
	In-sample		Out-sample		In-sample		Out-sample	
Run ID	CI	MAE [°C]	CI	MAE [°C]	CI	MAE [°C]	CI	MAE [°C]
05	32%	1.01	13%	1.15	91%	0.15	25%	0.62

4.3 Coarseness grid

When comparing the impact of grid resolution in the model, the key take-away is that for each step to a coarser grid, the model performed equally good or better, at the cost of detailed information on a geographical

scale. This became very clear when looking at a small part of Wageningen city (in-sample) for run 05, 06 and 07. These runs respectively had a grid size of 1, 10 and 20 m. Run 05 and 07 are shown in Figure 19. The complete comparison can be found in Appendix E.

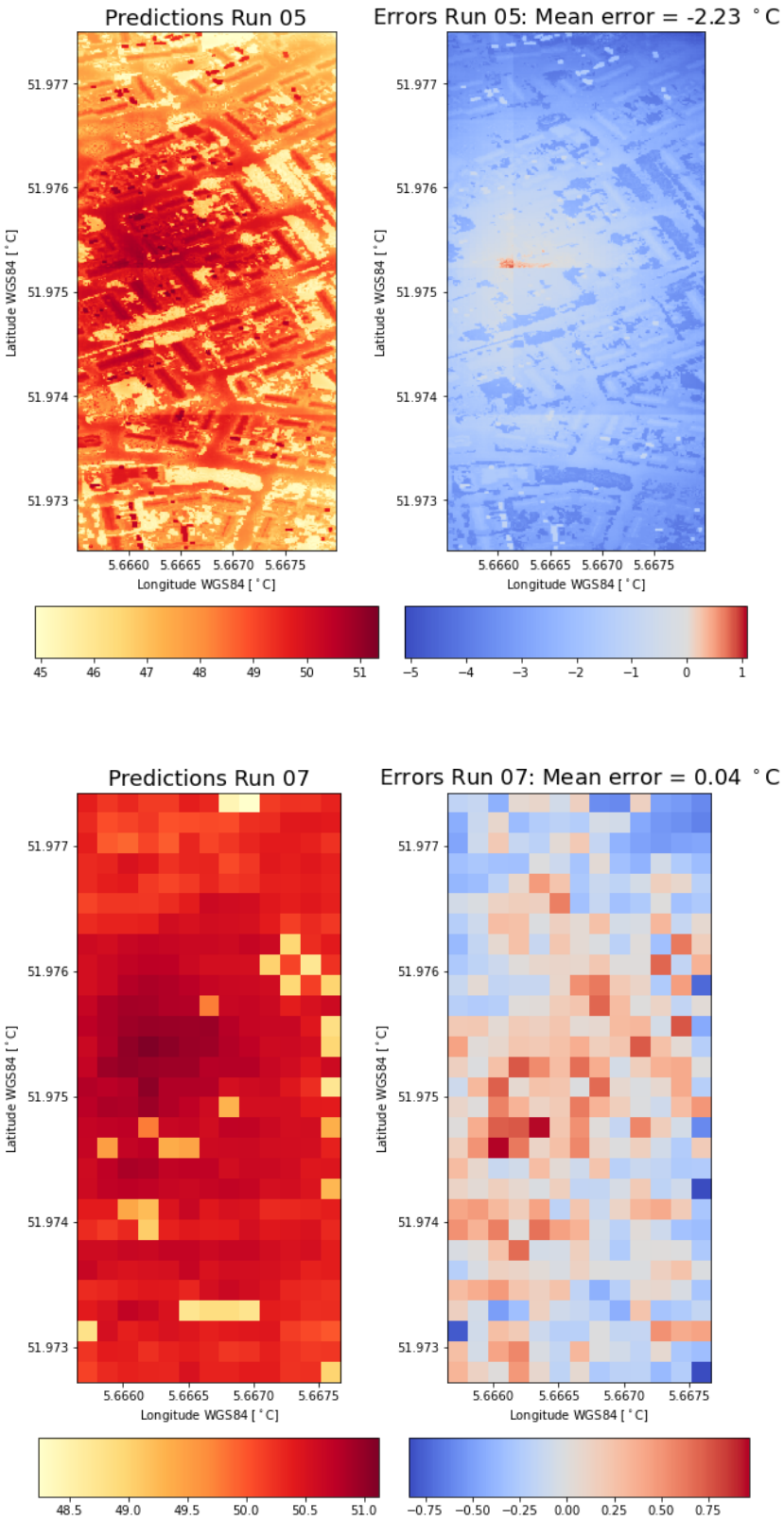


Figure 19: Comparison part of Wageningen city using a 1 m grid and a 20 m grid. The mean error was calculated only for the area visible in the image; although the mean error decreased with increasing coarseness, exact locations with heat concentration were less clearly visible

In Table 10 the performance for several grid sizes are shown. Run 05, 06 and 07 had a grid size of respectively 1, 10 and 20 m. For each step to a coarser grid, the model performed equally good or better, at the cost of detailed information on a geographical scale. Also two other scenarios were added. Run 08 was trained on 1 m resolution data, just like run 05. However, the out-of-sample test was done on data with a 10 m resolution, to emulate the realistic use case in which detailed data is available for training, but not for a new study area. We saw that the model did not score better than run 06, in which the test data also consisted of 10 m resolution, but with 10 m resolution training data. The same conclusion could be drawn when comparing the out-of-sample performance of run 09 with that of run 07.

Table 10: MAE and CI values for 05) 1 m 06) 10 m 07) 20 m 08) train 1 m/test 10 m 09) train 10 m/test 20 m

	PET_{sun}				PET_{shade}			
	In-sample		Out-sample		In-sample		Out-sample	
Run ID	CI	MAE [°C]	CI	MAE [°C]	CI	MAE [°C]	CI	MAE [°C]
05	32%	1.01	13%	1.15	91%	0.15	25%	0.62
06	58%	0.82	26%	1.04	92%	0.13	28%	0.56
07	67%	0.69	30%	0.96	92%	0.11	28%	0.54
08	32%	1.01	17%	1.16	91%	0.15	29%	0.56
09	58%	0.82	28%	1.04	92%	0.13	28%	0.55

A final comparison concerning the theory observations done in Section 4.1 was done with run 02 and 04. These two runs were the same as respectively run 01 and 03, but with a 10 m resolution grid instead of 1 m. This resulted in the same conclusions as in Section 4.1.

4.4 Missing data layers

We investigated what data layers from the base model are crucial to have in estimating PET_{sun} and PET_{shade}. The key finding in this section is that the model improved significantly compared to the base model after removal of each data layer containing a binary random variable. The sky view factor S_{vf} appeared to be especially important during training to get good results for PET_{sun}. Replacing the sky view factor with different configurations of the AHN2 layer could not solve this. The NDVI layer however seemed to be able to perfectly replace both vegetation layers F_{veg} and $F_{veg,UHI}$ and after removing *land use* the model performed even better. Finally information about the actual temperature layer T_a was also removed

during prediction, which resembles a potential use case of using a trained model with T_a in a new city for which the ECDF of T_a is not known. Consequently we obtained out-sample MAEs of 1.55 °C and 1.20 °C for PET_{sun} and PET_{shade} respectively, by only using $NDVI$ and a $AHN2$ layer from Rotterdam (run 22, Figure 20). This result really shows the added value of a NPBN compared to the current model, especially in areas with less data availability.

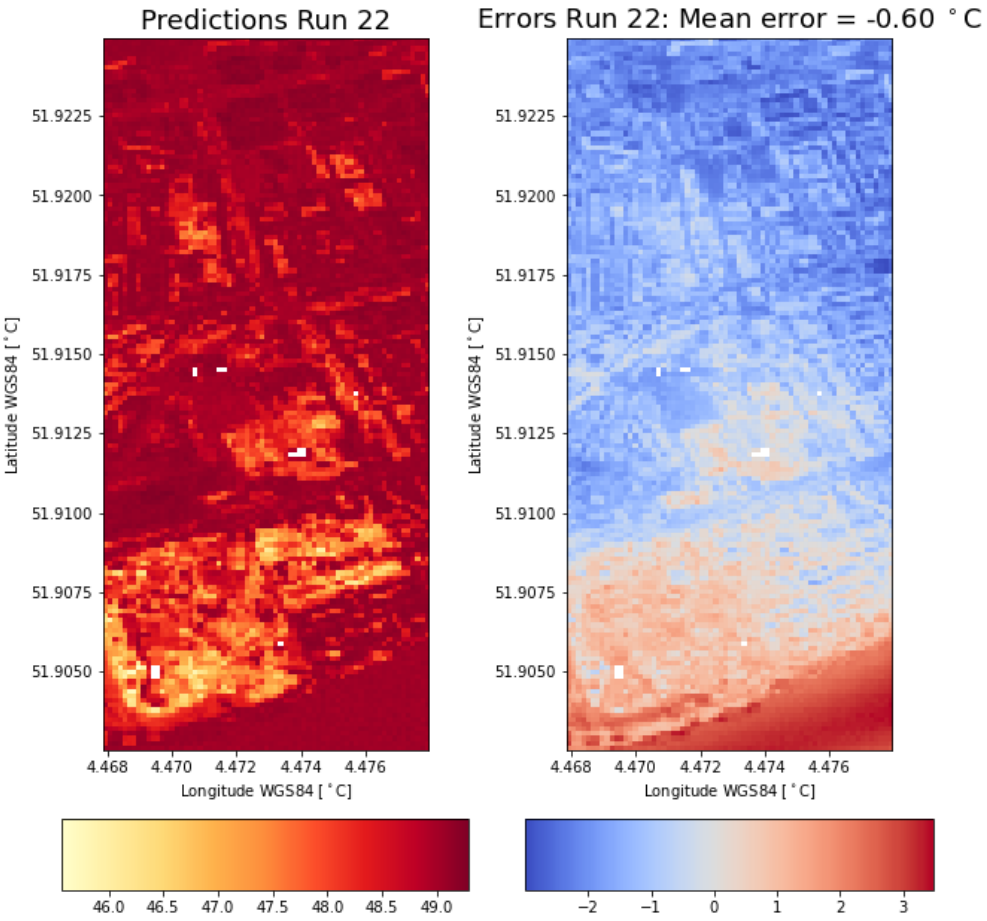
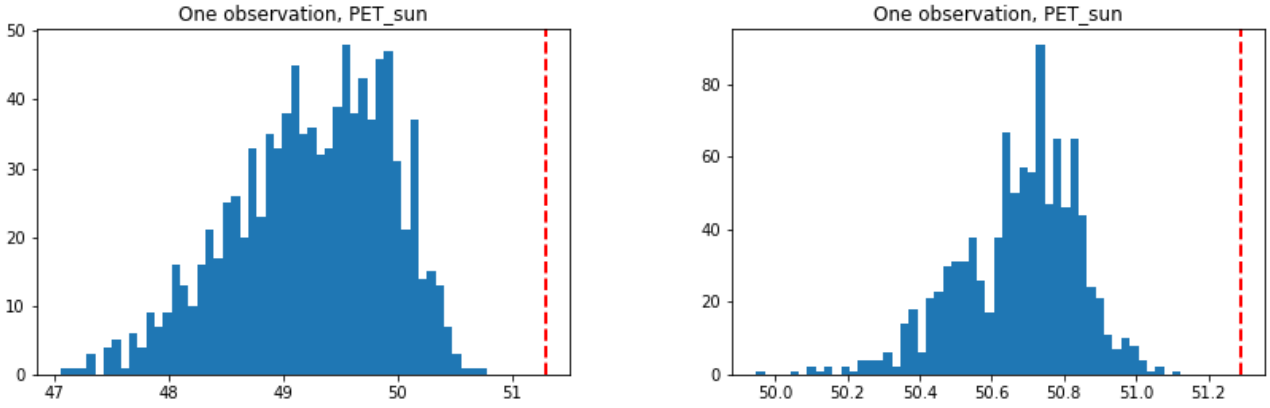


Figure 20: Out sample predictions and errors of run 22, only using information from Rotterdam regarding two layers: $NDVI$ and $AHN2$. For training on Wageningen city information on these two layers and information on T_a, PET_{sun} and PET_{shade} was used. The mean error was calculated only for the area visible in the image

4.4.1 Replacing sky view factor with $AHN2$ data

We replaced both S_{vf} and $S_{vf,UHI}$ with $AHN2$ in three steps, resulting in the performances shown in Table 11. Run 10 shows the results of training the NPBN with information on S_{vf}, but testing on Rotterdam without using information about S_{vf} over there. When comparing the MAE results with the base model in run 05, we saw that although the MAE values stayed the same, all of the four CI values are larger. This could be explained by looking at the inference plots of both runs for PET_{sun} of one observation (Figures 21a

and 21b). We saw a larger spread in the estimates of PET_{sun} for run 10, which could be explained by the direct connection in the DAG between S_{vf} and PET_{sun}. As we removed a 'piece of evidence (S_{vf})' during inference, this information could not be used in our estimates, resulting in a larger spread of estimates. The same happened for PET_{shade}.



(a) Inference one observation PET_{sun} run 10

(b) Inference one observation PET_{sun} run 05

Figure 21: Removing S_{vf} during out-sample prediction caused a much larger spread to occur in the inferred estimates, as could be seen when comparing the left figure (without S_{vf}) with the right figure (base model)

Table 11: MAE and CI values for different sky view settings

	PET_{sun}				PET_{shade}			
	In-sample		Out-sample		In-sample		Out-sample	
Run ID	CI	MAE [°C]	CI	MAE [°C]	CI	MAE [°C]	CI	MAE [°C]
10	61%	1.04	48%	1.17	97%	0.20	62%	0.60
11	60%	1.04	48%	1.17	97%	0.20	62%	0.61
14	55%	1.18	39%	1.31	95%	0.19	61%	0.58

In run 11 the $AHN2$ layer was added during training and testing to see if any correlation existed between $AHN2$ and S_{vf} that could compensate for the missing 'evidence' on S_{vf} in the test data, during inference. This resulted in very similar results to run 10, which was an indication that $AHN2$ did not add information when S_{vf} is not known. Finally, run 14 completely excluded both S_{vf} and $S_{vf,UHI}$ and trained and tested the NPBN using only $AHN2$ as a sky view-related variable. This resulted in worse performances, especially with respect to PET_{sun}. This result emphasised the importance of the sky view layers to estimate PET_{sun}.

Instead of the raw $AHN2$ data, runs were done using either $\hat{\sigma}_{AHN2}$ or $AHN2 - \hat{\mu}_{AHN2}$. These layers were derived from $AHN2$ to incorporate more of the variation in height in the surrounding of our area of estimation, leading to a result more comparable to the sky view factor from a physical point of view. However, both newly derived layers gave similar results as $AHN2$. These results (run 12, 13, 15 and 16) can be found in Table 21 in Appendix E. The visualisation of the NPBN of run 16 shows small rank correlation coefficients (0.09 and -0.09) between $AHN2 - \hat{\mu}_{AHN2}$ and respectively PET_{shade} and PET_{sun}, confirming the fact that the $AHN2$ data layer on its own did not inherit sufficient information to replace the sky view factors.

4.4.2 Removing vegetation random variables

We removed the layers F_{veg} and $F_{veg,UHI}$ in three steps. In run 17, F_{veg} and $F_{veg,UHI}$ were available during training, but not during testing of the NPBN. The PET_{sun} results were better than the base case in run 05, as can be seen in Table 12. When we left out F_{veg} completely (run 18), we see little worse results for PET_{sun}. After completely removing both F_{veg} and $F_{veg,UHI}$ in run 19, we obtained similar results as run 18. However, the results of all three runs were better than run 05 (base case). As results in Section 4.1 showed bad copula fits for the binary random variables, a better performance when removing F_{veg} makes perfect sense.

Table 12: MAE and CI values for different vegetation settings

	PET_{sun}				PET_{shade}			
	In-sample		Out-sample		In-sample		Out-sample	
Run ID	CI	MAE [°C]	CI	MAE [°C]	CI	MAE [°C]	CI	MAE [°C]
17	95%	0.50	45%	0.96	91%	0.15	25%	0.62
18	84%	0.59	36%	1.10	91%	0.15	25%	0.62
19	81%	0.60	35%	1.09	91%	0.12	29%	0.58

4.4.3 Refining *land use* layer

For the same reason mentioned above (a better performance of the model when removing the binary variable F_{veg}), we also explored ways to deal with the second binary variable *land use*. In runs 19a and 19b we split the model from run 19, see Table 13. Run 19a only modelled the observations where *land use* equals one. This value indicates either a water surface or a surface with a function that implies vegetation (agriculture, grass or forest). This concerned 142 observations in the training set and 212 observations in the validation

set. Run 19b was used to model observations for which *land use* equals zero, indicating buildings or other surfaces that do have a low evaporation rate. Especially for run 19b we saw an overall improvement compared to run 19, which indicates that splitting up data according to their values within a binary variable may improve the overall performance of a NPBN. A condition to do this, was that this binary random variable was also available as a conditional layer in the test set. Otherwise one would not have been able to split the data correctly over the two trained models.

Run 19c was done using an ordinal *land use* split, in an attempt to come closer to a continuous distribution. We split the data based on the evaporation rate, resulting in five classes increasing in evaporation rate: 'buildings', 'unpaved surface', 'water', 'grass/agriculture' and 'trees'. Although this model improved the in-sample performance for PET_{sun}, it did reduce the out-sample performance. This was possibly caused by the uneven distribution over these 5 classes, with still more than 1,300 of the 2,000 variables in the first class 'buildings'. Another run was done in which the binary random variable *land use* was removed completely. This resulted in better performance for PET_{sun} (run 20, Table 12).

Table 13: MAE and CI values for different *land use* refinements. 20) is without *land use*

	PET_{sun}				PET_{shade}			
	In-sample		Out-sample		In-sample		Out-sample	
Run ID	CI	MAE [°C]	CI	MAE [°C]	CI	MAE [°C]	CI	MAE [°C]
19	81%	0.60	35%	1.09	91%	0.12	29%	0.58
19a	82%	0.25	12%	0.93	95%	0.09	26%	0.65
19b	90%	0.42	44%	0.76	91%	0.07	29%	0.55
19c	93%	0.49	12%	1.00	91%	0.12	29%	0.58
20	94%	0.48	47%	0.96	91%	0.12	29%	0.58

4.4.4 Using unprocessed raw data layers: $AHN2$ and $NDVI$

In run 21 we used $AHN2$ and $NDVI$, leaving out F_{veg}, $F_{veg,UHI}$, *land use*, S_{vf} and $S_{vf,UHI}$. We compared the results with the previous run and base-case run 05 (Table 14). Compared with run 05, all CI values improved and the MAE values improved or stay roughly the same. Compared with run 20, there was a trade-off between little lower MAE values and larger CI values. Also this was the first run in which both in-sample CIs were 95% (or larger). A remarkable observation is the improvement of the out-sample CI from 29% to 61%, while keeping the same MAE value.

Table 14: MAE and CI values for 21) using $AHN2$ and $NDVI$, leaving out F_{veg}, $F_{veg,UHI}$, *land use*, S_{vf} and $S_{vf,UHI}$ 20) using $NDVI$, S_{vf} and $S_{vf,UHI}$ 05) base case

	PET_{sun}				PET_{shade}			
	In-sample		Out-sample		In-sample		Out-sample	
Run ID	CI	MAE [°C]	CI	MAE [°C]	CI	MAE [°C]	CI	MAE [°C]
21	96%	0.56	68%	1.04	95%	0.19	61%	0.58
20	94%	0.48	47%	0.96	91%	0.12	29%	0.58
05	32%	1.01	13%	1.15	91%	0.15	25%	0.62

When comparing these three runs regarding the theoretical performance measures (Table 15), we saw the best BNRC d-calibration score for run 20 and the best ERC d-calibration score for run 21.

Table 15: d-calibration scores using 21) raw input data layers 20) raw NDVI and sky view 05) base case

	BNRC			**ERC**		
Run ID	5th percentile	95th percentile	d-calibration score	5th percentile	95th percentile	d-calibration score
21	0.90	0.96	0.75	0.90	0.96	0.78
20	0.87	0.94	0.85	0.87	0.94	0.66
05	0.83	0.90	0.64	0.83	0.90	0.54

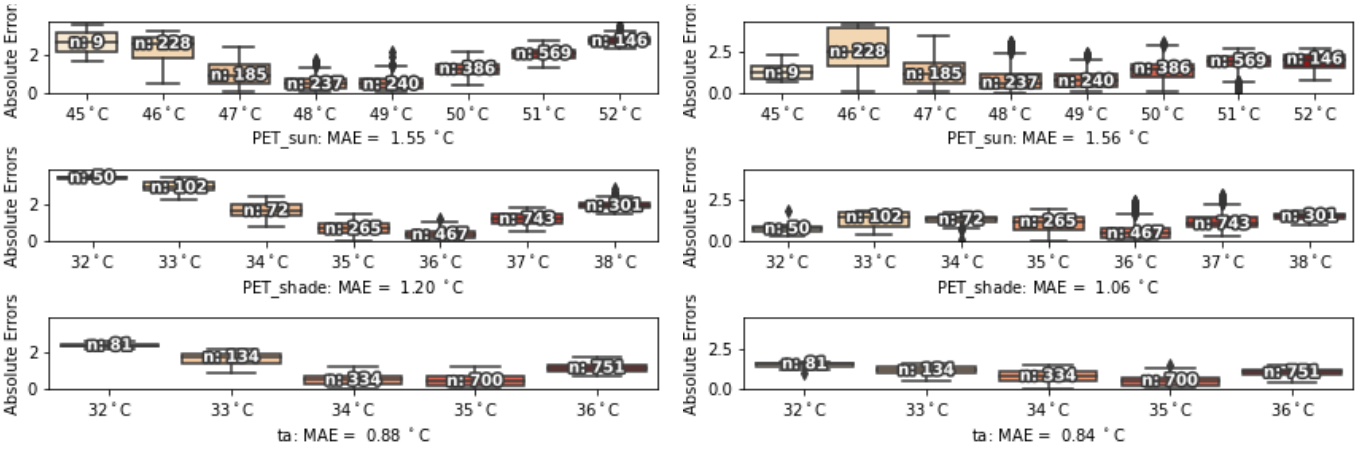
4.4.5 Changing T_a from conditionalising variable to prediction variable

All shown NPBN visualisations showed a strong conditional rank correlation between T_a and both PET_{sun} and PET_{shade}. In all previous tests T_a was regarded as a known variable we could conditionalise on. That is, until this point we always used T_a as an evidence variable during inference in a new city. However, as could be seen in these NPBN visualisations, T_a is directly dependent on $F_{veg,UHI}$ and $S_{vf,UHI}$ and indirectly on F_{veg}, *land use* and S_{vf} as well. If we treat T_a as a given (evidence) variable, we implicitly use information of other variables we treat as being unknown. As these variables also have other paths to influence both PET_{sun} and PET_{shade} these tests were all useful, especially to compare them to each other. However, if we want to know what run performs best without implicitly still using information from nodes we want to exclude, we should also treat T_a as a node we want to predict. We did this for the two best runs (21 and 20) and for the base case (run 05), resulting in respectively run 22, 23 and 24. In Table 16 the results of these runs can be found. The base case (run 24) had the lowest (best) out-of-sample MAE scores for all three variables to predict. However, it also had the lowest (worst) CI scores and the worst performance when looking at theory related scoring

measures (run 05 in Table 15). Run 22 (with only unprocessed raw data layers) performed similarly to run 23 (with sky view but without F_{veg} and $F_{veg,UHI}$). However, when looking closer at the individual absolute errors, the observations in the tails of the distribution were better estimated for run 23 (Figures 22a and 22b).

Table 16: MAE and CI values, also predicting T_a for 22) using $AHN2$ and $NDVI$, leaving out F_{veg}, $F_{veg,UHI}$, *land use*, S_{vf} and $S_{vf,UHI}$ 23) using $NDVI$, S_{vf} and $S_{vf,UHI}$ 24) base case

	PET_{sun}				PET_{shade}				T_a			
	In-sample		Out-sample		In-sample		Out-sample		In-sample		Out-sample	
Run ID	CI	MAE [°C]	CI	MAE [°C]	CI	MAE [°C]	CI	MAE [°C]	CI	MAE [°C]	CI	MAE [°C]
22	98%	0.80	67%	1.55	96%	0.67	79%	1.20	96%	0.47	77%	0.88
23	96%	0.69	62%	1.56	98%	0.46	65%	1.06	99%	0.35	66%	0.84
24	40%	1.00	17%	1.13	91%	0.16	25%	0.65	93%	0.08	20%	0.41

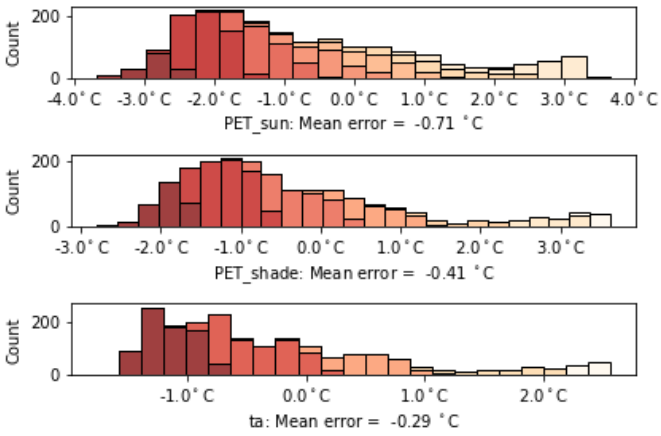


(a) Box plots absolute errors run 22

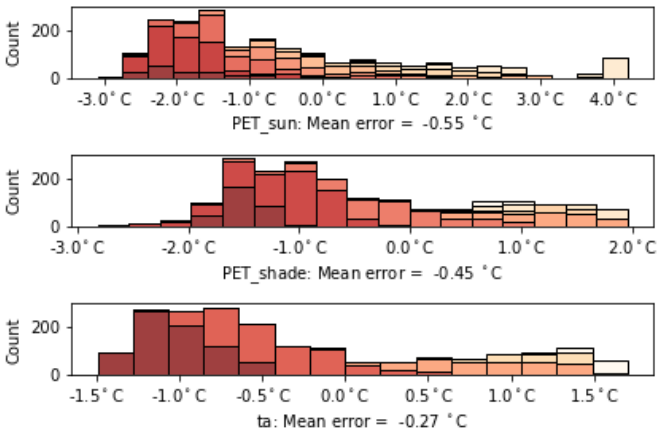
(b) Box plots absolute errors run 23

Figure 22: When comparing run 22, only using $NDVI$ and $AHN2$ as evidence, with run 23, also using S_{vf}, we saw smaller absolute errors for the latter case when comparing the tail observations

When plotting the distribution of the errors of run 22, we came to another important conclusion. We already had seen that temperatures were estimated to be colder than observed in a large part of Rotterdam city in Figure 20. In Figure 23a this was confirmed; errors are grouped and colour coded by the observed temperature. The errors above +2 °C, representing the observations in the river Nieuwe Maas seemed to introduce a bimodal distribution in this error plot, indicating the NPBN models water bodies less adequately.



(a) Distribution of the out-sample errors (Rotterdam) of run 22



(b) Distribution of the out-sample errors (Rotterdam) of run 23

Figure 23: When comparing run 22, only using $NDVI$ and $AHN2$ as evidence, with run 23, also using S_{vf}, we saw observed temperatures in the warmest places (darkest red) were underestimated, while temperatures in the coolest places (white) were overestimated for run 22. By incorporating the sky view factor (run 23), the errors in the tails have less extreme values than those in run 22

Figure 23b shows that in run 23, with sky view factor, the range of errors became smaller and also some of the extreme warm observations only had errors of around 1 °C, confirming the claim made above that this run estimated the extreme values in the tails of the distribution more adequately.

4.5 Number of training observations

Using the NPBN model with only unprocessed raw data layers (as in run 21), we created three models with a fewer number of training observations from the Wageningen City sample. Instead of using 2,000 observations, we used 200, 50 and 20 observations in respectively runs 25, 26 and 27. In Table 17 can be seen that the number of training observations did not have any significant impact on the performance of the model. Also when comparing the theoretical performance measures in Table 18, no prevalent differences were visible. However, these tests were only done for the smallest NPBN in terms of number of nodes. Sensitivity to the number of observations could have been higher if we had tested it on the more elaborate models as well (see Section 5).

Table 17: MAE and CI values, also predicting T_a using using 25) 200 training observations 26) 50 training observations 27) 20 training observations; these models have the same DAG as run 22 (2,000 training observations)

	PET_{sun}				PET_{shade}				T_a			
	In-sample		Out-sample		In-sample		Out-sample		In-sample		Out-sample	
Run ID	CI	MAE [°C]	CI	MAE [°C]	CI	MAE [°C]	CI	MAE [°C]	CI	MAE [°C]	CI	MAE [°C]
25	98%	0.81	69%	1.55	96%	0.66	79%	1.19	96%	0.47	78%	0.88
26	98%	0.82	73%	1.55	95%	0.68	77%	1.21	96%	0.47	74%	0.88
27	97%	0.83	70%	1.56	96%	0.67	76%	1.22	95%	0.47	71%	0.89

Table 18: d-calibration scores using 22) 2,000 training observations 25) 200 training observations 26) 50 training observations 27) 20 training observations. Using 200 training observations gave the best results for both metrics, with both d-calibration scores being the nearest to the 5^{th} percentile

	BNRC			**ERC**		
Run ID	5th percentile	95th percentile	d-calibration score	5th percentile	95th percentile	d-calibration score
22	0.90	0.96	0.75	0.90	0.96	0.78
25	0.90	0.96	0.77	0.90	0.96	0.83
26	0.90	0.96	0.73	0.90	0.96	0.78
27	0.90	0.96	0.73	0.90	0.96	0.73

4.6 Predict NDVI based on PET values

An important application of the urban heat stress tool is to determine the influence of possible NBSs on the PET values. Therefore we conditionalised on the PET values to see if we could predict the $NDVI$ value. In run 28 we still predicted $F_{veg,UHI}$ alongside, but in run 29 we only used $NDVI$ as a vegetation related factor. The results can be found in Table 19. We observed relatively low MAE values for both runs, but also the CI values are low. Also one should note that a MAE of 0.17 in $NDVI$, which has a range from 0 to 1, only gives information as soon as the NDVI can be coupled to a specific type of vegetation. This aspect was out of the scope of our research.

Table 19: MAE and CI values when conditionalising on the PET values

	T_a				$NDVI$				$F_{veg,UHI}$			
	In-sample		Out-sample		In-sample		Out-sample		In-sample		Out-sample	
Run ID	CI	MAE [°C]	CI	MAE [°C]	CI	MAE [-]	CI	MAE [-]	CI	MAE [-]	CI	MAE [-]
28	93%	0.10	35%	0.44	96%	0.06	71%	0.17	93%	0.03	37%	0.11
29	93%	0.07	36%	0.41	96%	0.07	71%	0.17	NA	NA	NA	NA

5 Discussion

During this research, we showed how a NPBN can be used to assess urban heat stress. This choice for this model was based on a knowledge gap in the current literature in the form of data availability. This form of Bayesian Networks can model non-linear relationships between variables while keeping both computational and memory costs relatively low. Although a lot of models are available to assess heat stress, we showed that an adequately trained NPBN can predict both PET_{sun} and PET_{shade} with Mean Absolute Errors (MAE) of respectively 1.6 °C and 1.2 °C in an out-sample city (Rotterdam), using only $NDVI$ and $AHN2$ layers of this out-sample city. Respectively 67% and 79% of the observations lay within the 95% confidence interval constructed by this NPBN. The present study produces two main findings, which are discussed below.

Firstly, we showed that a representative training sample is key in obtaining a NPBN with good theoretical and prediction performance. We found that the number of observations used in this training sample was less important. Using only 20 sample points resulted in the same performance as using 2,000 sample points. The NPBN in our study can help in standardising the assessment of heat stress outside the borders of the Netherlands, since a standardised method to assess heat stress is currently not globally available. Another advantage of the NPBN in our study is that the physics behind the network is transparent, which is not usual for data-driven models. Even expert knowledge on physics can be included in adjusting the NBPN to a specific context. Further research should focus on implementing our network in a different climate, through for example changing the meteorological variables that were constant in our study. Because as mentioned above, the success of applicability of a NPBN depends on a representative training sample. This can be seen when looking more closely to the differences in training on only Wageningen City data versus training on the Wageningen Rural sample, including both urban and rural data points (Section 4.1). A result that needs to be addressed from this section, is the fact that the conditional rank correlation between S_{vf} and PET_{shade} changed from negative to positive (Figures 12 and 14). When looking at the relationship in Equation 2, the latter term implies a positive conditional rank correlation between both variables, as was observed in the NPBN quantification of the Wageningen City sample (Figure 14). However, the Wageningen Rural sample gave a negative conditional rank correlation between both variables (Figure 12). This can be explained by the fact that this dataset contains both rural and urban data points. When comparing the distributions of S_{vf} from Figures 3a and 3b, rural observations in general have a larger sky view factor than urban observations. Part of the variance in the observed variable PET_{shade} that cannot be explained by the model, is addressed to a variable that directly influences this PET_{shade}, namely the variable S_{vf}. However, for the rural dataset this conditional rank correlation now not only contains the effect of S_{vf} on PET_{shade}, but also the effect of rural vs city data points, which is correlated with S_{vf}. This phenomenon is similar to the problem of

'omitted variable bias' in linear regression and stresses the importance of a representative training sample.

Secondly, we found that our NPBN can handle missing data layers up to a point in which only $NDVI$ and $AHN2$ layers are needed during prediction of expected heat stress in a new city. The network performs better at coarser grid resolutions up to 20 m, at the cost of the level of detail in the output. On the contrary, the current model of Nelen & Schuurmans is restricted to using multiple data layers at a 1 meter resolution. Ruangpan, Vojinovic, Di Sabatino et al. (2020) argued that knowledge is missing regarding the trade-off that needs to be made between high-resolution numerical models and the accuracy of results in NBS tools. However, we showed in this study that this trade-off can be avoided since our network can handle missing data layers. As one of the input layers, $NDVI$, is being influenced by most of the NBS solutions counteracting heat stress, this model could therefore be a framework to assess the effect of potential NBS solutions within a city. In the following two paragraphs we discuss some specific data layers and their effect on the model.

Regarding the sky view factor layer, we saw that including this layer generally led to better estimates in the tails of the PET distribution. When looking at the presented MAE box plots in Section 4, we saw for runs including the sky view factor that the extremer observations on both sides (coolest and warmest places) were estimated with lower absolute errors than the more frequent average observations. This is useful as especially the warmest spots are areas of interest regarding mitigation measures, whereas the coolest spots may be areas to learn from.

Both *land use* and F_{veg} can be seen as discrete ordinal random variables consisting of two values. According to Hanea, Kurowicka and Cooke (2007), these variables can be incorporated in a NPBN, as long as they are not dominating the DAG. Although we only used binary discrete ordinal variables, the model seemed to perform better without those variables. In relation to the claim made by Hanea, Kurowicka and Cooke (2007), it therefore seems to be important to distinguish between binary random variables and discrete ordinal variables with more than two values.

5.1 Reflection on methodology and limitations

Regarding our methodology, we took average values when resampling the datasets to coarser grids, except for *land use* (we used the mode). We did not test whether other resampling options are better. However, one should keep in mind the reason why coarse grids are tested; not only to speed up the process when data is available, but especially for areas when a fine grid is not available. In that case we cannot choose our resampling method and probably averaging would most closely resemble this situation.

A more general limitation regarding the validation method, is the lack of different benchmark models. We only compared our model to 'observed data' that had been obtained from the current standard method. However, the source data this standard method had been calibrated on, was not available for this research. Another approach to assess this particular dependence model, the NPBN, would be to compare the results of the NPBN with another statistical model, like a random forest. This was not in the scope of the research, but would be a good focus for further research. However, when using different dependence models, one should keep in mind one of the main advantages of using a NPBN. The current model is limited to heat stress, but as co-benefits are important in assessing urban mitigation measures, it makes sense to choose a model that incorporates the relationships between different phenomena based on some physical background. Qualification of a NPBN by drawing a DAG is exactly doing this and therefore it may be easier to expand a NPBN than e.g. a random forest, as this latter model does not take into account any physical aspects by construction.

One of the main limitations of the NPBN model, is the fact that marginal distributions of nodes can differ in different cities. Even when all variables that were assumed not to differ geographically, are indeed constant, we saw that adding information about the CDFs of the predicted nodes in Rotterdam improved the model significantly, up to CI values of 95% for this out of sample location as well. However, this also is an opportunity; when expert knowledge on PET ranges is available for a specific location, one can use this to update the marginal distribution for this new region.

Another limitation that is closely related to this, is the fact that a lot of variables of the original model cancelled out in the NPBN model as there was no geographical variability in these variables. If one of the variables (like humidity) differs for a new region, this variable should be added to the NPBN. A small nuance should be made to this; using the built NPBN model, we can still distinguish between the relatively cooler and warmer spots within an urban area if a variable is different in Wageningen, but constant again in the new area. This condition of a variable being constant is important, as for example coastal cities may deal with varying wind speeds geographically. Then the assumption of equal wind speed along the whole grid would not be valid anymore.

With this research we focused on quantifying heat stress during the current one-in-thousand summer day in the Netherlands, specifically in Wageningen and Rotterdam. We included several relevant variables, of which some were related to vegetation. Changing the variable $NDVI$ manually may be a starting point to assess the effect of new NBS implementations in the urban area when considering for example green

corridors. Although we particularly focus on heat stress in our study, including precipitation in the same NPBN could be a logical next step in quantifying the current situation and investigating how a NBS could influence this. We created a NPBN framework that could be helpful in mapping multiple co-benefits of one NBS, as it incorporates the dependency between multiple co-benefits by its construction. Further research on applying NPBN applications is needed and to ensure that this is possible, we open-sourced the Python package BANSHEE-y we built the NPBN in. This code is based on BANSHEE (Paprotny, Morales-Nápoles et al., 2020).

6 Conclusions

In this study, we examined the applicability of a Non-Parametric Bayesian Network (NPBN) to assess the influence of spatial surfaces on urban heat stress. By training a NPBN on data of Wageningen and testing our NPBN on data of Rotterdam, we considered both theoretical applicability and prediction performance in a new city. After confirming that the theoretical framework of NPBNs is suitable to quantify urban heat stress given a representative training sample, we explored how limited data availability would affect the performance. We assessed missing data layers, coarser resolution data layers and a limited number of observations available to train the model. An adequately trained NPBN predicted the Physiological Equivalent Temperature (PET) for sunny and shadow conditions with mean absolute errors of respectively 1.6 °C and 1.2 °C, only using $NDVI$ and $AHN2$ layers during prediction.

To our knowledge, this is the first research on how to cope with missing data in the context of assessing heat stress. Also in the field of NPBNs it is a first attempt to develop a model that could potentially be used globally. This can be achieved by quantifying the effect of using low-resolution data on the model, up to a resolution of 20 m, in the range of globally available datasets like SPOT, Sentinel and LANDSAT.

With this research we focused on quantifying heat stress during the current one-in-thousand summer day in the Netherlands, specifically in Wageningen and Rotterdam. An important conclusion is found in the direct applicability of a NPBN in a new urban area after being trained in an urban area exposed to a different climate. Although we obtained good results using the model trained on Wageningen data for predictions in Rotterdam, we saw an improvement in these predictions when using some information on the marginal PET distributions in this new city. Although relative heat stress is mapped accurately, future research should focus on how to obtain a marginal distribution of the PET in a new area, e.g. by using expert knowledge.

An important recommendation is regarding the extent of the model. Although we particularly focused on heat stress in our study, including precipitation in the same NPBN could be a logical next step in quantifying the current situation and investigating how a NBS could influence this. We created a NPBN framework that could be helpful in mapping multiple co-benefits of one NBS, as it incorporates the dependency between multiple co-benefits by its development. Further research on applying NPBN applications is needed and to ensure that this is possible, we open-sourced the Python code we built the NPBN in, under the name BANSHEE-y.

References

Adaptation Support Tool (2020). URL: https://crctool.org/ (visited on 06/10/2020).

Ale, B., Bellamy, L., Cooke, R., Duyvis, M., Kurowicka, D., Lin, P., Morales-Nápoles, O., Roelen, A. and Spouge, J. (2008). "Causal model for air transport safety (Final Report ISBN 10: 90 369 1724-7)". In: *Rotterdam, The Netherlands: Ministerie van Verkeer en Waterstaat*.

Alves, A., Gersonius, B., Kapelan, Z., Vojinovic, Z. and Sanchez, A. (2019). "Assessing the Co-Benefits of green-blue-grey infrastructure for sustainable urban flood risk management". In: *Journal of environmental management* 239, pp. 244–254.

Alves, A., Gersonius, B., Sanchez, A., Vojinovic, Z. and Kapelan, Z. (2018). "Multi-criteria approach for selection of green and grey infrastructure to reduce flood risk and increase CO-benefits". In: *Water Resources Management* 32.7, pp. 2505–2522.

Arifwidodo, S. D. and Chandrasiri, O. (2020). "Urban heat stress and human health in Bangkok, Thailand". In: *Environmental Research* 185, p. 109398. ISSN: 0013-9351. DOI: https://doi.org/10.1016/j.envres.2020.109398. URL: https://www.sciencedirect.com/science/article/pii/S0013935120302917.

Bornstein, R. D. (1968). "Observations of the urban heat island effect in New York City". In: *Journal of Applied Meteorology* 7.4, pp. 575–582.

CAIT Climate Data Explorer (2020). URL: https://cait.wri.org/ (visited on 06/10/2020).

Cheeseman, P. (1988). "An inquiry into computer understanding". In: *Computational Intelligence* 4.2, pp. 58–66.

Climate and Disaster Risk Screening Tools (2020). URL: https://climatescreeningtools.worldbank.org/ (visited on 06/10/2020).

Climate Change Knowledge Portal (2020). URL: https://climateknowledgeportal.worldbank.org/ (visited on 06/10/2020).

CREAT Climate Scenarios Projection Map (2020). URL: https://epa.maps.arcgis.com/apps/MapSeries/index.html?appid=3805293158d54846a29f750d63c6890e (visited on 06/10/2020).

De Nijs, T. et al. (2019). *Ontwikkeling Standaard Stresstest Hitte*. Rijksinstituut voor Volksgezondheid en Milieu (RIVM). DOI: 10.21945/RIVM-2019-0008.

Delgado-Hernández, D.-J., Morales-Nápoles, O., De-León-Escobedo, D. and Arteaga-Arcos, J.-C. (2014). "A continuous Bayesian network for earth dams' risk assessment: an application". In: *Structure and Infrastructure Engineering* 10.2, pp. 225–238. DOI: 10.1080/15732479.2012.731416. URL: https://doi.org/10.1080/15732479.2012.731416.

Dirksen, M., Ronda, R., Theeuwes, N. and Pagani, G. (2019). "Sky view factor calculations and its application in urban heat island studies". In: *Urban Climate* 30, p. 100498. ISSN: 2212-0955. DOI: https://doi.org/

10.1016/j.uclim.2019.100498. URL: https://www.sciencedirect.com/science/article/pii/S2212095519300604.

EcoActuary Tool (2020). URL: http://www.policysupport.org/ecoactuary (visited on 06/10/2020).

European Commission (2020). *Nature-Based Solutions for Flood Mitigation and Coastal Resilience. Analysis of EU-funded Projects*. European Commission. URL: https://op.europa.eu/en/publication-detail/-/publication/d6e80dca-d530-11ea-adf7-01aa75ed71a1/language-en/format-PDF/source-148331333.

Eurostat and OECD (2014). *Survey of National Practices in Estimating Net Stocks of Structures*. URL: https://ec.europa.eu/eurostat/documents/24987/4253483/Eurostat-OECD-survey-of-national-practices-estimating-net-stocksstructures.pdf. accessed: 06-10-2020.

Georegister, N. (2012). *Actueel Hoogtebestand Nederland 2*. URL: https://www.nationaalgeoregister.nl/geonetwork/srv/api/records/7939fd42-cf88-42c5-ab3b-232cf232a3ac.

Global Surface UHI explorer (2020). URL: https://yceo.users.earthengine.app/view/uhimap (visited on 06/10/2020).

Hanea, A. and Harrington, W. (2009). "Ordinal Data Mining for Fine Particles with Non Parametric Continuous Bayesian Belief Nets". In: *Information Processes Journal* 4, pp. 280–286.

Hanea, A., Morales-Nápoles, O. and Ababei, D. (2015). "Non-parametric Bayesian networks: Improving theory and reviewing applications". In: *Reliability Engineering & System Safety* 144, pp. 265–284.

Hanea, A., Kurowicka, D. and Cooke, R. (2007). "The population version of Spearman's rank correlation coefficient in the case of ordinal discrete random variables". In: *Proceedings of the Third Brazilian Conference on Statistical Modelling in Insurance and Finance*.

He, X., Miao, S., Shen, S., Li, J., Zhang, B., Zhang, Z. and Chen, X. (May 2014). "Influence of sky view factor on outdoor thermal environment and physiological equivalent temperature". In: *International journal of biometeorology* 59. DOI: 10.1007/s00484-014-0841-5.

Heusinkveld, B. G., Van Hove, L., Jacobs, C., Steeneveld, G., Elbers, J., Moors, E. and Holtslag, A. (2010). "Use of a mobile platform for assessing urban heat stress in Rotterdam". In: *Proceedings of the 7th Conference on Biometeorology*. Vol. 20, pp. 433–438.

Hintz, M., Luederitz, C., Lang, D. and Wehrden, H. von (October 2017). "Facing the heat: A systematic literature review exploring the transferability of solutions to cope with urban heat waves". In: *Urban Climate* 24. DOI: 10.1016/j.uclim.2017.08.011.

International Research Institute for Climate and Society (IRI) Climate and Society Map Room (2020). URL: http://iridl.ldeo.columbia.edu/maproom/ (visited on 06/10/2020).

IPCC Data Distribution Center (2020). URL: http://www.ipcc-data.org/ddc//ddc_provides.html (visited on 06/10/2020).

IPCC: Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V. and Midgley, P. M. (2013). "Summary for policymakers". In: *Climate Change 2013: the pysical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge: Cambridge University Press.

Jänicke, B., Holtmann, A., Kim, K., Kang, M., Fehrenbach, U. and Scherer, D. (November 2018). "Quantification and evaluation of intra-urban heat-stress variability in Seoul, Korea". In: *International Journal of Biometeorology* 63. DOI: 10.1007/s00484-018-1631-2.

Jesionek, P. and Cooke, R. (2007). "Generalized method for modeling dose-response relations application to BENERIS project". In: *European Union project.*

Kadaster (2018). *Topografie Nederland (TOP10NL): 2018-07-16.* URL: https://www.kadaster.nl/zakelijk/producten/geo-informatie/topnl.

— (2019a). *Basisregistratie Adressen en Gebouwen (BAG): 2019-05-09.* URL: https://www.kadaster.nl/zakelijk/registraties/basisregistraties/bag.

— (2019b). *Basisregistratie Grootschalige Topografie (BGT): 2019-05-09.* URL: https://www.kadaster.nl/zakelijk/registraties/basisregistraties/bgt/bgt-documentatie.

Kahn Jr, C. E., Roberts, L. M., Shaffer, K. A. and Haddawy, P. (1997). "Construction of a Bayesian network for mammographic diagnosis of breast cancer". In: *Computers in biology and medicine* 27.1, pp. 19–29.

KNMI (2020). *Sky View Factor of The Netherlands.* URL: https://data.overheid.nl/dataset/5657-sky-view-factor-van-nederland. accessed: 20-12-2020 (through Nelen & Schuurmans data portal.

Koningsveld, M. (van and Slinger, J. ((2015). *Building with nature video #03 - The Building with Nature Philosophy @ TU Delft 2015.* TU Delft. URL: http://dx.doi.org/10.4121/uuid:dbb2fe8f-bb3b-4fc4-b6ac-8c088b1d1047.

Koopmans, S., Droste, A., Heusinkveld, B. and Steeneveld, G.-J. (2020). *Validatierapport Landelijke Hittekaart.* Wageningen University.

Koopmans, S., Ronda, R., Steeneveld, G.-J., Holtslag, A. A. and Klein Tank, A. M. (2018). "Quantifying the effect of different urban planning strategies on heat stress for current and future climates in the agglomeration of The Hague (The Netherlands)". In: *Atmosphere* 9.9, p. 353.

Kurowicka, D. and Cooke, R. (2005). "Distribution-free continuous Bayesian belief". In: *Modern statistical and mathematical methods in reliability* 10, p. 309.

Landsberg, H. E. (1981). *The urban climate.* Vol. 28. International Geophysics Series. New York, United States of America: Academic press.

Langseth, H., Nielsen, T. D., Rumı, R. and Salmerón, A. (2009). "Inference in hybrid Bayesian networks". In: *Reliability Engineering & System Safety* 94.10, pp. 1499–1509.

Li, H., Zhou, Y., Jia, G., Zhao, K. and Dong, J. (2021). "Quantifying the response of surface urban heat island to urbanization using the annual temperature cycle model". In: *Geoscience Frontiers*, p. 101141. ISSN: 1674-9871. DOI: https://doi.org/10.1016/j.gsf.2021.101141. URL: https://www.sciencedirect.com/science/article/pii/S1674987121000050.

Li, H., Zhou, Y., Li, X., Meng, L., Wang, X., Wu, S. and Sodoudi, S. (2018). "A new method to quantify surface urban heat island intensity". In: *Science of The Total Environment* 624, pp. 262–272. ISSN: 0048-9697. DOI: https://doi.org/10.1016/j.scitotenv.2017.11.360. URL: https://www.sciencedirect.com/science/article/pii/S0048969717334186.

Li, H., Zhou, Y., Wang, X., Zhou, X., Zhang, H. and Sodoudi, S. (2019). "Quantifying urban heat island intensity and its physical mechanism using WRF/UCM". In: *Science of The Total Environment* 650, pp. 3110–3119. ISSN: 0048-9697. DOI: https://doi.org/10.1016/j.scitotenv.2018.10.025. URL: https://www.sciencedirect.com/science/article/pii/S0048969718338993.

Li, X., Zhou, Y., Asrar, G. R., Imhoff, M. and Li, X. (2017). "The surface urban heat island response to urban expansion: A panel analysis for the conterminous United States". In: *Science of The Total Environment* 605-606, pp. 426–435. ISSN: 0048-9697. DOI: https://doi.org/10.1016/j.scitotenv.2017.06.229. URL: https://www.sciencedirect.com/science/article/pii/S0048969717316492.

Matzarakis, A., Rutz, F. and Mayer, H. (2010). "Modelling radiation fluxes in simple and complex environments: basics of the RayMan model". In: *International journal of biometeorology* 54.2, pp. 131–139.

Mayrhuber, E. A.-S., Dückers, M. L., Wallner, P., Arnberger, A., Allex, B., Wiesböck, L., Wanka, A., Kolland, F., Eder, R., Hutter, H.-P. and Kutalek, R. (2018). "Vulnerability to heatwaves and implications for public health interventions – A scoping review". In: *Environmental Research* 166, pp. 42–54. ISSN: 0013-9351. DOI: https://doi.org/10.1016/j.envres.2018.05.021. URL: https://www.sciencedirect.com/science/article/pii/S0013935118302706.

Morales, O., Kurowicka, D. and Roelen, A. (2008). "Eliciting conditional and unconditional rank correlations from conditional probabilities". In: *Reliability Engineering & System Safety* 93.5, pp. 699–710.

Morales-Nápoles, O., Hanea, A. and Worm, D. (2014). *Experimental results about the assessments of conditional rank correlations by experts: Example with air pollution estimates*. London: Tayolor & Francis, pp. 1–24.

Morales-Nápoles, O., Delgado-Hernández, D. J., De-León-Escobedo, D. and Arteaga-Arcos, J. C. (2014). "A continuous Bayesian network for earth dams' risk assessment: methodology and quantification". In: *Structure and Infrastructure Engineering* 10.5, pp. 589–603. DOI: 10.1080/15732479.2012.757789. URL: https://doi.org/10.1080/15732479.2012.757789.

Moran, P. (1948). "Rank correlation and permutation distributions." In: *Biometrika* 44, pp. 142–144.

NBS Selection Tool (2020). URL: https://www.urbangreenup.eu/resources/nbs-selection-tool/nbs-selection-tool.kl (visited on 06/10/2020).

Nelen & Schuurmans (2020). *Hittestresskaart*. URL: https://hittestress.nu/.

Oke, T. R. (1982). "The energetic basis of the urban heat island". In: *Quarterly Journal of the Royal Meteorological Society* 108.455, pp. 1–24.

Oniésko, A., Lucas, P. and Druzdzel, M. J. (2001). "Comparison of rule-based and Bayesian network approaches in medical diagnostic systems". In: *Conference on Artificial Intelligence in Medicine in Europe*. Springer, pp. 283–292.

Paprotny, D., Kreibich, H., Morales-Nápoles, O., Terefenko, P. and Schröter, K. (2020). "Estimating exposure of residential assets to natural hazards in Europe using open data". In: *Natural Hazards and Earth System Sciences* 20.1, pp. 323–343. DOI: 10.5194/nhess-20-323-2020. URL: https://nhess.copernicus.org/articles/20/323/2020/.

Paprotny, D., Morales-Nápoles, O. and Jonkman, S. N. (2018). "HANZE: a pan-European database of exposure to natural hazards and damaging historical floods since 1870". In: *Earth System Science Data* 10.1, pp. 565–581. DOI: 10.5194/essd-10-565-2018. URL: https://essd.copernicus.org/articles/10/565/2018/.

Paprotny, D., Kreibich, H., Morales-Nápoles, O., Castellarin, A., Francesca and Schröter, K. (2020). "Exposure and vulnerability estimation for modelling flood losses to commercial assets in Europe". In: *Science of the Total Environment* 737. DOI: 10.31223/osf.io/r6dfg.

Paprotny, D. and Morales-Nápoles, O. (2017). "Estimating extreme river discharges in Europe through a Bayesian network." In: *Hydrology & Earth System Sciences* 21.6. URL: https://doi.org/10.5194/hess-21-2615-2017.

Paprotny, D., Morales-Nápoles, O., Worm, D. T. and Ragno, E. (2020). "BANSHEE–A MATLAB toolbox for Non-Parametric Bayesian Networks". In: *SoftwareX* 12, p. 100588. URL: https://doi.org/10.1016/j.softx.2020.100588..

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference—Judea Pearl*.

PEDDR Ecosystem Opportunity Mapping (2020). URL: https://pedrr.org/mapping-eco-drr-opportunities/ (visited on 06/10/2020).

Pogačar, T., Casanueva, A., Kozjek, K., Ciuha, U., Mekjavić, I. B., Bogataj, L. K. and Črepinšek, Z. (2018). "The effect of hot days on occupational heat stress in the manufacturing industry: Implications for workers' well-being and productivity". In: *International journal of biometeorology* 62.7, pp. 1251–1264.

Rijksoverheid (2019). *Basisregistratie Personen (BRP): 2019-05-15*. URL: https://www.kadaster.nl/-/basisregistratie-personen.

Rijkswaterstaat (2019). *Nationaal Wegen Bestand (NWB): 2019-05-15*. URL: https://www.pdok.nl/introductie/-/article/nationaal-wegen-bestand-nwb-.

Rose, A. (2004). "Economic principles, issues, and research priorities in hazard loss estimation". In: *Modeling spatial and economic impacts of disasters*. Springer, Berlin, pp. 13–36.

Ruangpan, L., Vojinovic, Z., Di Sabatino, S., Leo, L. S., Capobianco, V., Oen, A. M. P., McClain, M. E. and Lopez-Gunn, E. (2020). "Nature-based solutions for hydro-meteorological risk reduction: a state-of-the-art review of the research area". In: *Natural Hazards and Earth System Sciences* 20.1, pp. 243–270. DOI: 10.5194/nhess-20-243-2020. URL: https://nhess.copernicus.org/articles/20/243/2020/.

Ruangpan, L., Vojinovic, Z., Sabatino, S. D., Leo, L. S., Capobianco, V., Oen, A. M., McClain, M. E. and Lopez-Gunn, E. (2020). "Nature-based solutions for hydro-meteorological risk reduction: a state-of-the-art review of the research area". In: *Natural Hazards and Earth System Sciences* 20.1, pp. 243–270.

Schuurmans, N.

bibinitperiod (2019). *Land use Netherlands*. accessed: 20-12-2020 (through Nelen & Schuurmans data portal.

Shenoy, C. and Shenoy, P. P. (2000). "Bayesian network models of portfolio risk and return". In: *Y. S. Abu-Mostafa, B. LeBaron, A W. Lo, and A. S. Weigand (eds.), Computational Finance 1999*. The MIT Press, pp. 87–106.

Steeneveld, G.-J., Koopmans, S., Heusinkveld, B., Van Hove, L. and Holtslag, A. (2011). "Quantifying urban heat island effects and human comfort for cities of variable size and urban morphology in the Netherlands". In: *Journal of Geophysical Research: Atmospheres* 116.D20.

Stull, R. B. (1988). *An introduction to boundary layer meteorology*. ISBN: 978-94-009-3027-8.

ThinkHazard Tool (2020). URL: https://thinkhazard.org/ (visited on 06/10/2020).

TNO (2021). *Factsheet Climate Proof Cities: Hittestress*. URL: https://www.tno.nl/media/3959/factsheet-hittestress.pdf. accessed: 05-03-2021.

Unger, J., Sümeghy, Z., Gulyás, Á., Bottyán, Z. and Mucsi, L. (2001). "Land-use and meteorological aspects of the urban heat island". In: *Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling* 8.2, pp. 189–194.

Van Der Gaag, L. C. (1996). "Bayesian belief networks: odds and ends". In: *The Computer Journal* 39.2, pp. 97–113.

Waterschadeschatter (2020). URL: https://www.stowa.nl/publicaties/waterschadeschatter (visited on 27/08/2020).

Waterschapshuis (2018). *Aerial photograph of the Netherlands*. URL: https://www.hetwaterschapshuis.nl/beeldmateriaal. accessed: 20-12-2020 (through Nelen&Schuurmans data portal).

Wesenbeeck, B. K. (van and Slinger, J. ((2015). *Building with nature video #06 - Ecological processes in Building with Nature @ TU Delft: Part 1 2015*. TU Delft. URL: http://dx.doi.org/10.4121/uuid:d4b327cf-e3c9-49b6-b210-feea772186ca.

Worm, D., Morales-Nápoles, O., Haak, W. v. d. and Bakri, T. (2011). "Continuous Dynamic Non-Parametric Bayesian Networks: Application to traffic prediction". In: *TNO-report project number 043.01000*.

Appendices

A Current online tools

Online there can be found multiple tools that are related to NBSs and natural hazard mapping. The current tools available give insights in one of the following:

- Global overview of hazards experienced by urban centres;
- Insight into risk related to natural hazards;
- Available nature-based solutions and interventions.

We give some examples of both categories in the following paragraphs.

A.1 Hazard tools

The first category of tools that are online available is those who give a (global) overview of natural hazards. Tools can be found related to natural hazards like flooding (coastal, fluvial and pluvial), heat stress, landslides and wildfires. Some of these tools only map one hazard, while others have incorporated multiple. Especially flooding and heat stress are of interest for this research. Considering coastal and river flooding, two tools that deal with these on a global scale are aqueduct and the fathom. Both are providing current and future flood risk predictions globally, taking into account several climate change scenarios. The Aqueduct tool is based on a hazard-exposure-vulnerability-risk split. The resolution of this tool is 1km, making it quite coarse for recommendations on a urban scale. Although the Fathom flood hazard model has a 90m resolution and also includes pluvial flood risk, this model requires a license.

Considering heat stress, a tool that visualises this phenomenon due to the UHI effect is the *Global Surface UHI explorer* (2020) by Yale University. Related to heat stress, the *IPCC Data Distribution Center* (2020), the *CAIT Climate Data Explorer* (2020) and CIL all model climate data. The *Climate Change Knowledge Portal* (2020), the *IPCC Data Distribution Center* (2020), the *CREAT Climate Scenarios Projection Map* (2020) and the *International Research Institute for Climate and Society (IRI) Climate and Society Map Room* (2020) also offer tools to access datasets on climate.

Some tools give insight into multiple natural hazard categories simultaneously. The *ThinkHazard Tool* (2020) provides qualitative insights in flooding, heat stress, landslides, wildfires and other natural hazards for territories. Each hazard is classified from very low to high risk. The *Climate and Disaster Risk Screening Tools*

(2020) also cover all selected hazards. However, this tool relies on user input data and returns a qualitative assessment.

It is noted that several alternative methods for the urban heat island exist, including the Hittestresskaart (heat stress map) for the Netherlands (Nelen & Schuurmans), the urban heat index tool, and the global imperviousness dataset. However, the hittestresskaart and the urban heat index tool require extensive information on the city characteristics, while the imperviousness is only an indirect indicator for heat stress.

A.2 Risk tools

As was mentioned in the previous section, the Aqueduct tool and the Fathom flood hazard model both give insight into the risk of coastal and fluvial floods as well. Besides this, the *EcoActuary Tool* (2020) offers a link between natural hazards and a monetary valuation. It maps the economic benefits of mitigation of flood, heat stress and landslides by a large scale NBS. The *Waterschadeschatter* (2020) (water damage estimator) is provides detailed damage estimates specifically for flood risk. However, it was developed only for the Netherlands. Although it could be scaled globally, a lot of country specific socio-economical variables needs to be gathered in order to do so.

A.3 Intervention tools

The third category of tools focuses on NBSs. The *PEDDR Ecosystem Opportunity Mapping* (2020) maps the best locations for the restoration of coral reefs, mangroves, and forests. The resolution is coarse (approximately 10 km) and city characteristics are not taken into account. The costs and co-benefits are also not considered in the suitability analysis. However, the tool is valuable for the identification of coral reefs, mangroves and forests opportunities on a global scale.

There are two tools that give advice on using NBSs at a city level; the *NBS Selection Tool* (2020) by Deltares and the *Adaptation Support Tool* (2020). They both advise on the use of NBSs in cities, but in order to do so, the user needs to input data such as city characteristics himself. The former tool requires filling out a questionnaire that also covers hazards in the city, besides those city characteristics, making the tool strongly rely on the data provided by the user.

However, no tools are available that combine both insights using open data to come to a recommendation on a city level that proposes useful NBSs based on the local natural hazards. Also most of the current tools

do not include a cost-benefit analysis.

B Mathemathical background BNs

Bayesian Networks (BNs) were introduced in Section 2. A Bayesian Network is a graphical representation of a set of random variables and their conditional dependencies. The nodes in such a network represents the random variables and the directed edges represents the conditional dependencies, where the direction of the edge indicates the direction of causality. This property implies that a BN is a directed acyclic graph (DAG).

At the mathematical core of BNs we find Bayes' theorem, postulated by Thomas Bayes in 1763, after whom both the equation and the school of Bayesian statistics was named:

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)} \tag{21}$$

Starting with a parameter vector θ and data y, the term $p(\theta|y)$ is the probability distribution of θ given the data y. This term is also called the posterior distribution and literally is the updated belief on the parameter θ based on the data.

A Bayesian Network uses Bayes' theorem to model dependency between different random variables. By mapping the relation between random variables, one can update his/her belief on the probability of a random variable given evidence on related random variables (this can be seen as the data y from the previous passage). Below an example is given to visualise this process.

Figure 24: Bayesian Network of three events: $Snow$ (S), $TrainDelay$ (T) and a late arrival $Late$ (L) of a student at a lecture

Let us, for illustrative purposes, consider the following case. A student travels to class by train. Whether the student arrives late at class, depends (partly) on the probability of a delayed train. This event, in turn, is influenced by another event, namely whether it snows or not. In Figure 24 a DAG is shown for these three events. The structure of this DAG implicates $Snow$ only affects being $Late$ via $TrainDelay$ and assumes no other modes of transport, which could also be affected by $Snow$, are chosen by the student. Also, let us for illustrative purposes consider all these three nodes as being binary random variables.

We define the event $Late$ as L, $TrainDelay$ as T and $Snow$ as S. If $\hat{L}$ is an estimate of the event of being late, we are interested in $p(\hat{L} = 1)$. We can derive all probabilities of events if we know the full joint distribution

$P(S,T,L)$. As mentioned in the last sentences of the previous section, this involves 2^N parameters and therefore $2^N - 1$ entries (observations) are needed to estimate all parameters. In this example this boils down to seven observations. Regarding this aspect, the advantage of a BN becomes clear when writing out this probability using Equation 21 twice, the other way around:

$$\begin{aligned} P(S,T,L) &= P(S) \cdot P(T,L|S) \\ &= P(S) \cdot P(T|S) \cdot P(L|T,S) \end{aligned} \tag{22}$$

By construction of the DAG in Figure 24, in which we model L as being dependent on S only via event T, we state that if we know T then L is independent of event S, or phrased differently, L and S are conditionally independent given T: $(L \perp\!\!\!\perp S|T)$. This changes the final term of Equation 22:

$$P(S,T,L) = P(S) \cdot P(T|S) \cdot P(L|T) \tag{23}$$

Now we can use conditional probability tables for the three remaining to quantify this BN. Using the complement rule, the binary random variable $P(S)$ only has one unknown parameter, $P(T|S)$ has two and $P(L|T)$ has two, leading to a total of five unknown parameters to be estimated. In such a small example the number of parameters to be estimated, has decreased with almost 30% from seven to five. For problems with more nodes, this can lead to large reductions in parameters to be estimated.

As touched upon above, this decrease in number of parameters comes with the cost of a more restrictive model. By excluding a direct edge between snowfall and being late, we assume conditional independence between this nodes. However, in the context of a student which only uses the train as a mode of transport, this is a reasonable restriction. Another advantage of using this model, that is based on Bayes' theorem, is that it can be used very intuitively the other way around. We can, for example, ask our self the following question. What is the chance it has been snowing ($S = 1$), given that the train is delayed ($T = 1$)? We can again solve this problem with Bayes' theorem (Equation 21) combined with splitting up the joint density as done in Equation 22:

$$\begin{aligned} P(S=1|T=1) &= \frac{P(S=1,T=1)}{P(T=1)} \\ &= \frac{\sum_L P(S=1,T=1,L)}{\sum_S \sum_L P(T=1)} \\ &= \frac{P(S=1) \cdot P(T=1|S=1) \cdot \sum_L P(L|T=1,S=1)}{\sum_S \{P(S) \cdot P(T=1|S) \cdot \sum_L P(L|T=1,S)\}} \end{aligned} \tag{24}$$

In this example another advantage of using a BN arises, namely in the computational speed. We can again use $(L \perp\!\!\!\perp S|T)$, which simplifies the final result of Equation 24:

$$
\begin{aligned}
P(S=1|T=1) &= \frac{P(S=1)\cdot P(T=1|S=1)\cdot \sum_L P(L|T=1)}{\sum_S \{P(S)\cdot P(T=1|S)\}\cdot \sum_L P(L|T=1)} \\
&= \frac{P(S=1)\cdot P(T=1|S=1)}{\sum_S \{P(S)\cdot P(T=1|S)\}}
\end{aligned}
\tag{25}
$$

Because the final term in the denominator, $\sum_L P(L|T=1)$ is now independent from S, it can be written out of the summation over S and is therefore much easier to calculate. Especially since the sum over one variable given its predecessor(s) is always equal to one by construction, this term is now determined very quickly. We see that in this specific example the complete term even cancels out with the same term in the nominator.

C D-separation

It is important to realise that the influence of a given node works in two direction across the BN. This is dependent on the way the edges are connecting the nodes. In the example above, the path between *Snow* and *Late* is called 'inactive' because we condition on $TrainDelay$. As this is the only path between S and L and it is inactive, S and L are conditionally independent. Within the framework of BNs, one can also state that S and L are 'd-separated'. For larger networks this same analysis can be made per triple of nodes.

Theorem 1 *Let both X and Y be node that represent a discrete random variable in a BN. If all paths from X to Y are inactive, X to Y are conditionally independent. We can also say X to Y are d-separated.*

For larger BNs, several configurations active and inactive triples can occur. See Figure 25 for an overview. One should note that as soon as we find an active path between node X and Y, independence is not guaranteed. Only conditional independence can be guaranteed.

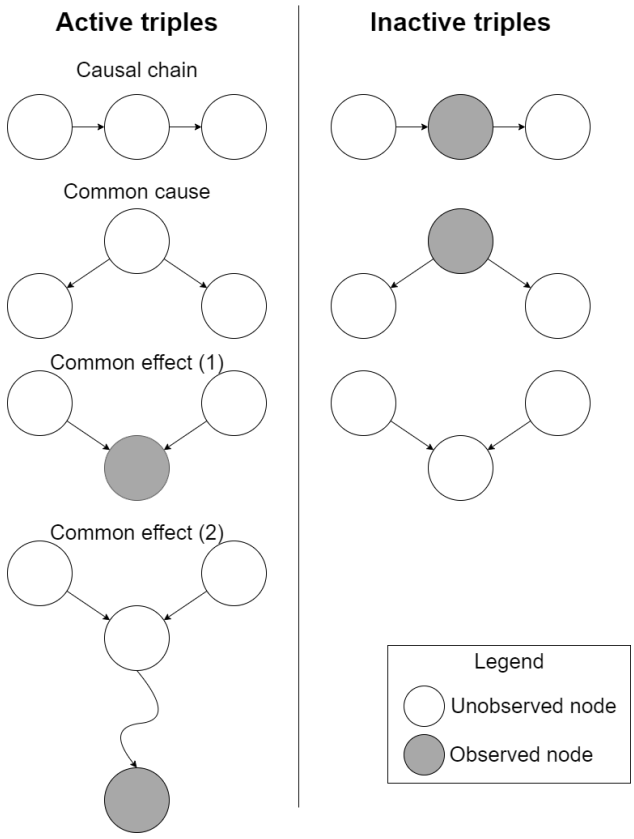


Figure 25: An overview of the possible active and inactive triples (part of a BN consisting of three nodes and their edges between each other) that can be encountered in a BN. A path is active if each triple on this path is active. Note we have two ways of a 'common effect' 1) if the direct descendant is observed or 2) if one of its descendants is observed.

In Figure 24 we have already used one of these triple formations, the 'causal chain'. We have seen how *Snow* and *Late* become conditionally independent if we observe $TrainDelay$. An example of a 'common

cause' could be another weather related variable, *Rain*. *Rain* causes both $Wetgrass$ and it causes people to carry $Umbrellas$. If we do not know anything on the common cause *Rain*, a larger probability of people carrying $Umbrellas$ will affect the probability we observe $Wetgrass$. However, if the common cause node is observed, this triple becomes inactive. If we know whether it *Rain*s or not, a larger probability of people carrying $Umbrellas$ will not change our belief on the probability we will observe $Wetgrass$, as this belief is dictated now by our observation of *Rain*. The third form a triple can take, is the 'common effect'. Both *Rain* and $Sprinklers$ can cause $Wetgrass$. If we observe $Wetgrass$ to be true, this increases both the chance of *Rain* to have happened and the chance of $Sprinklers$ to have been switched on. However, if we do not observe $Wetgrass$, these two nodes are conditionally independent and the triple is inactive. This 'updating of our belief' based on observations is also a reason BNs are often called Bayesian belief networks (BBNs). During the qualification process of a BN, one can check if the chosen edges agree with the data using the d-calibration score (Morales-Nápoles, Hanea and Worm, 2014).

D Applications of NPBNs

Hanea, Morales-Nápoles and Ababei (2015) classify applications in the fields of "risk analysis, reliability of structures, properties of materials" (Hanea, Morales-Nápoles and Ababei, 2015). Considering risk analysis, a first application of NPBNs is in assessing the safety of earth dams in Mexico (Morales-Nápoles, Delgado-Hernández et al., 2014 and Delgado-Hernández et al., 2014). They use a NPBN to prioritise the allocation of resources for maintenance activities in earth dams. Various failure mechanisms such as overtopping, breaching and piping are combined in one model, making it possible to assess risks at the level of an earth dam as a system. The marginal distribution of some of the model variables are assessed by using data from seven case studies in Mexico. Expert judgement is used for model variables of which no or little data is available. The results of the model are comparable with results which were reported in the literature. Other applications in the field of risk analysis are 1) establishing the relationship between sulfur dioxide emissions and concentrations of $PM_{2.5}$ (Hanea and Harrington, 2009) and 2) understanding the causal factors underlying the risks related to air transport and their relation to different possible consequences (Ale et al., 2008). Furthermore, Hanea, Morales-Nápoles and Ababei (2015) mentions applications in disciplines like traffic prediction Worm et al., 2011) and food consumption (Jesionek and Cooke, 2007), underlining the broad spectrum of disciplines in which NPBNs can lead to useful results.

D.1 NPBN in hazard modelling

Paprotny and Morales-Nápoles (2017) use a NPBN to estimate extreme river discharges in Europe. They assemble discharge data of 1841 river gauge stations. The measured daily discharges in these time series are converted to annual maximum discharges (Q_{AMAX}). For long data series ("at least three full decades of uninterrupted data"), which are available for 1125 stations, return periods can be calculated to obtain a validation set. These annual maximum discharges, calculated by the NPBN, are used to calculate return periods using a marginal extreme value distribution. This distribution can consequently be compared to a extreme value distribution fitted on the observed annual maximum discharges for the 1125 stations with long data series. Also Paprotny and Morales-Nápoles (2017) quantified the NPBN using only part of the data, in order to consequently validate the model with data of stations which are not used to quantify the model. This validation result is even better than the full quantification result. As a performance measure they use both the Pearson's coefficient of determination (R^2) and the Nash-Sutcliffe efficiency (I_{NSE}), which is a measure of bias of the model. Both have a maximum value of 1, but where (R^2) can take values between 0 and 1, I_{NSE} can become negative, indicating that the mean of the observations is a better predictor than the value obtained from the model. Using a validation set leads to a value of $R^2 = 0.94$ and $I_{NSE} = 0.93$. When using these performance measures to compare the performance of NPBNs with other models from the literature, this NPBN outperforms all of them, while not having geographical restrictions in applicability.

However, it should be noted that although this holds theoretically, in practice we walk into the problem of restricted data availability outside of Europe. As was already stated by Paprotny and Morales-Nápoles (2017), the performance of the NPBN decreased when applying a general circulation model with coarser resolution. Most of the datasets used have a resolution of 100 m or better; when looking at datasets globally available, most of the datasets have a resolution of 100 m or worse.

D.2 NPBN in exposure and vulnerability modelling

The research mentioned before uses a NPBN to estimate extreme river discharges, which makes a contribution towards estimating flood hazards. On the other hand, efforts are made to use NPBNs in estimating potential flood losses in both rural and urban areas. Paprotny, Kreibich, Morales-Nápoles, Terefenko et al. (2020) estimate the exposure of residential assets to natural hazards in Europe. To get a good estimation of the total exposure, the usable floor space area needs to be derived. Therefore the height or the number of floors of residential buildings is needed. Especially in urban areas, which can have tall residential tower blocks, this can be important to take into account when calculating potential flood losses. Because those variables (number of floors and building height) are rarely recorded in OpenStreetMap, Paprotny, Kreibich, Morales-Nápoles, Terefenko et al. (2020) use a NPBN to estimate the number of levels of a building, on the scale of individual buildings.

Then, using Eurostat and the Perpetual inventory method (PIM) the value per m^2 was calculated at a country level. Multiplying this by the estimated total dwelling area affected, the flood exposure can be calculated. Combining this with the PIM, one could arrive at an estimate of the exposure globally. Using a simplification of the PIM, only an average service life of dwellings and the investments in dwellings over this average service life time are needed to come to an estimate of the exposure. This latter vale, the investments done in dwellings, is found either by extracting the Gross fixed capital formation (GFCF) for asset type "dwellings" from Eurostat, or by extrapolation from the Gross domestic product (GDP) (Eurostat and OECD, 2014). In a similar way the household contents per m^2 are calculated on a country-level.

In a different study Paprotny, Kreibich, Morales-Nápoles, Castellarin et al. (2020) estimates the flood losses of commercial buildings. Although the methodology shows similarities with the study previously mentioned (Paprotny, Kreibich, Morales-Nápoles, Terefenko et al., 2020), now not only the flood exposure is considered, but also vulnerability. A NPBN is built to estimate the relative loss to the building structure and the relative loss to the machinery/equipment. However, besides variables related to flood hazard (inundation depth and inundation duration), also a variable related to vulnerability (the use of precautionary measures) is used as an explanatory variable in this model. Besides these, two exposure-related variables are also used as

explanatory variables. These two variables are used to summarise the economic situation of regions, defined using Nomenclature of Territorial Units for Statistics (NUTS). The Gross value added (GVA) per capita by NUTS3 region is used. This is the value of output minus the value of intermediate consumption; this variable is useful in the context of a basic economic principle in the assessment of natural disasters impacts: "to avoid the double counting of stocks and flows of goods and services because the value of an asset (stocks) is the discounted flow of its net returns" (Rose, 2004). Beside this variable, the GFCF per person employed by NUTS2 region is used.

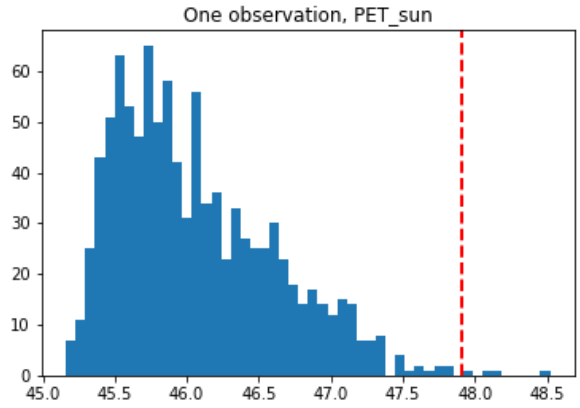
E Results

Table 20: All d-calibration scores

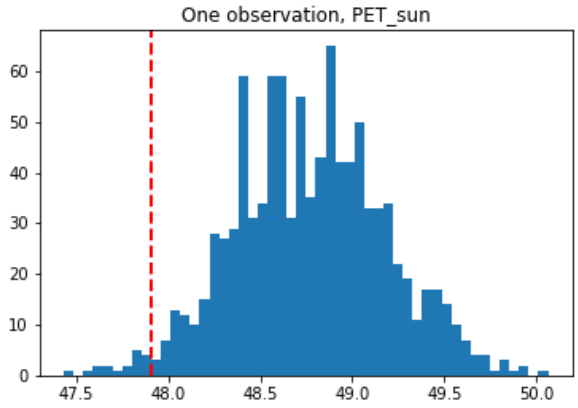
	BNRC			ERC		
Run ID	5th percentile	95th percentile	d-calibration score	5th percentile	95th percentile	d-calibration score
01	0.822	0.898	0.277	0.822	0.896	0.52
02	0.82	0.899	0.255	0.82	0.897	0.503
03	0.829	0.9	0.643	0.827	0.896	0.535
04	0.83	0.898	0.587	0.826	0.895	0.587
05	0.829	0.9	0.643	0.827	0.896	0.535
06	0.83	0.898	0.587	0.826	0.895	0.587
07	0.829	0.898	0.565	0.827	0.895	0.591
08	0.829	0.9	0.643	0.827	0.896	0.535
09	0.83	0.898	0.587	0.826	0.895	0.587
10	0.829	0.9	0.643	0.827	0.896	0.535
11	0.817	0.883	0.624	0.817	0.884	0.535
12	0.815	0.884	0.639	0.814	0.885	0.534
13	0.818	0.882	0.642	0.816	0.883	0.534
14	0.85	0.908	0.63	0.853	0.908	0.611
15	0.851	0.908	0.683	0.854	0.908	0.604
16	0.852	0.909	0.652	0.854	0.91	0.61
17	0.829	0.9	0.643	0.827	0.896	0.535
18	0.838	0.909	0.693	0.84	0.909	0.645
19	0.862	0.925	0.768	0.86	0.924	0.651
19a	0.873	0.938	0.674	0.871	0.938	0.597
19b	0.87	0.939	0.782	0.871	0.939	0.666
19c	0.864	0.927	0.747	0.862	0.925	0.658
20	0.873	0.939	0.851	0.874	0.939	0.664
21	0.899	0.956	0.752	0.897	0.955	0.781
22	0.899	0.956	0.752	0.897	0.955	0.781
23	0.873	0.939	0.851	0.874	0.939	0.664
24	0.829	0.9	0.643	0.827	0.896	0.535
25	0.898	0.956	0.77	0.897	0.955	0.825
26	0.898	0.955	0.729	0.897	0.955	0.775
27	0.897	0.955	0.728	0.897	0.955	0.732
28	0.852	0.925	0.75	0.848	0.922	0.657
29	0.873	0.939	0.851	0.874	0.939	0.664

Table 21: All MAE and CI values

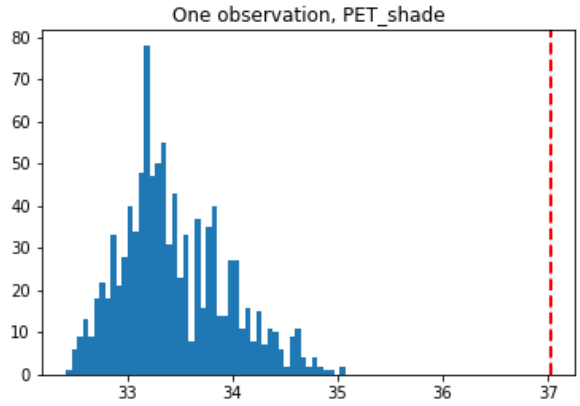
	PET_{sun}				PET_{shade}			
	In-sample		Out-sample		In-sample		Out-sample	
Run ID	CI	MAE [°C]	CI	MAE [°C]	CI	MAE [°C]	CI	MAE [°C]
01	39%	0.96	NA	NA	70%	0.32	NA	NA
02	47%	0.82	NA	NA	64%	0.30	NA	NA
03	32%	1.01	NA	NA	91%	0.15	NA	NA
04	58%	0.82	NA	NA	92%	0.13	NA	NA
05	32%	1.01	13%	1.15	91%	0.15	25%	0.62
06	58%	0.82	26%	1.04	92%	0.13	28%	0.56
07	67%	0.69	30%	0.96	92%	0.11	28%	0.54
08	32%	1.01	17%	1.16	91%	0.15	29%	0.56
09	58%	0.82	28%	1.04	92%	0.13	28%	0.55
10	61%	1.04	48%	1.17	97%	0.20	62%	0.60
11	60%	1.04	48%	1.17	97%	0.20	62%	0.61
12	54%	1.03	39%	1.17	96%	0.17	57%	0.56
13	61%	1.04	46%	1.17	97%	0.19	61%	0.60
14	55%	1.18	39%	1.31	95%	0.19	61%	0.58
15	51%	1.16	35%	1.30	94%	0.16	55%	0.54
16	54%	1.18	37%	1.31	95%	0.19	61%	0.57
17	95%	0.50	45%	0.96	91%	0.15	25%	0.62
18	84%	0.59	36%	1.10	91%	0.15	25%	0.62
19	81%	0.60	35%	1.09	91%	0.12	29%	0.58
19a	82%	0.25	12%	0.93	95%	0.09	26%	0.65
19b	90%	0.42	44%	0.76	91%	0.07	29%	0.55
19c	93%	0.49	12%	1.00	91%	0.12	29%	0.58
20	94%	0.48	47%	0.96	91%	0.12	29%	0.58
21	96%	0.56	68%	1.04	95%	0.19	61%	0.58
22	98%	0.80	67%	1.55	96%	0.67	79%	1.20
23	96%	0.69	62%	1.56	98%	0.46	65%	1.06
24	40%	1.00	17%	1.13	91%	0.16	25%	0.65
25	98%	0.81	69%	1.55	96%	0.66	79%	1.19
26	98%	0.82	73%	1.55	95%	0.68	77%	1.21
27	97%	0.83	70%	1.56	96%	0.67	76%	1.22
28	NA	NA	NA	NA	NA	NA	NA	NA
29	NA	NA	NA	NA	NA	NA	NA	NA



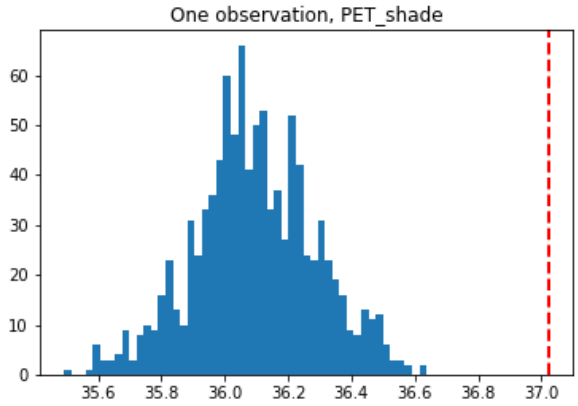
(a) Inference one observation PET_{sun} run 01



(b) Inference one observation PET_{sun} run 03



(c) Inference one observation PET_{shade} run 01



(d) Inference one observation PET_{shade} run 03

Figure 26: Difference in inference results for the same observation, using either Wageningen (rural and city) as training sample (run 01, left) or Wageningen city (run 03, right). As temperatures in rural area are in general lower due to the absence of the heat island effect, estimations from run 01 are underestimating the temperature in a city observation. In both cases the distribution of estimations of run 03 is closer to the observed value than run 01.

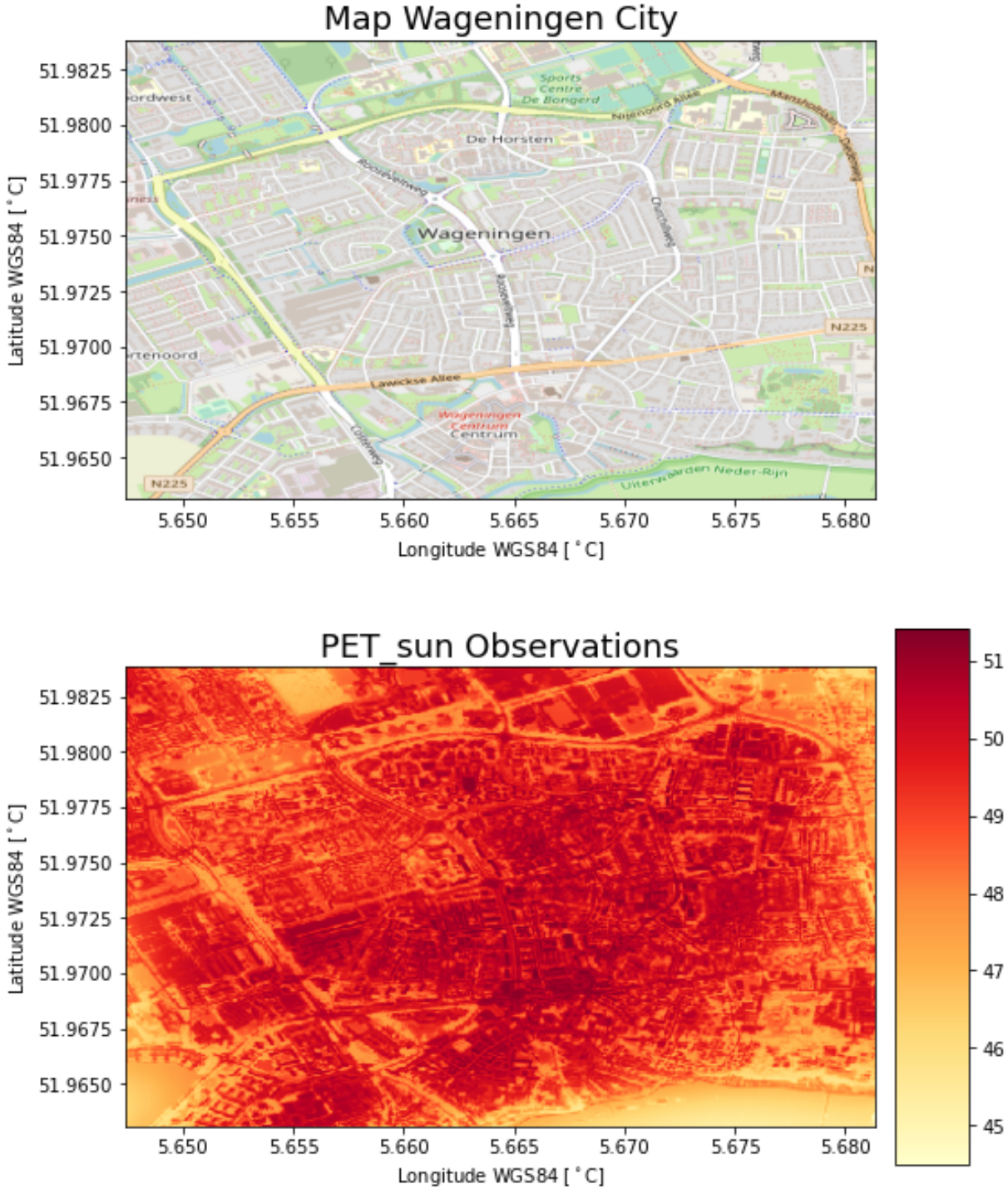


Figure 27: Map of Wageningen City with the observed PET_{sun} values (resolution = 10 m)

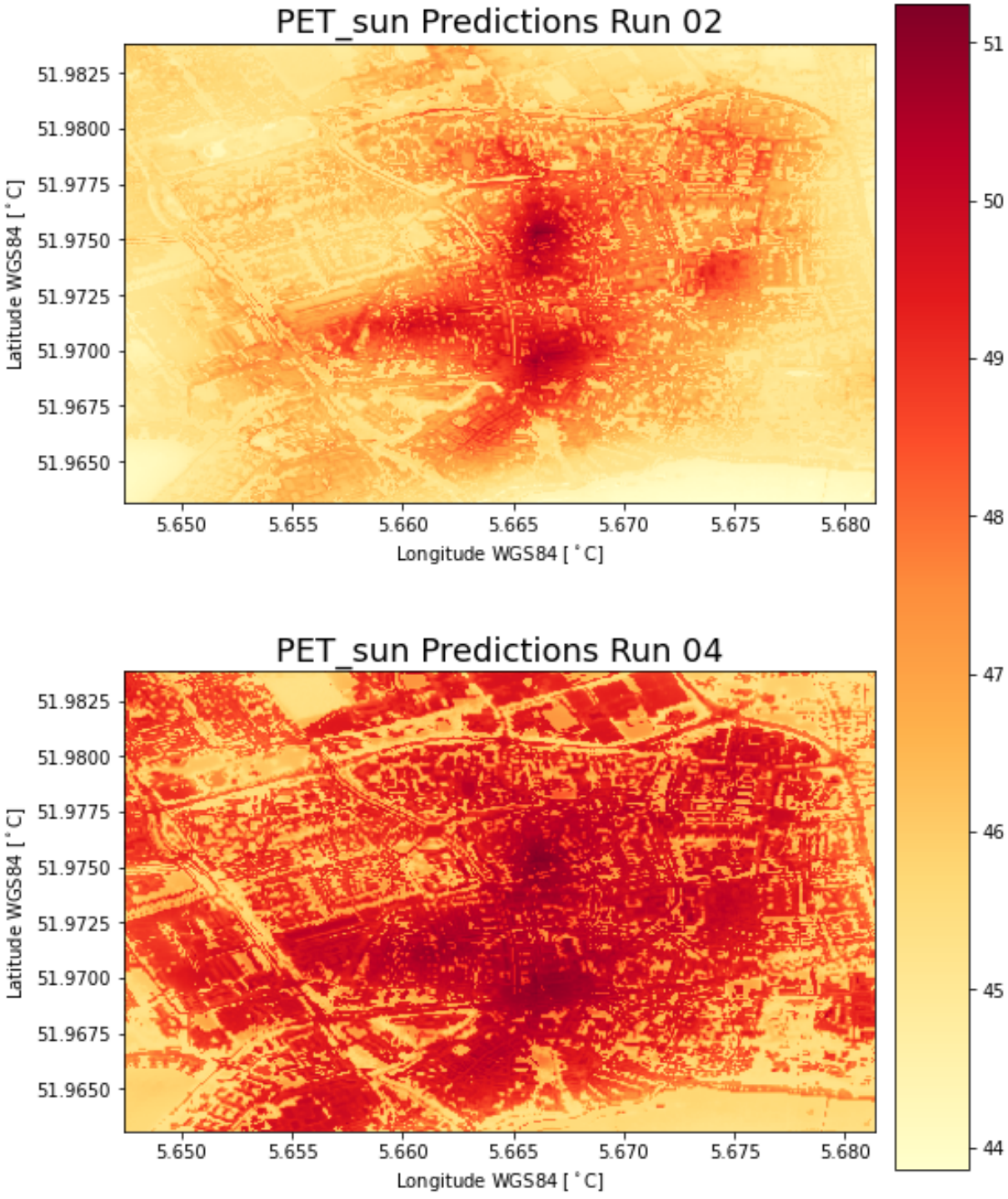


Figure 28: PET_{sun} predictions from run 02 (trained on both rural and urban data) and run 04 (only trained on urban data). Resolution = 10 m

E.1 Coarseness grid

When comparing the impact of grid resolution in the model, the key take-away is that for each step to a coarser grid, the model performs equally good or better, at the cost of detailed information on a geographical scale. This becomes clear when looking at Wageningen city for run 05, 06 and 07.

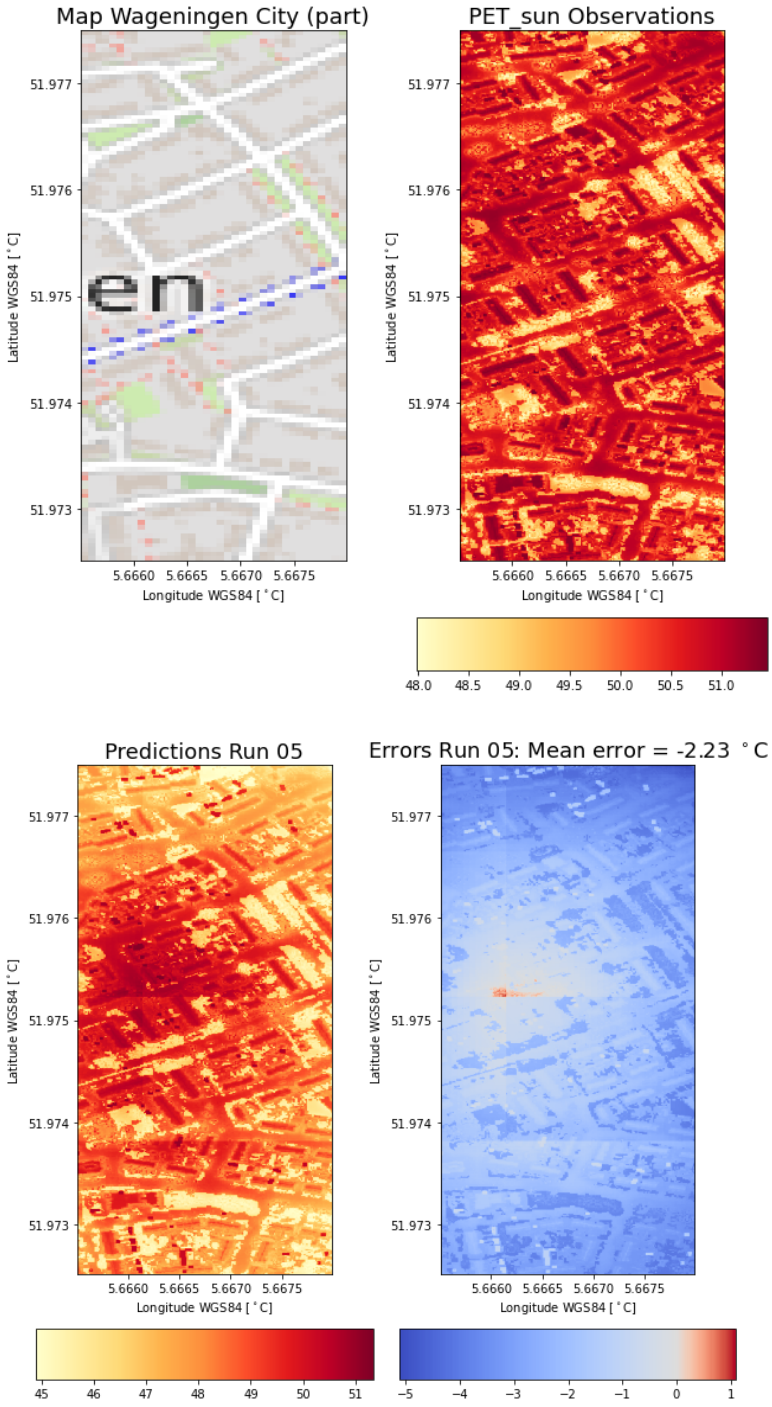


Figure 29: Comparison part of Wageningen city using a 1 m grid against the original observations.

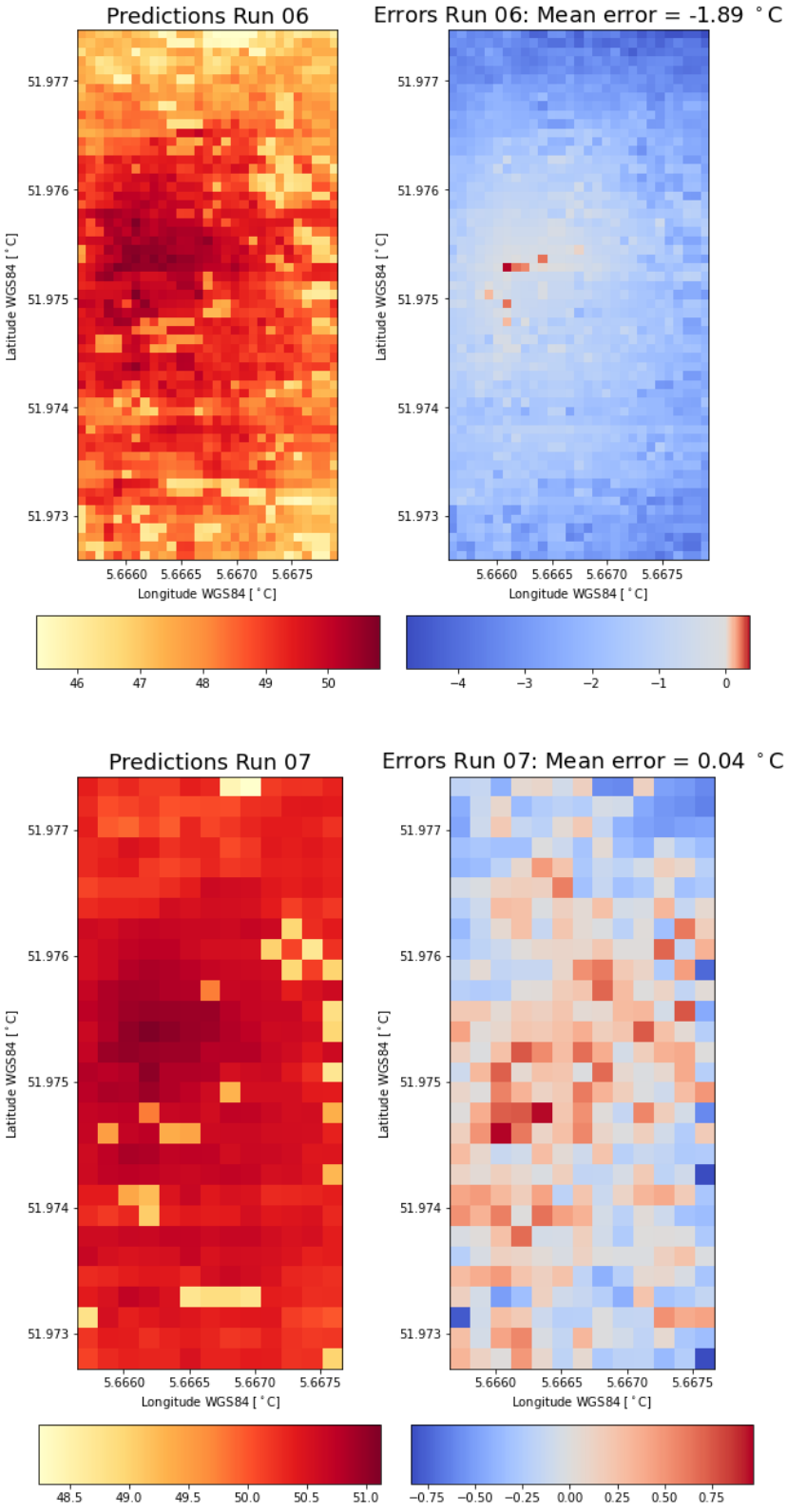


Figure 30: Comparison part of Wageningen city using a 10 m grid and a 20 m grid. Although the mean error decreases with increasing coarseness, exact locations with heat concentration are less clearly visible