Integration of expert knowledge and predictive learning: Modelling water flows in agriculture

Vladimir Kuzmanovski

Master Thesis
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia, September 2012

Evaluation Board:

Prof. Dr. Marko Bohanec, Chairman, Jožef Stefan Institute, Ljubljana, Slovenia

Dr. Florence Leprince, Member, ARVALIS Institute, France

Prof. Dr. Marko Debeljak, Member, Jožef Stefan Institute, Ljubljana, Slovenia

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Vladimir Kuzmanovski

Integration of expert knowledge and predictive learning: Modelling water flows in agriculture

Master Thesis

Integracija ekspertnega znanja z napovednim učenjem: Modeliranje vodnih tokov v poljedelstvu

Magistrsko delo

Supervisor: Prof. Dr. Marko Debeljak

Co-Supervisor: Prof. Dr. Sašo Džeroski

Ljubljana, Slovenia, September 2012

Index

Abstract

The pollution of water with phytopharmaceuticals used in agriculture can make significant damage to environment and human health. The water can be polluted by two types of pesticide inputs: diffuse and point source pollution. Both have to be considered separately and have to be covered in mitigation strategies for water protection.

The main paths of pesticide transfer from diffuse sources into the water are the water flows from a field, including runoff, drainage, lateral seepage and infiltration of water into the ground through the soil profile of the field. The water flows from a field play an important role in the process of water pollution. To reduce the water pollution, the amount of outflow of polluted water from the fields has to be reduced. The main question raised here is how to successfully reduce the outflow of polluted water from a field.

The answer consists of several steps. First, we need to have accurate models for predicting the outflow. Second, while taking into account predictions of these models, a proper set of mitigation strategies needs to be proposed. Many studies offer solutions mainly based on mechanistic approaches, but these solutions are either too expensive for parameterization or too general in their predictions, which results in inappropriately proposed solutions.

In this thesis, we address the problem of predicting the water flows from a field by methodology based on machine learning and data mining techniques. Our focus is mainly on the description of the process of water drainage, estimation of the critical periods of drainage events in an agricultural campaign and accurate prediction of the amount of drained water. The methodology being proposed is based on machine learning and data mining techniques.

The proposed methodology exploits both, the available expert knowledge structured in decision tables and the data collected from the experimental fields of ARVALIS, located in the La Jaillière region in France. The models built from expert knowledge and field data can improve the general recommendations for reduction of water flows from the fields.

The results of the thesis support the hypothesis that the available expert knowledge should be used in the process of supervised learning. Namely, a wide palette of data mining and machine learning methods, including regression trees, model trees, ensembles and polynomial equations has been implemented. The models were mainly learned in two different scenarios, namely using data from whole campaigns (12 months) and drainage season only, where a drainage season is defined as the period of a campaign with most intensive drainage events. Namely, the most accurate predictions have been achieved with the integrated predictive model for the drainage season (scenario two).

Povzetek

Onesnaževanje vode s fitofarmacevtskimi sredstvi, ki se uporabljajo v kmetijstvu, lahko zelo škoduje okolju in zdravju ljudi. Onesnaževanje vode s pesticidi poteka preko točkovnih ali prostorsko razpršenih virov zato mora obravnava problematike onesnaževanja potekati ločeno glede na vrsto vira, ki ga morajo upoštevati tudi strategije za varstva voda.

Glavne poti prenosa pesticidov iz polja so vodni tokovi, ki polje zapustijo z izhlapevanjem, površinskim odtokom, bočnim pronicanjem in infiltracijo vode skozi talni profil. Vodni tokovi s polj tako predstavljajo pomembno vlogo v procesu onesnaževanja voda. Da bi vplive onesnaževanja zmanjšali, je potrebno najprej zmanjšati količino odtoka onesnažene vode s polj, pri tem pa ostaja vprašanje, kako bi to lahko uspešno dosegli.

Odgovor je sestavljen iz nekaj korakov. Najprej je potrebno pripraviti točne napovedi količine iztočne vode, nato pa na osnovi napovedi predlagati ukrepe za zmanjševanje količine iztoka vode. Številne študije ponujajo rešitve predvsem na podlagi mehanističnih pristopov, vendar pa so te rešitve največkrat predrage zaradi stroškov določanja vrednosti parametrov modelov ali pa so v svojih napovedih preveč splošne, kar povzroči, da so predlagane rešitve neprimerne.

Naloga obravnava napovedovanje vodnih tokov s polja z uporabo metodologije, ki temelji na strojnem učenju in tehniki podatkovnega rudarjenja. Pri tem se osredotoča predvsem na opis procesa odvajanja vode, ugotavljanje obdobji intenzivnejših odtokov in na natančno napovedovanje količine odtečene vode.

Predlagana metodologija temelji na razpoložljivem strokovnem znanju, ki je podano v obliki odločitvenih tabel in na podatkih, zbranih na poskusnih poljih, ki se nahajajo v regiji La Jaillière, Francija. Modeli, ki temeljijo tako na ekspertnih, kot dejanskih podatkih, lahko izboljšajo splošna priporočila glede ukrepov za zmanjšanje odtoka količine vode s polj.

Rezultati naloge podpirajo hipotezo o učinkovitosti vključitve razpoložljivega strokovnega znanja v proces nadzorovanega učenja. Široka paleta metod podatkovnega rudarjenja in metod strojnega učenja, vključno z regresijskimi drevesi, modelnimi drevesi, ansambli in polinomskimi enačbami, je bila namreč že uveljavljena. Modeli so bili zgrajeni na podlagi dveh scenarijev: z uporabo podatkov iz enoletnih obdobji spremljanja količine drenirane talne vode (12 mesecev) in z uporabo podatkov iz obdobji intenzivnega dreniranja, ki so bila časovno opredeljena na osnovi strokovne ocene ARVALIS-a. Največjo točnost napovedi smo dosegli z integriranim napovednim modelom za obdobje intenzivnega drenirana (drugi scenarij).

Abbreviations

ARVALIS	=	ARVALIS Institut du Végétal (France)
KW	=	Kinematic wave
1D-DPM	=	One-dimensional dual-porosity model
2D	=	Two-dimensional
3D	=	Three-dimensional
DP-MIM	=	Dual-permeability/Mobile-immobile
EK	=	Expert knowledge
ES	=	Expert system
RMSE	=	Root mean squared error
RRSE	=	Root relative squared error
Std.Dev.	=	Standard deviation
CIPER	=	Constrained Induction of Polynomial Equations for Regression

1 Introduction

The pollution of water with phytochemicals used in agriculture can make significant damage to aquatic ecosystems and human health. The water can be polluted by two types of pesticide inputs, diffuse and point source pollution, which have to be considered separately (Reichenberger et al, 2007). Furthermore, the risk of water pollution can be reduced by appropriate mitigation strategies (Kreuger & Nilsson, 2001). Mitigation of pesticide inputs into water includes the reduction of both diffuse and point source pollution. The main paths of pesticide transfer from diffuse sources into the water are the water flows from a field, including runoff, drainage, lateral seepage and infiltration of water into the ground. The water flows in a field play an important role in the process of water pollution. To reduce water pollution, the amount of outflow of polluted water from the fields has to be reduced. The main question raised here is how to successfully reduce the outflow of polluted water from a field.

The answer consists of several steps. First, we need to have accurate models for making predictions of the outflow. Second, while taking into account these predictions, a proper set of mitigation strategies needs to be considered. The proper set of mitigation strategies needs to cover strategies valid for the analyzed type of water flow and its general characteristics. Finally, we need to make the final selection of the most appropriate mitigation strategy, which has to respect local field characteristics.

Many studies offer solutions mainly based on mechanistic approaches, but these solutions are either too expensive for parameterization or too general in their predictions, which results in inappropriately proposed solutions.

In this thesis, we address the problem of predicting the water flows from a field. Our focus is mainly on the description of the process of water drainage, estimation of the critical periods of drainage events in an agricultural campaign, and accurate prediction of the amount of drained water. The methodology being proposed is based on machine learning and data mining techniques.

Machine learning approaches often give effective and accurate solutions for this kind of problems (Debeljak & Džeroski, 2011). Namely, machine learning is a very promising research area with numerous applications in the field of environmental sciences. There exist three types of machine learning: supervised, unsupervised and semi-supervised learning. The problem of prediction is addressed by supervised machine learning (also known as predictive machine learning). Supervised machine learning tries to build a predictive model (in the form of decision trees, decision rules, linear equations, etc.) that will accurately predict the values of the dependent target variable (class or output). The process of learning the model is based on experience, given in the form of learning examples and described with a feature (attribute) set.

The proposed methodology exploits both the available expert knowledge structured in decision tables and the data collected from experimental fields of ARVALIS, located in the La Jaillière region in France. These are used by our machine learning methodology in order to obtain water flow models with high accuracy, resulting in efficient proposed mitigation measures. The built model improves the general recommendations made from expert knowledge by making them more specific and adjusting the knowledge, based on the data collected and the predictions of the models learned by data mining.

The thesis is organized as follows.

In Section 1, we briefly presented some background on the topic of the thesis and gave an overview of the related domain knowledge in the field of understanding the water flow through the soil in agriculture. In Section 2, we describe the related work on the topic of modeling the water flow in agriculture. In addition, a description of the topic of the thesis, its goals and the hypothesis are given.

In Section 3, the data from the La Jaillière experimental site are described, followed by a description of the procedure of data collection. The dataset contains data for soil properties, climate data and crop and land management. At the end, the expert knowledge, provided in the form of decision tables, is described.

Next, in Section 4, we present the techniques for data pre-processing that we used with some data mining and machine learning methods for feature ranking and selection. First, we describe the pre-processing of the dataset from the experimental field, including database manipulation and feature set creation. We then describe the propositionalization of the expert knowledge and its decision tables.

In Section 5, we present the machine learning techniques used in this thesis. These include classification, regression and model trees, and ensemble methods. The experiments are performed in two machine learning environments: WEKA and CLUS. In Section 6 the complete experimental setup is presented.

Then, in Section 7, we present the evaluation of the results from different experiments performed in our study and presented in the experimental setup. This section also discusses the significance of the results relative to existing and default models, and the influence of the independent variables. Towards the end of Section 7, we summarize the content of the thesis.

Finally, in Section 8, we state our conclusions; summarize the contributions and outline possible improvements in further work.

2 Problem description

Modeling water flows or preferential flows has been the interest of many academic areas in the past 30 years, with the underlying aim to prevent groundwater from contamination. A number of laboratory and field experiments have been conducted to qualitatively describe water flow through unsaturated and saturated soil, and understand the mechanisms that control this type of flow (Ersahin *et al*. 2002). Even though many mechanistic and physically-based models have been developed to quantitatively describe water flow through soils, but very few of them are complete in terms of incorporation of data like soil structure and soil matrices.

2.1 Background

The consumption of water in the world is approximately doubling every 20 years i.e., much faster than the population increases. On the other hand, new sources of water are becoming scarcer and polluted water is becoming more expensive to remediate. Taking these facts into consideration, it is urgent to protect the existing water.

Soil is the first filter of the earth's water. The soil's ecosystem processes of buffering and filtering are very important for achieving the quality of the surface and sub-surface water reserves. Land management can affect the ability of the soil to defend the receiving water and disturb the process of filtering. Statistics show that about 70 % of the world's fresh water is consumed by agriculture (Clothier *et al*. 2008). Therefore, it is necessary to protect the water reserves from the land and crop management practices in agriculture. In addition, the performance of land-management practices can be both tracked and improved, so that policies for sustainable management can be developed.

The traditional land management and agricultural production were based on traditional economics. It has viewed capital as simply being cash, investments and economic instruments. However, sustainable development is now seen to rely on four types of capital: the traditional capital finance; the manufactured capital of infrastructure; human capital in the form of intelligence, culture and organization; and the natural capital of the renewable and non-renewable stocks of natural resources that support life and economic activities (Hawken *et al*. 1998). Therefore, there is a need to develop an integrated and accepted system of valuing or measuring natural capital and ecosystem services. Furthermore, Fenech *et al*. (2003) point out that turning the idea of natural capital into a practical means of measuring or modeling both economic and ecological systems requires considerable study and innovation.

The above statements define the needs of developing a system that will improve the land and crop management practices in order to protect the quality of the ecosystem, including water and environmental resources. The quality of the water resources is highly dependent on the water pollution with phytochemicals from agriculture. There exist two main types or sources of water pollution: point and diffuse source pollution.

In this thesis, we focus on diffuse source water pollution. The main paths of pesticide transfer from diffuse sources into the water are the water flows from a field, including runoff and erosion, drainage, lateral seepage and infiltration of water into the ground. The water flows in a field play an important role in the process of water pollution. To reduce

the water pollution, the amount of outflow of polluted water from the fields has to be reduced.

To achieve the aim of protecting the water, we need to consider the outflow of the polluted water and select sustainable mitigation strategies. The key step in selecting a sustainable mitigation strategy is the understanding of the outflow of the polluted water, and its dependence on the field's properties. This could be obtained from predictions of high accuracy models.

The study of water flow in soil begins nearly a century ago. In 1898, Colonel Moore of the Royal Engineers stated that "in undrained clay land, cracks of one and two inches wide and five feet deep are sometimes met with, with the result that direct passage of sewage and surface water into them has occurred, so that the effluent is not purified as intended. It is thus very unsuitable for irrigation, unless the surface is specially prepared" (Moore, 1898).

Ever since, scientists and experts have studied the phenomena of water and solute movement along certain pathways, while bypassing or going through fractions of the porous matrix in the soil. First, they tried to understand the process and its characteristics. Second, they developed models for describing the process of movement using soil characteristics. In the next stage, some conceptual models were built, based on analytical and statistical observations. In the last 20 years, better mechanistic models were developed to incorporate water flow processes (Gerke & van Genuchten, 1993).

A significant time was spent debating about the approaches that need to be considered regarding the main causes of water flow in the soil, the kind of characteristics that should be considered while developing models, as well as the kind of methodology that needs to be used for developing models. At the beginning, the experts thought that there is a direct relationship between the quality of soil and those soil characteristics that would initiate and sustain water flow and transport through it. In this group of experts the most numerous were pedologists. They noted that the soil characteristics are enough to understand and eventually predict the water flow in the soil (Clothier & Green, 1997). On the other hand, some soil physicists considered that flow through soil is not uniform phenomenon that would enable pedological theories to predict it easily (Köhne *et al.* 2009). Furthermore, the main causes of water and solute flow in soil are not the forces of capillarity and gravity acting alone in a porous medium. Rather it is the constantly changing distributed pattern of the soil's pressure created by plant roots that hold the major control over the hydrology of surface soil (Clothier & Green, 1997).

In this thesis, we consider both the soil properties and crop management which are responsible for water flow through the soil. We, additionally, consider climate influence and land management.

Furthermore, the conceptualization of the soil helps the scientists to recognize the trend of the flow through the soil. The traditional conceptualization defines the soil as a porous medium with continuous properties (Figure 1). Therefore, the water flow is uniform and at a local equilibrium is usually described with Richards's equation (Richards, 1931) (1):

$$\frac{d\theta}{dt} = \frac{d}{dz}[K(\theta)(\frac{d\psi}{dz} + 1)] \tag{1}$$

where K is the hydraulic conductivity (cm/s), ψ is static pressure head (m), z is elevation above a vertical datum (m) and θ is the water content of the soil.

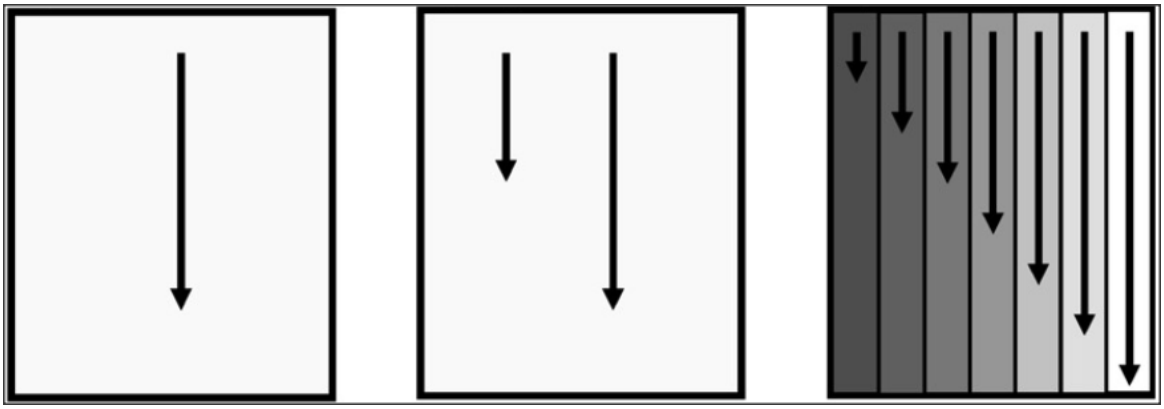


Figure 1: *Continuum conceptualization*. A visualization of the continuum conceptualization of the soil regarding the water flow (Köhne *et al*. 2009). Left-to-right: Uniform flow, Dynamic flow and Stream tubes.

Moreover, by defining dynamic flow and stream tubes flow, the experts provide additional descriptions of the possible water flow in the continuum conceptualization of the soil.

Other conceptualizations of the soil are the bi- and multi-continuum which define soil as a composition of two or more domains (Figure 2). Therefore, the water flow can be horizontal (lateral seepage) and vertical (infiltration). In case of a horizontal flow, domain to domain exchange of water is noticed, while in a vertical flow only in-depth flow.

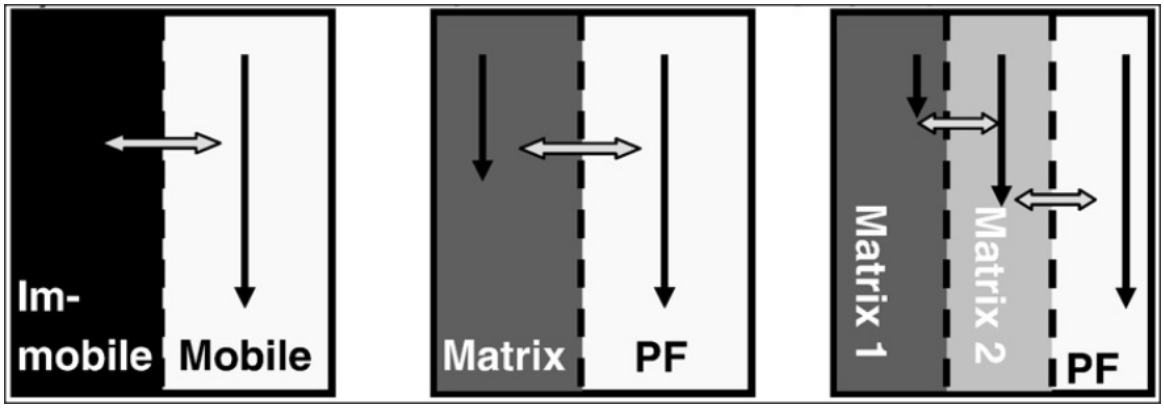


Figure 2: *Bi- and Multi-continuum conceptualization*. A visualization of the bi- and multi-continuum conceptualizations of the soil regarding the water flow (Köhne et al. 2009). Left-to-right: Mobile-Immobile approach, dual-permeability approach and triple-permeability approach

The bi- and multi-continuum conceptualizations assume that the porous medium consists of two overlapping pore domains, with water flowing relatively fast in one domain (often called the macro-pore, fracture, or inter-porosity domain) and slow in the other domain (often referred to as the micro-pore, matrix, or intra-porosity domain) (Köhne *et al*. 2009). Based on this knowledge and available schemes and conceptualizations, many models were developed for simulation and prediction of the water flow through the soil. Basically, all of them are modeling schemes with which we can, with caution, predict the features and impacts of water flow and transport in soils. However, the uncertainty in such prediction is unknown. Furthermore, when taking into account the deep structure of the soil, these models are very expensive for parameterization. Yet another major concern about the existing models is that their validation status is quite low (Dubus *et al*. 2002). This may be due to lack of data, poor parameter identification techniques, or the use of subjectivity in the parameterization process.

A strategy to overcome these problems is to build simplified models which are understandable and less complex. Furthermore, the model should not require expensive input data, which need to be provided by special analyses. In other words, the model must maintain the optimal relation between the complexity of the model, the cost of the input and the predictive accuracy of the model. We have achieved this goal by using machine learning and data mining techniques.

Machine learning and data mining techniques, such as regression trees, artificial neural networks and support vector machines have been widely used in many applications (Witten and Frank, 2005). These techniques exploit the accumulated vast quantities of data because they rely on effective learning procedures. With the rapid development of these techniques, we can achieve more reliable and accurate results. Machine learning is part of the broader area of artificial intelligence. In artificial intelligence, one of the basic approaches, knowledge engineering (Feigenbaum and McCorduck, 1983) is to extract domain knowledge from an expert and encode it in computer-readable form.

This study investigates the possibility of integrating expert knowledge within the process of learning predictive models from data. This approach gives more accurate and precise predictions about the water flow from a field, as compared individually with either the existing expert knowledge or the models learned from data only by using machine learning techniques.

2.2 Related work

Researchers have attempted to model processes from experimental observations into conceptual approaches and mathematical models. The governing equations can be implemented in computer simulation tools, which can then be tested, modified, and/or validated against new experimental data. Increasingly sophisticated models have been developed to analyze water flows in various environmental systems, where there exist considerable differences in spatial and temporal scales.

A comprehensive literature survey on the existing models for describing (prediction and simulation) water flows emphasizes two major modeling approaches: conceptual and computer models.

Classical physically-based conceptual models for water flows and solute transport in structured soils can be broadly classified into continuum, bi- continuum or multi-continuum models, as we described before (Köhne *et al*. 2009). Each of these conceptualizations depends on a number of the considered domains. Namely, continuum conceptualization consider flow in the entire soil as being controlled by both capillary and gravity forces, while bi-continuum approaches assume that flow in the flowing domain is controlled by gravity only and is always directed downwards. On the other hand, the newest approaches from multi-continuum conceptualization, such as capacity or routing controlled approaches comprise simplified descriptions of macro-pore flow. Furthermore, in each of these conceptualizations there exist different approaches: uniform flow and stream tubes (Figure 1), mobile-immobile, dual and triple permeability approach. These approaches are based on the type of the water flow through the soil: “domain to domain” water exchange and in-depth water infiltration.

In the continuum conceptualization, scientists model the water flow based on Richard’s equation (1), unlike the bi-continuum approaches which are based on the Kinematic Wave – KW (Lighthill and Whitham, 1955) equation (2) for the water flow in macro-pore regions and Richard’s equation for the water flow in the matrix regions of the soil. The KW equation is in the following form:

$$\frac{dh}{dt} + C\frac{dh}{dx} = D\frac{d^2h}{dx^2} \tag{2}$$

where h is the debris flow height, t is the time, x is the downstream channel position, C is the pressure gradient (depth dependent nonlinear variable *wave speed*) and D is a flow (height dependent variable *diffusion term*).

These models, described above, are only conceptual and their accuracy is not satisfactory for the practical use. Additionally, Brederhoeft (2005) noted that while the foundations of modeling are the conceptual models, new data typically render invalid predictions from these conceptual models. Moreover, he suggests that the solution of this problem is twofold: to collect as much data as feasible, and for the model developer to keep open the possibility to change the conceptual models.

The principle of conceptual model refinement was used during the development of computer models, which are still mechanistic. They are usually based on the dual-permeability approach of Gerke and van Genuchten (1993). Hereafter, we give a brief description of the commonly used models.

DUAL (Gerke and van Genuchten, 1993) is a 1-Dimensional Dual-Permeability Model (1D-DPM) based on two Richard's equations (1) and convection dispersion equations for matrix and fracture pore systems. These are coupled by first-order terms for bi-directional exchange of water and solute. The original research model was later adapted for application to field conditions (Gerke and Köhne, 2004).

HYDRUS-1D (Šimůnek et al., 1998, 2003, 2005, 2008a, b, c; Šimůnek and van Genuchten, 2008) describes water, heat, and solute movement in the vadose zone. It simulates water flow, solute and heat transport in one-dimension and is public domain software. HYDRUS (2D/3D) extends the simulation capabilities to the second and third dimensions, and is distributed commercially.

The KW one-region model (Beven and Germann, 1982; Germann, 1985) is based on the boundary layer flow theory and was used for describing water flows. The KW model assumes that the wetting front proceeds by convective film flow in the mobile region and does not exchange water with the immobile region.

MACRO (Jarvis, 1994; Larsbo and Jarvis, 2003; Larsbo *et al*., 2005) is a 1D-DPM that combines a KW equation (2) for describing the water flow and solute convection for the macro-pore region with Richards's equation (1) for water flow and solute convection dispersion in the matrix. Water transfer into the matrix is treated as a first-order approximation to the water diffusion equation and is proportional to the difference between actual and saturated matrix water contents.

The Root Zone Water Quality Model, RZWQM (Ahuja *et al*., 2000) utilizes a dual-permeability/mobile-immobile (DP/MIM) description of 1D vertical soil water flow and chemical movement. Three transport regions are assumed to exist in the soil: cylindrical macro-pores, the mobile soil matrix, and the immobile soil matrix. In the macro-pores, water flow is calculated using the Poiseulle equation (3) and solutes are displaced by convection. The Poiseulle equation has the following form:

$$\Delta P = \frac{8\mu L Q}{\pi r^4} \tag{3}$$

where ΔP is the pressure drop, L is the length of pipe, μ is dynamic viscosity, Q is the volumetric flow rate, r is radius and d is the diameter.

In the mobile matrix region, water flow is described using the Green-Ampt equation (4) approach during infiltration and the Richard's equation (1) during redistribution, while

solute moves by convection (Köhne *et al*. 2009). The Green-Ampt approach is described by the following equation:

$$F = \frac{K_a S_w (\theta_a - \theta_i)}{i - K_a} \tag{4}$$

where θ_a and θ_i are the saturated and initial volumetric water contents, respectively, S_w is the soil water suction at the wetting front, i is rainfall intensity and K_a is the saturated hydraulic conductivity.

There are other models and approaches for qualitatively describing water flow through unsaturated and saturated soil, and understanding the mechanisms that control this type of flow. Among these, worth emphasizing is the FOOTWAYS (2009) web-based system which is basically developed to help farmers to protect the water and the environment from phytochemical products that come from agriculture (diffuse sources). As a part of the system, there exists one module for prediction and simulation of the water flow through and over the surface of the soil. Its structure is a composition of the MACRO and Pesticide Root Zone (PRZM) models, dealing with water flow through the soil and surface water flow, respectively.

2.3 Problem

In the previous section, we gave a brief description of existing conceptual and computer models for prediction and simulation of the water flow through the soil. All of the mentioned models are either physically-based or mechanistic. Basically, all of them are useable modeling schemes with which we can, with caution, predict the features and impacts of water flow and transport in soils. However, the uncertainty of these predictions is unknown. In addition, they are either too expensive for parameterization or too general in their predictions, which results in inappropriately proposed solutions. Therefore, the biggest concern in such a case is how much data we need to fit the input parameters in order to get the output, prediction or simulation. Yet another major concern about the existing models is that their validation status is quite low (Dubus *et al*. 2002). This may be due to lack of data, poor parameter identification techniques, or the use of subjectivity in the parameterization process.

A strategy to overcome these problems is to build models that will establish an optimal tradeoff between the complexity of the induced models, expensiveness of the input data, and the model's predictive accuracy. The input of the model should comprise data that can be taken from the field without any specific analyses of the soil and soil matrices, as well as hydraulic conductivity and moisture contents. On the other hand, the model's output should be accurate. Our understanding of the mechanism of water flow that we gained from the experimental approaches and the observations from the La Jaillière experimental field, led us to a better concept of modeling and better parameterization of our models. We achieved this goal by using machine learning and data mining techniques on data available from the La Jaillière experimental site and available expert knowledge.

2.4 Hypothesis

The main hypothesis of our study is that the integration of structured expert knowledge with models built from data gives more accurate and precise predictions about the water flows through the soil and from the fields, as compared individually with the existing expert knowledge or the models built from data using machine learning techniques. The

outputs from the expert system (e.g., expert decision tables) are too general and cannot be adjusted to local specifics, while the models built from data can give wrong predictions if they are not correctly validated (supported by expert knowledge), or if they are overfitted to local specifics. Therefore, the solution proposed in the thesis generates outputs that are adjusted to local specifics (like soil properties, surface properties and agriculture practice) and supported by expert knowledge.

3 Data description

Today's technology allows us to collect data from different sources and for different purposes. This fact encourages scientists to use a wide spectrum of data and progressively combine them for learning models by data mining in order to achieve more accurate results. Namely, our study is concerned with learning models based on real data and expert knowledge. Therefore, we used two types of input: the data collected from experimental fields and the available expert knowledge captured in decision tables.

First, we describe the location and properties of the La Jaillière experimental site, where data were collected in the past 25 years. Second, a brief description of the procedure of collecting data is given. Then, we present the data of the soil properties, climate and on-field practices. Finally, the expert knowledge is described.

3.1 Experimental site

The experimental site, where data were collected, is located in western France (Figure 3). It is situated at the southern end of the Armorican massif, in the La Jaillière province. It is owned by ARVALIS - Institut du Végétal*.

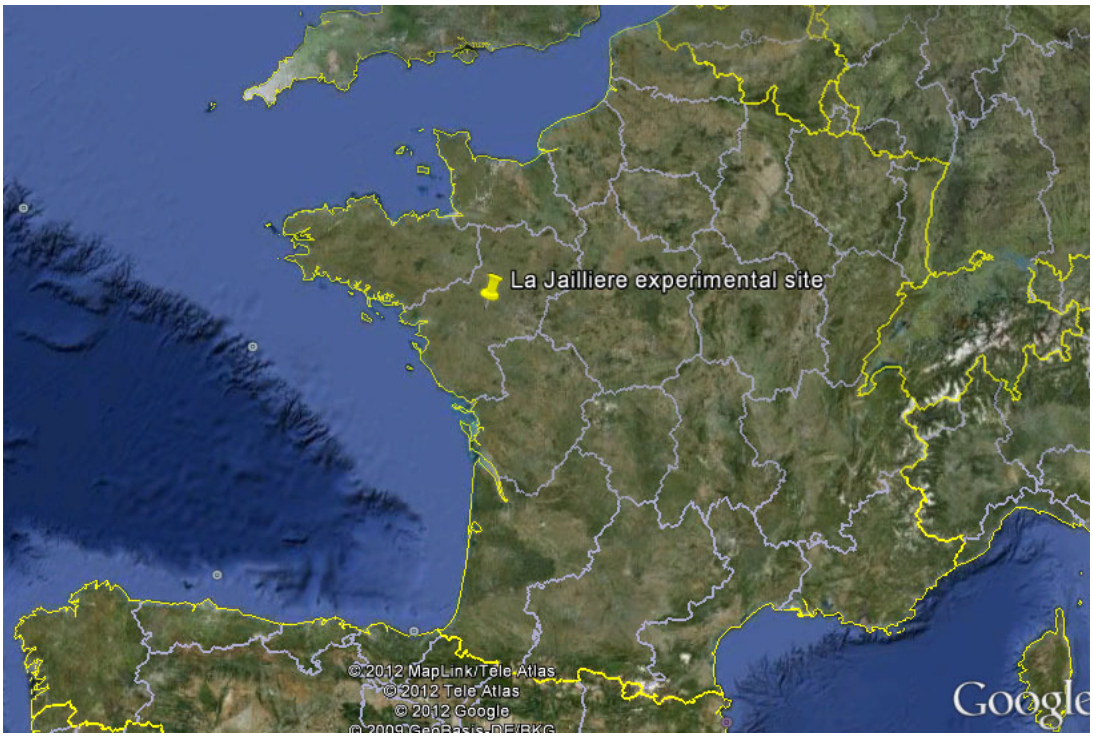


Figure 3: *France*. Location of the La Jaillière experimental site.

* More details at http://www.arvalisinstitutduvegetal.fr

The site has been dedicated to the study of the influence of agricultural management practices on water quality since 1987. It is a reference site for the European Commission FOCUS working group (FOCUS, 2001). The La Jaillière site is considered as a representative of the agricultural regions in Europe with shallow silt clay soils.

Soils are hydromorphic brown with a silt clayed texture, and shallow schistose bedrock situated at about 0.90 m below the surface. The average clay content is 22 % (Madrigal, 2004), but variations from 18 % to 30 % were observed depending on the soil horizons (Arlot, 1999). Organic matter content was found to be on average 2 % in the superficial soil horizon (Madrigal, 2004).

The climate at the site is of oceanic type. The mean annual precipitation of 617 mm is evenly distributed along the year (monthly values between 40 and 62 mm). The mean annual potential evapotranspiration is 610 mm.

The site contains many fields divided in north and south parts. Furthermore, each part contains blocks of fields (Figure 4). Each block is used for a different type of experimental analyses. Our data were collected from the fields in block A11 (Figure 5).

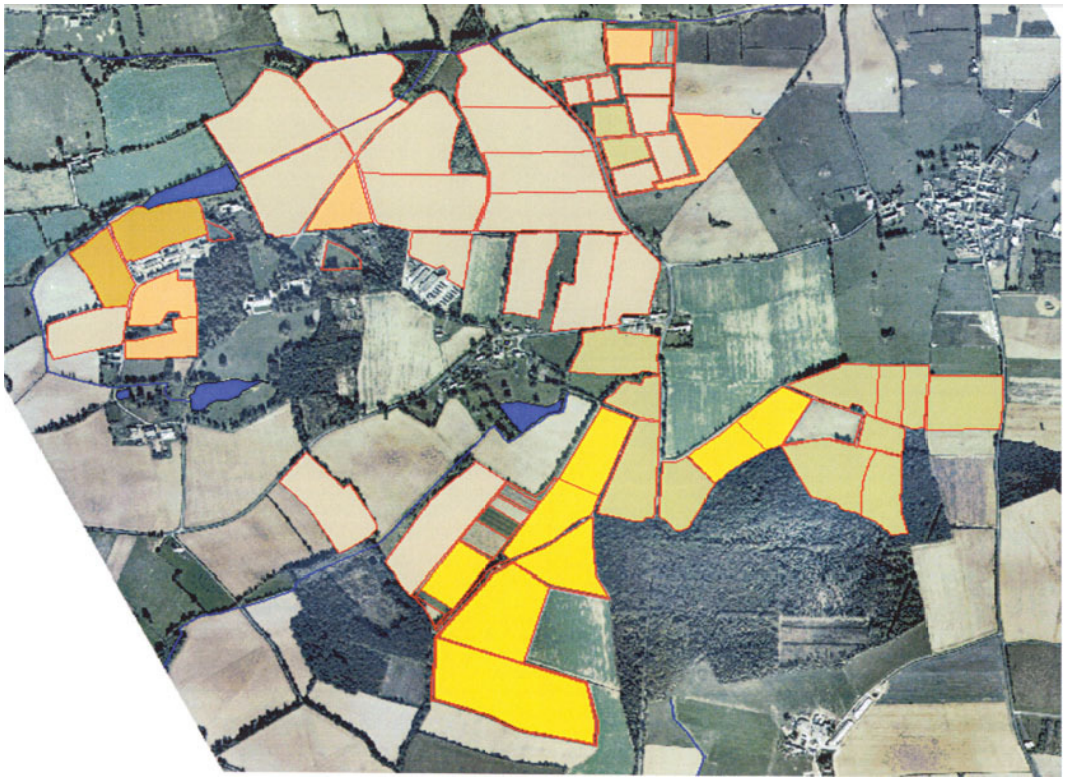


Figure 4: *La Jaillière site*. All fields on the site (Selected papers from ARVALIS No.11).

Figure 5: *Block A11*. The fields within block A11 of the La Jaillière site (Selected papers from ARVALIS No.11).

Table 1: *Fields and water collection*. The description and size for each field considered in our study, together with the type and starting year of water collection

Field	Block	Surface (ha)	Type of water collection	Starting year
T1	A11	0.83	Runoff	1987
T2	A11	0.90	None	1987
T3	A11	1.04	Drainage/Runoff	1987
T4	A11	1.08	Drainage/Runoff	1987
T5	A11	0.85	Drainage	1987
T6	A11	1.01	Drainage/Runoff	1987
T7	A10	0.43	Drainage/Runoff	1989
T8	A10	0.43	Drainage/Runoff	1989
T9	A10	0.34	Drainage/Runoff	1989
T10	A11	0.42	Drainage/Runoff	1991
T11	A11	0.42	Drainage/Runoff	1991

In our study, we have included data from 8 fields within block A11, plus 3 fields from block A10 (Table 1). Each field is about, or less than, 1 ha of surface area and is cultivated following a traditional winter wheat/corn crop rotation. They are equipped with an independent tile drainage system and surrounded by metal cuttings for hydraulic isolation from the other farm fields and with a collecting trap for surface runoff measurement (Figure 6). Tile drains are located at depth $d = 0.9$ m below the soil surface, with a spacing of 10 m (Branger et al. 2006).

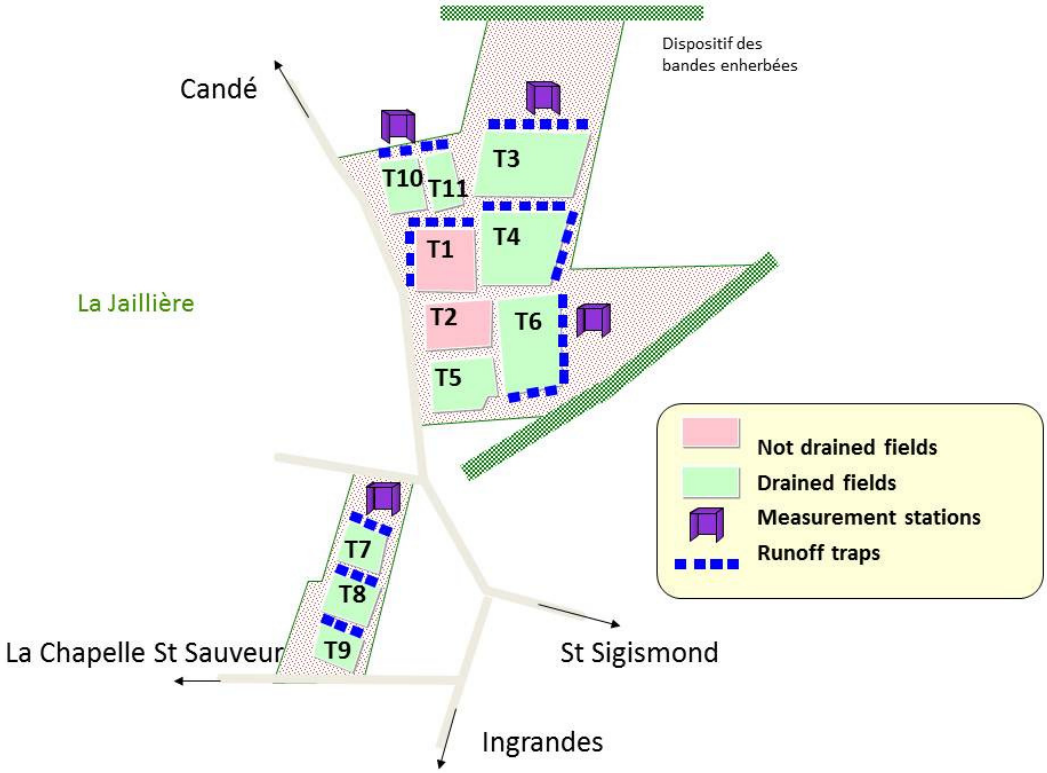


Figure 6: *Field's scheme*. The layout of the fields with their names, runoff and drainage system characteristics and the location of the stations where water is collected (Beard, 2005)

Three stations where water is collected are located in the block A11 (Figure 6). The water is collected from drainage and runoff separately for each field. Since 2005, a small meteorological station is installed on the site. This station gives information on the temperature, evaporation, and amount of rainfall. Furthermore, from the available meteorological data, the following is derived: minimal, average and maximal temperature per day, evapotranspiration per day, and amount of rainfall per day.

Below we describe the procedure of collecting water from the experimental site in the La Jaillière region. Almost all fields, except field T1 and T2, have a drainage system installed. The drainage pipes are located 0.9 m below the soil surface, with a spacing of 10 m (Branger *et al*. 2006). All of the pipes from one field are connected in a single pipe that ends up at the nearest station (Figure 7).

On the other hand, at the experimental site, there is a system for channelizing the runoff water or water that flows over the surface horizon. As we are able to see from Figure 6, with the exception of T2 and T5 fields, all the remaining fields are equipped with runoff traps. The runoff traps are located on lower-elevation sides along the edges of the fields and end up in water station, as well.

Figure 7: *Water collection at the La Jaillière fields*. Schemes for the drainage and runoff systems: Piège à ruissellement – Runoff traps; Réseau de drainage: Tile drainage network (Selected papers from ARVALIS No.11).

Figure 8: *Water collector*. Drainage and runoff collectors for measuring the flow rates and collecting water samples (Selected papers from ARVALIS No.11).

One water station consists of water collectors for measuring the flow rates and collecting water samples (Figure 8). Water collectors are separated for drainage water and runoff water. In a period of flow, an average sample is taken every week from each of the plots and is then sent to the laboratory for analysis. Otherwise, the flow is monitored at an hourly time step but recorded in the database as a cumulative daily amount of water for runoff and drainage, separately.

3.2 The data

As we mentioned before, our study is based on data for the traditional agricultural practices performed on the fields and the amount of water flown out of the fields (the PCQE database), the soil properties, and the meteorological data. Hereafter, we describe each type of data collected at the La Jaillière experimental site.

3.2.1 Soil properties

Soil properties are recorded in a soil database. This database covers different types of soil that can be found in France. The soil properties for the La Jaillière experimental site are included in this database.

Namely, each soil type registered in the database has a unique code by which some type of soil can be recognized. The characteristics of the soil in the La Jaillière site can be found under the reference PL0133500. Although soil variability has been determined at six points across the site, the soil in La Jaillière is referenced under the same code, as the variability is not sufficient to represent different types of soils on the site.

The soil is medium loamy over clay, affected by seasonal waterlogging, which is evident from the grey mottles in deeper layers. Furthermore, in the subsurface layer the soil is slightly to moderately stony, while at the depth of about 60 cm it is getting to become very to extremely stony. The soil across the site varies slightly in the texture, too. There are small differences in the stone content of top-soils, but it is not considered sufficient to affect the classification of the soil profiles.

The database contains data about the physical, hydrological and chemical characteristics of the soil for each horizon. The substratum is described with its nature, texture, consistence and level of permeability.

3.2.2 The PCQE database

In the present work, we used the data available in the PCQE database owned by ARVALIS. The complete database consists of three types of data: agricultural practices, water flow monitoring data, and crop protection practices. The data were collected during the period of 1987–2011.

In our study, we are trying to learn a model that will predict the amount of water going out of the field, based on the data from the PCQE database. For the task of water flow prediction, we need data related to water inputs (like rainfalls and irrigation), evapotranspiration, crop and land management, and water outputs (like drainage and runoff water flow). We used only these data from the PCQE database, while data related to crop protection practices were excluded. Furthermore, we used the meteorological data presented in Section 3.2.3.

Hereafter, we present the data related to agricultural (crop and land management) practices. Next, we describe the monitored flow of drained water which was considered as the output of our study. Finally, we describe the meteorological data considering rainfall, temperature and evapotranspiration.

3.2.2.1 Agricultural practices

The PCQE database records all crop and land management practices that have been performed on the fields. Since 1987, monitoring field practices is a standard procedure in order to produce better data, which can be further used in experiments and analyses. Each record from the database describes one operation: It contains the date when operation was performed, the type of the operation, the materials used, and some additional information. Moreover, the crop present on the field where the operation was performed is monitored and registered in the database. In Table 2, we present the complete set of information collected for a performed operation.

Table 2: *Field practices*. The information recorded when some agricultural practice is performed.

Practice	Field	Crop present	Date	Material	Dose
Fertilization	✓	✓	✓	✓	✓
Irrigation	✓	✓	✓	×	✓
Phytochemical protection	✓	✓	✓	✓	✓
Harvesting	✓	✓	✓	✓	×
Tillage	✓	✓	✓	✓	×

The process of water flow is highly dependent on irrigation, harvesting and tillage practices. Therefore, these practices have been considered in the process of learning our model for predicting the amount of drained water flow. Although some of these practices are not suitable to be included directly in the process of learning the model, the problem has been overcome with pre-processing techniques, which will be described in Section 4.

3.2.2.2 Water flow

As we described before, at the La Jaillière experimental site, two types of water flow are observed: drainage and surface (runoff) water flow. In this study, the drainage water flow has been considered as the output, and runoff as an input attribute, because it is a water flow on the surface of the soil and it is easily noticed. In addition, during a drainage season, the presence of runoff can be a significant aftereffect of extreme drainage events.

The drainage and surface runoff rates are routinely monitored at an hourly time step, but the data recorded in the PCQE database are based on cumulative values per day. A day for water flow observations is defined as the time period that starts at 00:00 (midnight) and lasts for 24 hours. Note that the day for meteorological data is defined differently as we will describe in the next sub-section. Furthermore, the water flow has been observed for each field, separately.

We have observations for drained water flow for each day in a campaign. A campaign is defined as the time period that starts on September 1st, and finishes on August 31st. A total of 25 campaigns were observed for each field where a drainage system is installed (Fields T3 – T11, Table 1). The observations for runoff are similar and are available for each field where runoff traps bound the field's edges (Fields T1, T3, T4, T6 – T11).

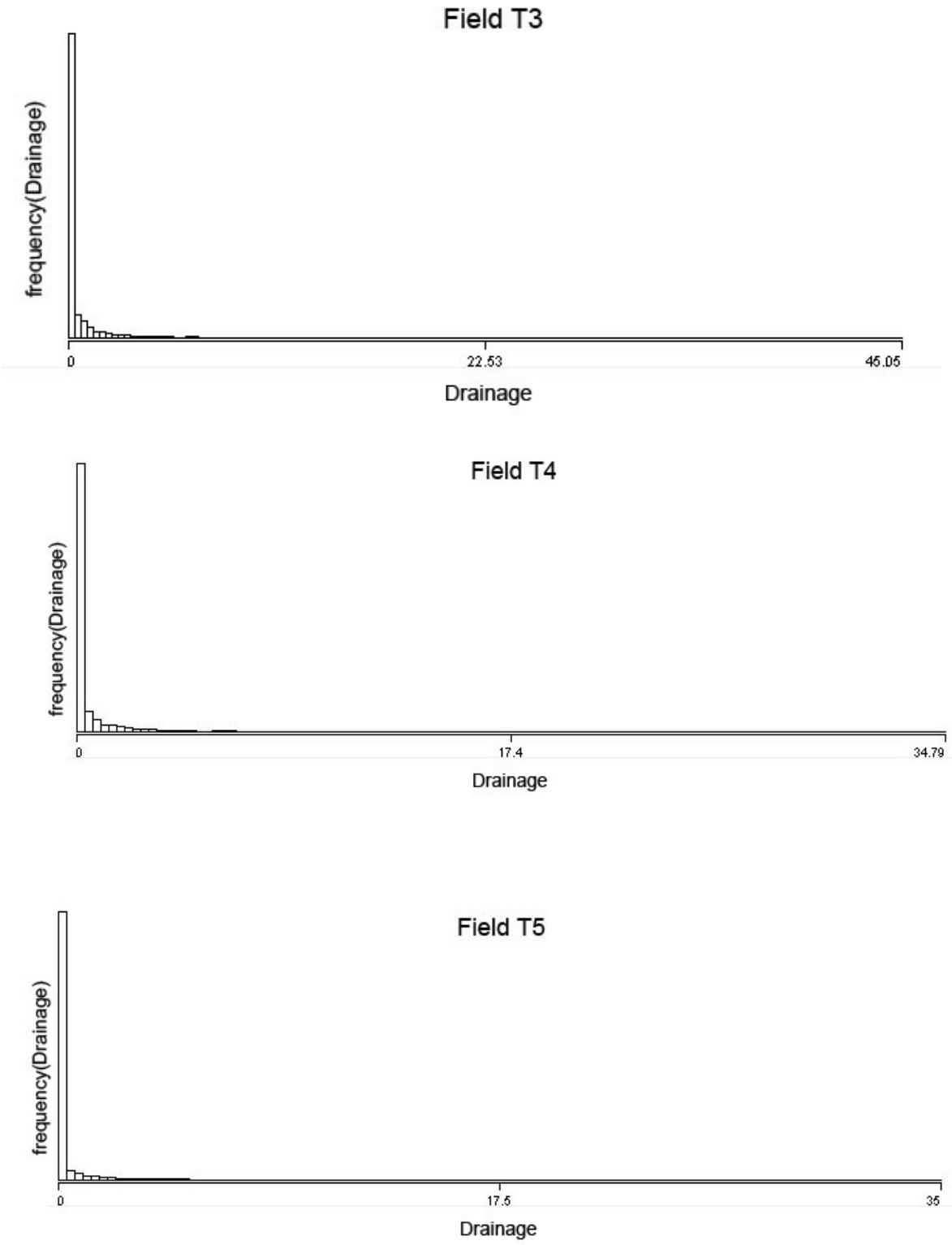


Figure 9: *Drainage quantity value distributions*. The distribution of the drainage quantities over each of the 25 campaigns. The frequency is the number of days when a particular drainage quantity was observed.

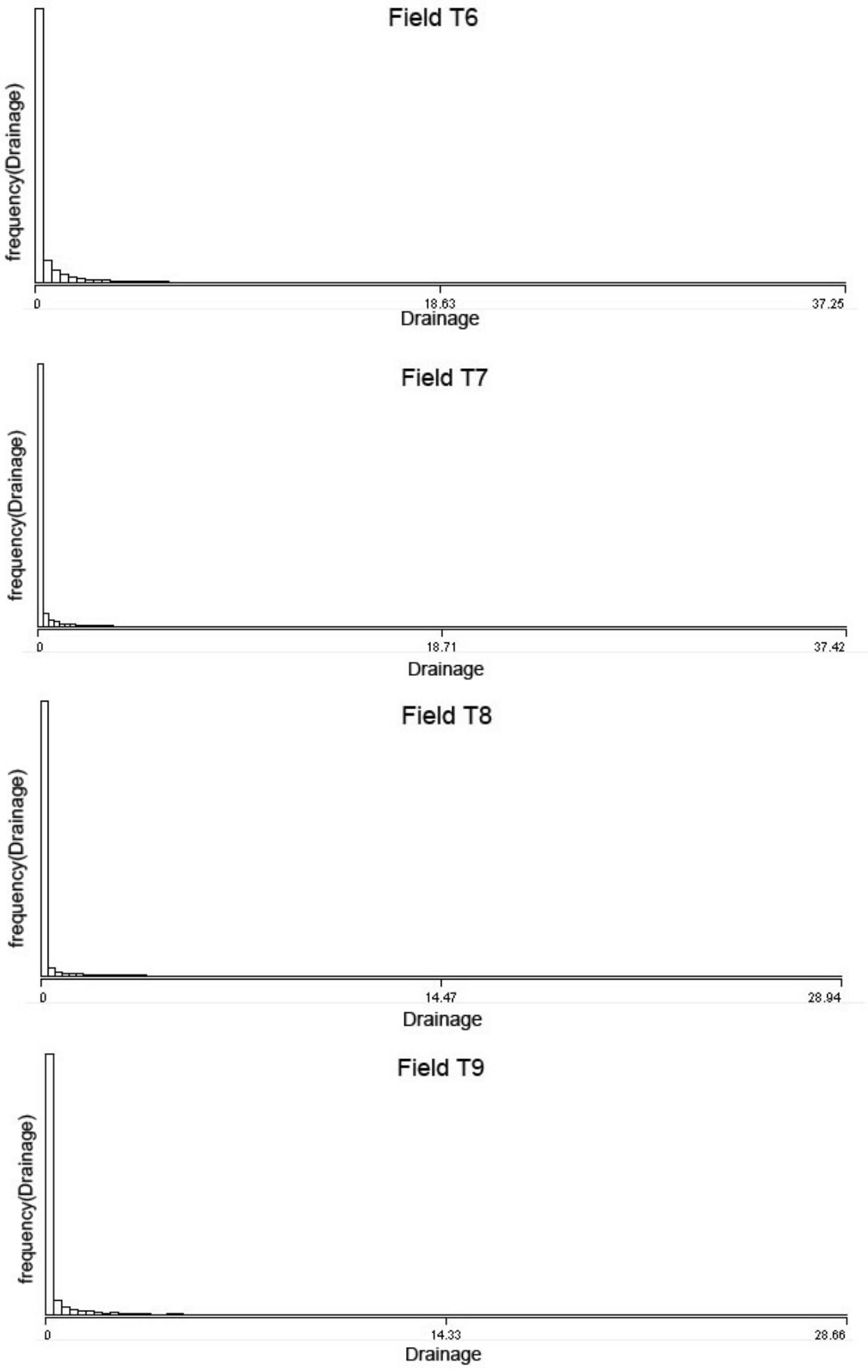


Figure 9 *(Continued)*: *Drainage quantity value distributions*.

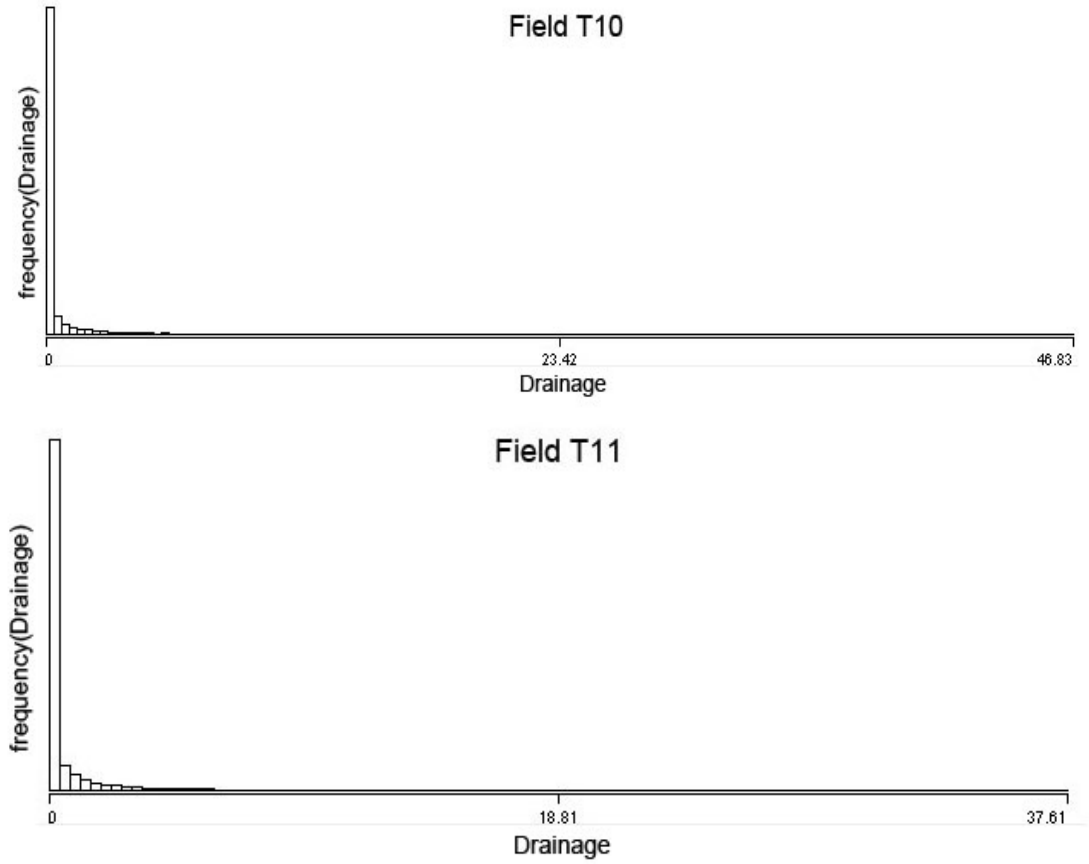


Figure 9 *(Continued)*: *Drainage quantity value distributions*.

Figure 9 presents the distribution of the drainage quantity values over the 25 campaigns for each field, separately. The x-axis depicts the possible drainage quantity values, while the y-axis gives the number of the days with a specific drainage quantity. The histograms depict the drainage quantity values and their frequencies, showing that all of the fields have an exponential distribution of drainage quantity values. Furthermore, in Table 3, we present the basic properties of the distributions. The distributions of values for Fields T1 and T2 are not presented because there is no drainage system installed on these fields.

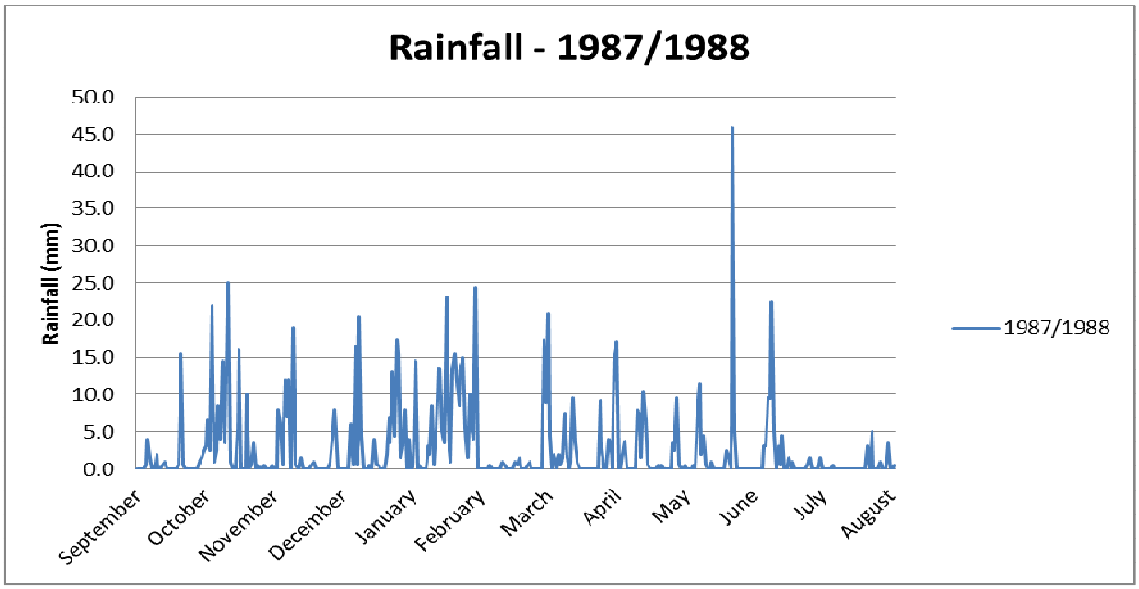
Table 3: *Properties of drainage quantity value distributions*. Basic statistics of the Drainage attribute given across all 25 campaign.

Field	Min Value	Max Value	Mean Value	Standard Deviation
T3	0	45.05	0.631	1.976
T4	0	34.79	0.591	1.881
T5	0	35	0.499	1.913
T6	0	37.25	0.675	2.279
T7	0	37.42	0.357	1.471
T8	0	28.94	0.338	1.499
T9	0	28.66	0.459	1.711
T10	0	46.83	0.542	2.064
T11	0	37.61	0.672	2.253

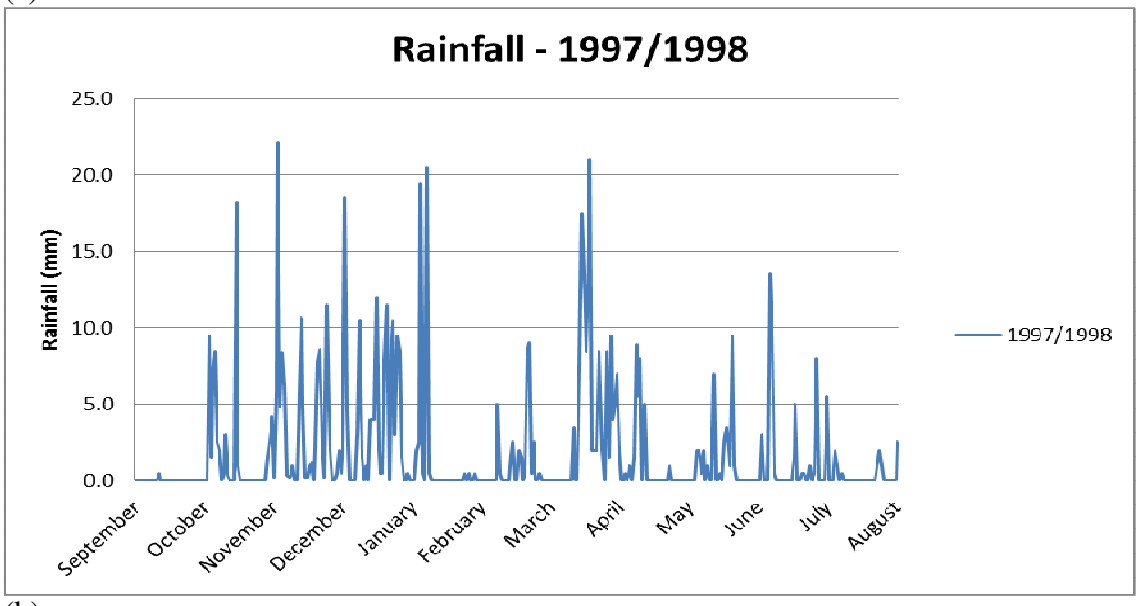
3.2.3 Meteorological data

The meteorological data were collected from two sources. Starting from 1987, data were taken from Météo France, the French national meteorological agency. Météo France has a wide range of stations in France, some of them being in the surroundings of La Jaillière. Therefore, the data were collected from the nearest meteorological station to La Jaillière, referenced with number 4499. During the year 2005, ARVALIS installed their own meteorological station at the La Jaillière experimental site. Since January 1st, 2006, the data were collected from the ARVALIS meteorological station located at the site.

The meteorological data are distributed as a separate database and contain information about the minimal, mean and maximal temperature per day, cumulative evapotranspiration per day and cumulative rainfall per day. It is important to note that day here is defined as the period that starts at 06:00 and lasts for 24 hours, which is different from the PCQE database, described before, where it start at 00:00.

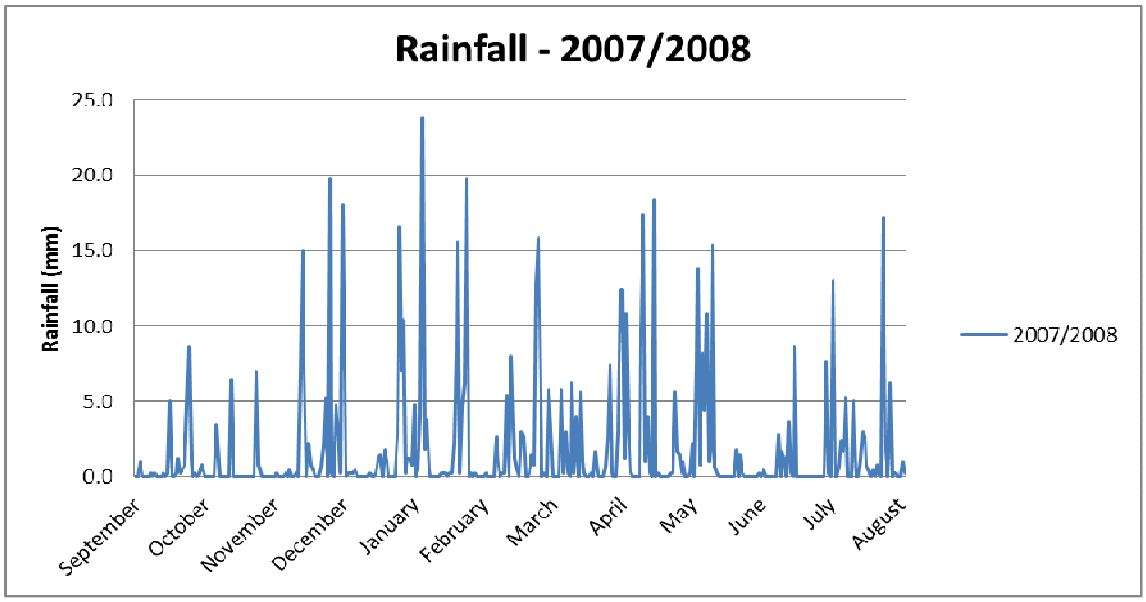


(a)



(b)

Figure 10: *Rainfall*. The rainfall quantities for 3 different campaigns.

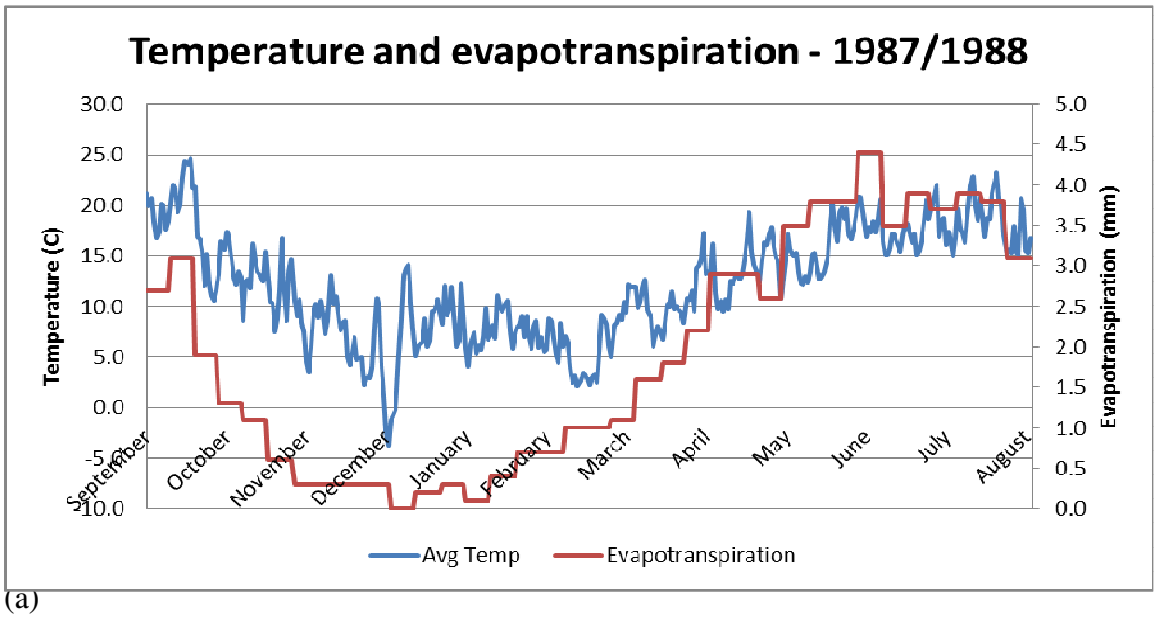


(c)

Figure 10 *(continued)*: *Rainfall*.

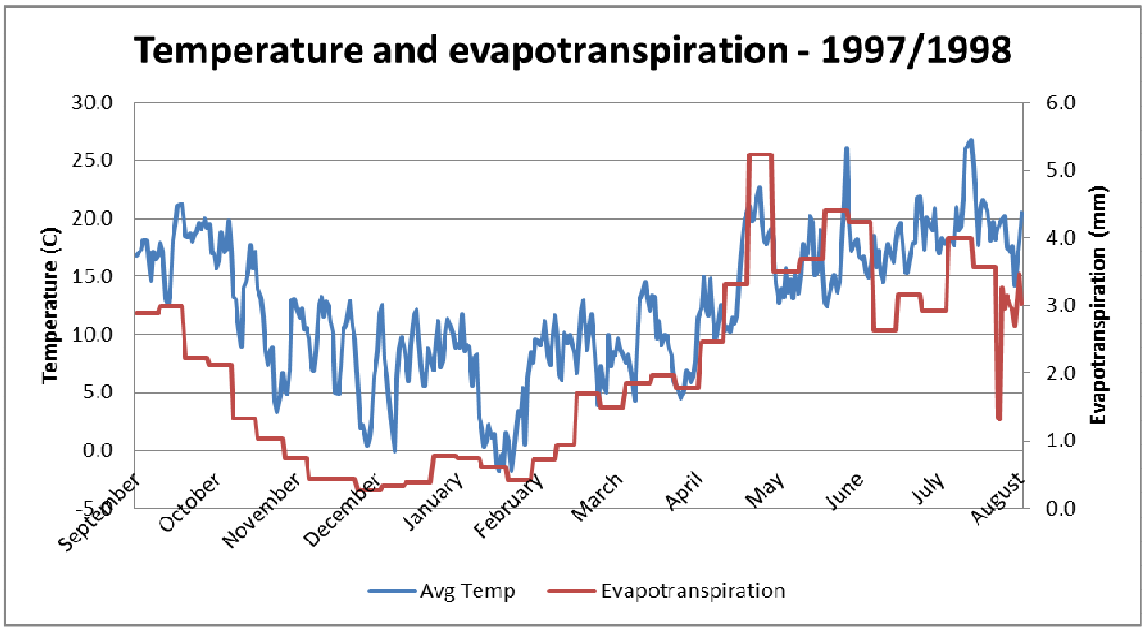
In Figure 10, we present the distribution of rainfall over three campaigns, where each graph showing one campaign. One campaign is a period that starts on September 1st, and finishes on August 31st, the following year. One campaign is representative for one agricultural season. Each chart in Figure 10 is plotted from daily cumulative rainfalls measured in mm and stored in the meteorological database by ARVALIS. The 3 presented campaigns are: 1987/1988, 1997/1998 and 2007/2008.

In addition, Figure 11 presents the average temperature per day and evapotranspiration per day for three campaigns 1987/1988 (Figure 11a), 1997/1998 (Figure 11b) and 2007/2008 (Figure 11c). It is worth mentioning that evapotranspiration is a derived value, which depends on temperature. On the charts, it is easy to perceive the dependence between temperature and evapotranspiration.

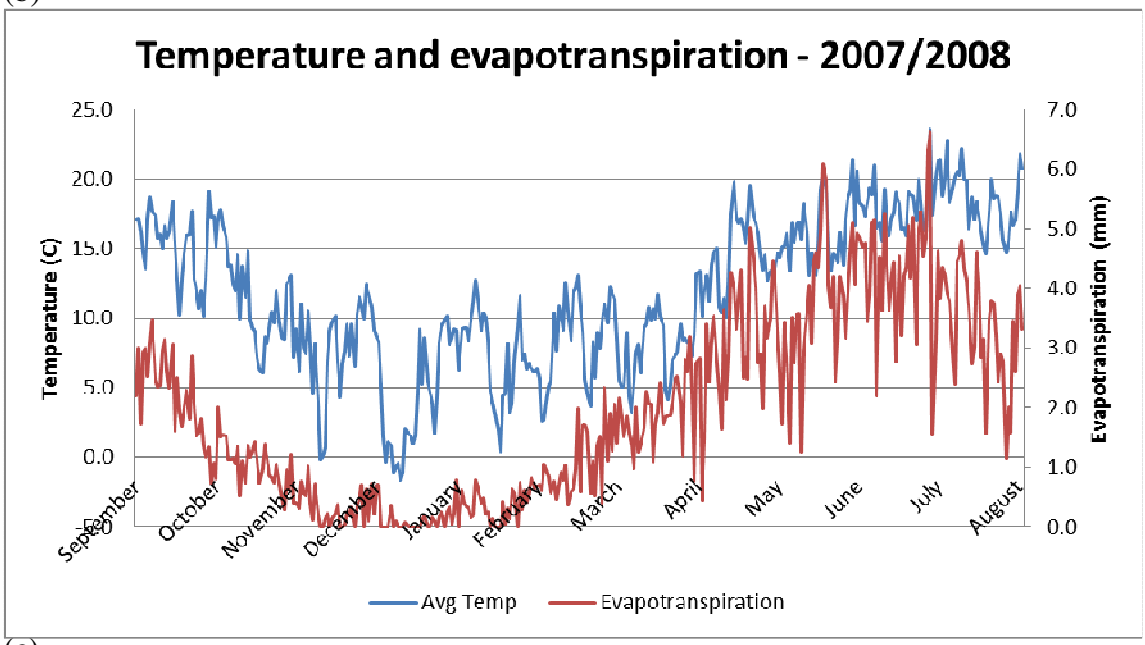


(a)

Figure 11: *Average temperature and evapotranspiration*. The average temperature and evapotranspiration for 3 different campaigns.



(b)



(c)

Figure 11 *(continued)*: *Average temperature and evapotranspiration.*

3.3 Drainage seasons

Within a campaign, a few periods can be identified during which extreme drainage events are registered. These periods are named drainage seasons. Usually there are two drainage seasons during a campaign: a winter and a spring drainage season. Given the names of the drainage seasons, it is easy to roughly define the time of their appearance, but the beginning and ending days of the winter and spring drainage seasons are not the same with the starting and ending days of the summer and winter seasons. Instead, there are some rules and conditions for defining a drainage season.

Based on their experience, the domain experts (B. Real & J. Maillet-Mezeray) defined a rule for the start of the winter drainage season. Namely, the winter drainage season starts on a day when the cumulative amount of drained water since the start of the

campaign (September 1st) exceeds the threshold of 5 mm. The end of the winter drainage season depends on several conditions, but only one is well defined: a winter drainage season finishes when the cumulative amount of drained water in the previous 7 days has not exceeded the threshold of 1 mm. On the other hand, the actual dates obtained for the start and the end of a drainage season according to these rules, are not the same as those obtained from the experts. Therefore, we analyzed the dates obtained from the experts in order to extract new knowledge which is related to the conditions for defining the start and the end of a drainage season. The dates obtained from the experts are given in a data set which is presented in Appendix A. It contains the dates for the start and the end of the drainage periods for each field, for each campaign in the period 1987–2011.

3.4 Expert knowledge

In this study, we have considered that the available expert knowledge (EK) provided in the form of tables should be taken into account and integrated in the process of learning a predictive model from data. Therefore, besides using the data described in the previous sub-section, the process of learning predictive models for predicting the drained water flow from a field has been supported by the available expert knowledge. The available expert knowledge is manually formulated in the form of decision rules and written in tables. It is an integral part of the Aquanouveau system owned by ARVALIS - Institut du Vegetal.

The complete structured expert knowledge contains 7 modules. In our study, we will use Module 1, Module 2, Module 4 and Module 6 (Figure 12). But, for this particular thesis we focus our attention on the tables from Module 1 and 2 only. These concern the diagnosis of water flows from the fields.

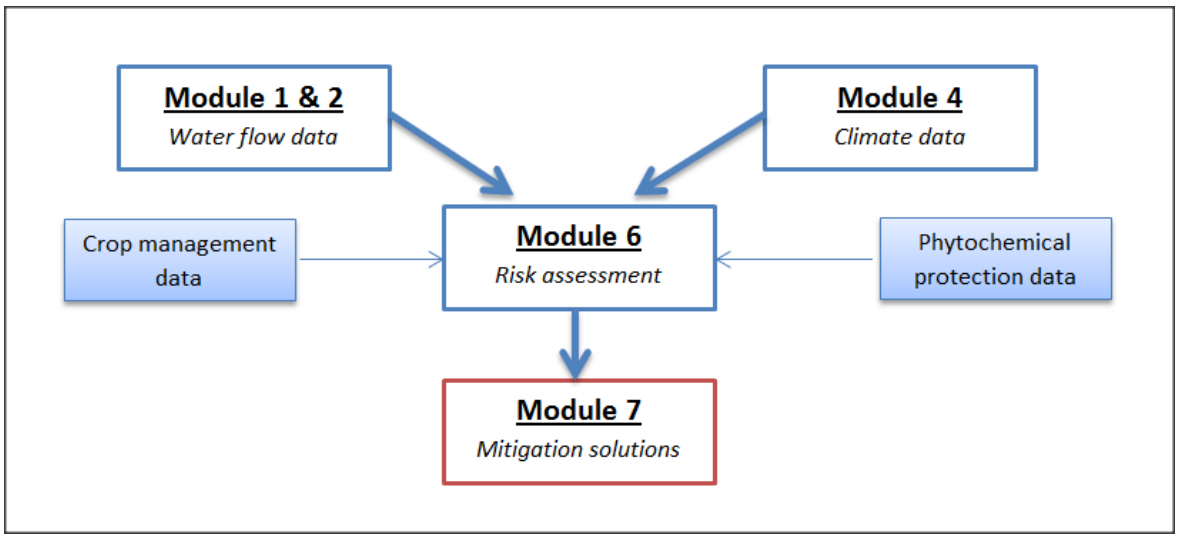


Figure 12: *The scheme of the Aquanouveau system*. The modules of the available expert knowledge written in the form of decision rules in tables and their interactions.

Module 1 consists of decision rules that assess the risk of different types of water flow from a field, which is cultivated with none of the known agricultural practices. The types of water flow considered in Module 1 are: runoff, lateral seepage (sub-surface flow), infiltration and erosion.

Unlike Module 1, Module 2 refers to assessing the risk of water flow from both cultivated and uncultivated fields. The fields with a drainage network installed are considered in Module 2. Therefore, we include the rules from Module 2 in the process of

learning predictive models for predicting drained water flow.

Namely, Module 2 contains 34 partial tables (such as the one given in Table 4), divided in 3 parts depending on the weather season: autumn-winter, spring and summer. Beside these tables, Module 2 contains some additional information, given in textual documents, which will be used as input attributes. The complete process of using the available expert knowledge is described in the Section 4.

Table 4: *Part of the EK*. Assessment of the types of flow in the soil in summer, on permeable substrate with break in permeability, with drain performing poorly and with plough pan

Depth of permeability disruption	Cracks	Lateral seepage on plough pan by cracks	Infiltration by cracks	Transfer by drainage
<40	No	Nonsense	Nonsense	Nonsense
<40	Yes	Nonsense	Nonsense	Nonsense
40 to 80	No	Null	Null	Null
40 to 80	Yes	Low	High	Low
> 80	No	Null	Null	Null
> 80	Yes	Null	High	Low

4 Data preprocessing

The pre-processing phase starts with the tasks of available expert knowledge and converting these partial tables and accompanying text into a single table. A single table is a suitable format to use machine learning techniques in order to restructure the expert knowledge.

The phase continues in the direction of database manipulation and data pre-processing. First, a feature set must be defined from the available data. Second, feature ranking and feature selection are performed in order to achieve a reduced feature set. The reduced feature set contains only attributes that are most relevant for the given problem. Finally, the pre-processing phase finishes by completing the feature set with features that describe the soil properties and some additional features recommended by the experts.

4.1 Expert knowledge pre-processing

As mentioned before, the expert knowledge is hand-crafted and all of the modules are distributed in the form of documents (ARVALIS internal report 1 from the IDV024 project). The documents contain text, which gives additional information that will be used as input attributes (Table 4). Furthermore, the hand-crafted rules are divided in 3 groups. Each group represents one weather season (autumn-winter, spring and summer).

In the first stage, techniques for feature engineering i.e. invention have been employed and new features have been introduced into the set of rules. These features were derived from the available information present in the text within the documents. Some invented features contained complex information. Such complex information can usually be represented by two or more features. Therefore, we used the pre-processing technique of feature extraction to represent these types of information contained in the complex features.

Table 5: *Propositionalization of the expert knowledge*. An example of the features constructed from the text and feature extracted from complex valued features.

Season	Soil Type	Permeability	Cracks in permeability
Autumn-Winter	Permeable substrate with cracks in permeability	Yes	Yes
Autumn-Winter	Impermeable substrate with cracks in permeability	No	Yes
Spring	Permeable substrate with cracks in permeability	Yes	Yes
Spring	Impermeable substrate with no cracks in permeability	No	No
Summer	Permeable substrate with cracks in permeability	Yes	Yes
Summer	Impermeable substrate with cracks in permeability	No	Yes
Summer	Permeable substrate with no cracks in permeability	Yes	No

For example, the introduced feature “SoilType” has a complex value, which contains information of the permeability of the soil and the presence or absence of cracks in permeability (Table 5).

The phase of pre-processing the rules resulted in a dataset that contains all available partial tables provided by the expert and all important additional information given in the text documents. The last step of the rule preparation is the simple step of adding the already existing features, defined in the tables of expert knowledge. The number of instances, input and target attributes of the complete dataset are given in Table 6.

Table 6: *Complete dataset constructed from the available expert knowledge*. A quantitative description of Modules 1 & 2 of the available expert knowledge.

	Module 1	Module 2
No. of tables	12	34
No. of input attributes	12	13
No. of target attributes	11	12
No. of instances	312	6816

The final dataset is a single table that contains all the information that has been provided by the experts. This format is suitable for further processing of the expert knowledge, which includes the restructuring of the expert knowledge in the form of decision trees. Finally, the restructured expert knowledge has been validated with 100% correctly classified examples over the training set, due to the fact that it will not be used for classification of new examples, but only optimized for further integration (Kuzmanovski *et al.*, 2012). We present the results of the completed processing of the expert knowledge in Section 7.3.

4.2 Database pre-processing

The data is stored in a relational database system, which represents information about the entities and the relationships among them. The representation of the entities and their relationships is in form of relational tables that consists of rows and columns. The rows and columns represent examples and features, respectively.

First of all, we transformed all of the relational tables into one single table. This format is suitable for the further use of machine learning techniques. Then, we introduced new attributes from the existing once with the purpose of generating attributes that will describe the output in the best way. Finally, with feature ranking and feature selection techniques, we defined the final feature set, which describes the dependent variable or the target attribute, in the best way.

4.2.1 Independent variables

In this sub-section, we describe the explanatory or independent variables that have been defined from the data in the PCQE database. Next, we present the extraction of the meteorological data. Finally, we present the data from the soil properties database that has been included in the dataset.

4.2.1.1 Crop management and agricultural practices

Since the problem has been defined as the problem of predicting drained water flow per day, it is clear that the basic time step is one day. Therefore, one example (also referred to as an instance or record) was defined as a vector of feature recordings for one day. The

first day recorded in the dataset was September 1st, 1987, while the last day that will be considered in the analyses and recorded in the database is August 31st, 2011. Moreover, each day in a campaign has been recorded in the PCQE database. Therefore, we have examples for each day in a campaign. Data for 22–25 campaigns were available from the PCQE database, for each of the 11 fields in the La Jaillière experimental site.

In total, 96426 examples were extracted from the relational PCQE database. Moreover, from the information extracted before, four features were introduced: *Field*, *Campaign*, *Date*, and *Day* (in a campaign). The *Date* and *Campaign*, which identify a day and the corresponding data record, due to possibility of overfitting, have been considered as features, only during the preprocessing phase and have been excluded from the dataset in the experimental phase of applying machine learning.

All operations and practices performed on the field are recorded in a separate relational table in the PCQE database. As described before, the following 4 main practices have been recorded in the database since the experimental site has been established: tillage, irrigation, plant protection, and harvesting. Under tillage, the practices of sowing and ploughing were considered. Moving and production are recorded as a part of general harvesting practice.

Each of the practices performed on the field is saved in the PCQE database, together with the date and the material used. Therefore, we have information of what crop is found on the field on each day that a practice is performed. That crop is on the field during the time period between the sowing and harvesting dates. This led to the introduction of a new feature into the dataset.

The crop has a different influence on the water in the soil in different periods after sowing. Because of this, a *crop development coefficient* has been introduced as a new feature in the dataset, which describes the amount of water that the crop takes from the soil, while the crop is present on the field. The crop development coefficient depends on the type of the crop (Table 7). The list of all explanatory (independent) variables that have been defined from the data of PCQE database is given in Table 8.

Table 7: *Crop development coefficient*. This coefficient expresses the water taken by a crop at different stages of crop development, i.e. at different months during the year.

Crop	1	2	3	4	5	6	7	8	9	10	11	12
Wheat	1.00	1.00	1.00	1.10	1.20	1.10	0.70	0.60			0.60	0.80
Maize					0.60	0.80	1.15	1.15	0.90	0.90	0.80	0.70
Spring barley			0.60	1.00	1.10	1.20	0.70	0.60	0.60			

Table 8: *Features constructed from PCQE database*. Features defined by the data from the PCQE database and included in the constructed dataset.

Source	Feature	Description
PCQE database	Field	Name of the field
	Campaign	A campaign is the period from September 1st, and finishes August 31st, following year.
	Date	
	Day	The consecutive day in the campaign, 1st of September being the first day.
	Crop	Crop that is on the field at the time of observation.
	CDCoef	Crop development coefficient

4.2.1.2 Meteorological and water input data

The meteorological data are available in separate database. The time step is a day, as for the transformed data from the PCQE database. Therefore, no additional transformation was needed in order to incorporate the meteorological data in the constructed dataset.

The process of water flow through the soil mostly depends on the water input as well as the crop and soil properties. The defined water input is described through the rainfall, irrigation and evapotranspiration (Figure 13). On its way from the input to the output, i.e. during its flow through the soil, the water is delayed. The delay is different for different geographical regions and soils. Therefore, we do not have exact information about the delay. This means that data about the input in the previous periods are required in order to approximate the time delay using machine learning and data mining techniques. Hence, we included the daily and cumulative water input information for several periods of time:

- 1 day ago (yesterday)
- 2 day(s) ago
- 3 day(s) ago
- 4 day(s) ago
- 5 day(s) ago

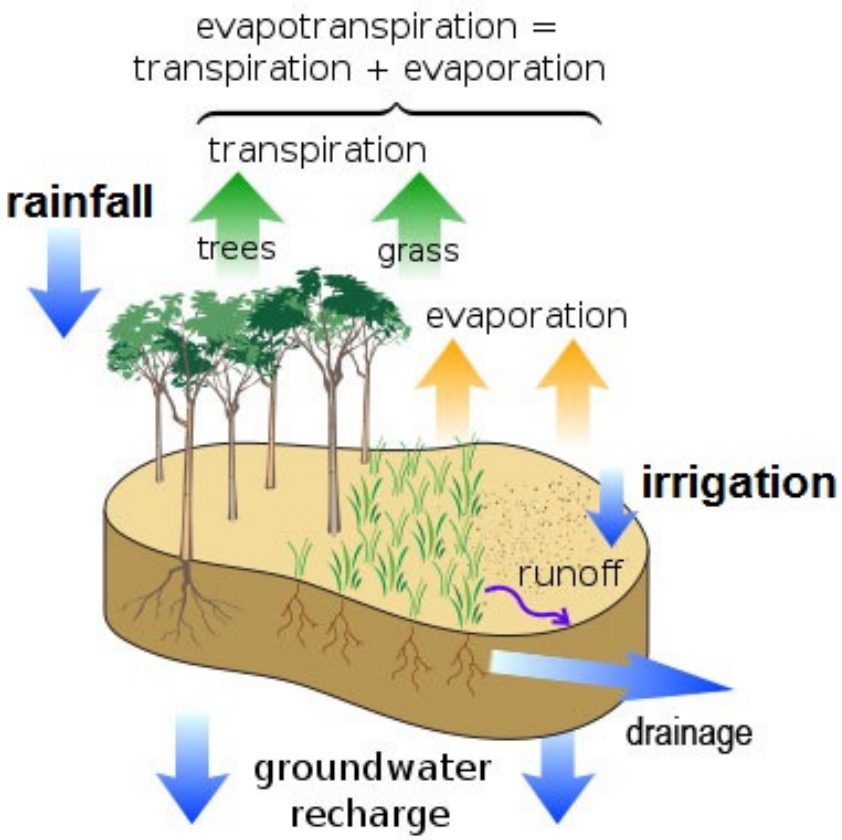


Figure 13: *The water cycle for a field*. The roles of rainfall, irrigation and evapotranspiration in the global water cycle (http://en.wikipedia.org/wiki/Evapotranspiration).

The water input, as described above, includes data on the cumulative amount of rainfall per day, cumulative evapotranspiration per day, and cumulative amount of irrigation per day.

A comprehensive survey of the literature on the topic of water flow through the soil (referenced in background section) emphasizes that temperature influences the process of water flow and the amount of water that stays in the soil. The temperature is included in the calculation of the evapotranspiration, as we noted in the temperature and evapotranspiration charts, where we mentioned that there was a linear dependence between temperature and evapotranspiration values. However, we include the average temperature per day and the averages from previous periods (days) in the dataset. The temperature for the previous periods are calculated in the same way as the water input, but instead of cumulative values, for temperature we used average values per day.

Furthermore, beside the temperature and water input, the experts recommend that runoff should be considered as an input variable. Although, in reality, runoff is a product of saturated or capping soil, we considered that it gives important information about the drained water. Namely, when runoff appears on the surface of the soil because of soil saturation, then in the soil the drained water is flowing with full capacity. There are certain exceptions of this but until now we considered this to be sufficient information, which can bring some improvements to the learned predictive model.

4.2.1.3 Soil properties data

The soil at the La Jaillière experimental site is uniform. Therefore, this data will not give any advantages in the process of learning the predictive model. Hence, it is excluded from the process of learning from data, but will be considered when expert knowledge is integrated within the learned predictive model.

The only relevant information from the soil properties is the slope. Namely, each field has a different slope of the surface layer. Therefore, the slope is included in the constructed dataset.

4.2.2 Dependent variables

In this thesis, we consider the drained water flow as a dependent variable or target attribute.

The literature referenced in the introductory section, describes the process of drained water flow as a continuous process. Hence, the present state of the drained water is related to the previous states (Figure 14). Therefore, we include the data about the drained water flow at the previous time points as independent or descriptive variables. The previous values of drained water flow were derived by the same scheme as we derived features from meteorological data. This means that we include data for drained water for:

- 1 day ago (yesterday)
- 2 days ago
- 3 days ago
- 4 days ago
- 5 days ago

This completes the feature set for the constructed dataset. The complete list of attributes that were ranked by the feature ranking techniques is presented in Table 9.

Table 9: *Features of the constructed dataset*. Features derived from the different databases and included in the constructed dataset.

Source	Feature	Description
PCQE database	Field	Name of the field
	Campaign	A campaign is the period from September 1st of the current year, and finishes August 31st, of the following year.
	Date	
	Day	The consecutive day in the campaign, 1st of September being the first day.
	Crop	Crop that is on the field at the time of observation.
	CDCoef	Crop development coefficient
	Irrigation	Today's amount of irrigation
	IrrigationN1	Yesterday's amount of irrigation
	IrrigationN2	Amount of irrigation 2 days ago
	IrrigationN3	Amount of irrigation 3 days ago
	IrrigationN4	Amount of irrigation 4 days ago
	IrrigationN5	Amount of irrigation 5 days ago
	IrrigationA1	Cumulative irrigation for today and yesterday
	IrrigationA2	Cumulative irrigation for today and the 2 days before
	IrrigationA3	Cumulative irrigation for today and the 3 days before
	IrrigationA4	Cumulative irrigation for today and the 4 days before
	IrrigationA5	Cumulative irrigation for today and the 5 days before
	Runoff	Today's measured runoff
	RunoffN1	Yesterday's measured runoff
	RunoffN2	Runoff measured 2 days ago
	RunoffN3	Runoff measured 3 days ago
	RunoffN4	Runoff measured 4 days ago
	RunoffN5	Runoff measured 5 days ago
	RunoffA1	Cumulative runoff for today and yesterday
	RunoffA2	Cumulative runoff for today and the 2 days before
	RunoffA3	Cumulative runoff for today and the 3 days before
	RunoffA4	Cumulative runoff for today and the 4 days before
	RunoffA5	Cumulative runoff for today and the 5 days before
	DrainageN1	Yesterday's measured drained water
	DrainageN2	Drained water measured 2 days ago
	DrainageN3	Drained water measured 3 days ago
	DrainageN4	Drained water measured 4 days ago
	DrainageN5	Drained water measured 5 days ago
Soil properties	Slope	Slope of the observed field
Meteorological data	Temp	Today's temperature
	TempN1	Yesterday's temperature
	TempN2	Temperature measured 2 days ago
	TempN3	Temperature measured 3 days ago
	TempN4	Temperature measured 4 days ago
	TempN5	Temperature measured 5 days ago
	TempA1	Average temperature for today and yesterday
	TempA2	Average temperature for today and the 2 days before
	TempA3	Average temperature for today and the 3 days before

Table 9 *(continued)*: *Features of the constructed dataset*.

	TempA4	Average temperature for today and the 4 days before
	TempA5	Average temperature for today and the 5 days before
	Evp	Today's evapotranspiration
	EvpN1	Yesterday's evapotranspiration
	EvpN2	Evapotranspiration measured 2 days ago
	EvpN3	Evapotranspiration measured 3 days ago
	EvpN4	Evapotranspiration measured 4 days ago
	EvpN5	Evapotranspiration measured 5 days ago
	EvpA1	Cumulative evapotranspiration for today and yesterday
	EvpA2	Cumulative evapotranspiration for today and the 2 days before
	EvpA3	Cumulative evapotranspiration for today and the 3 days before
	EvpA4	Cumulative evapotranspiration for today and the 4 days before
	EvpA5	Cumulative evapotranspiration for today and the 5 days before
	Rainfall	Today's rainfall
	RainfallN1	Yesterday's rainfall
	RainfallN2	Rainfall measured 2 days ago
	RainfallN3	Rainfall measured 3 days ago
	RainfallN4	Rainfall measured 4 days ago
	RainfallN5	Rainfall measured 5 days ago
	RainfallA1	Cumulative rainfall for today and yesterday
	RainfallA2	Cumulative rainfall for today and the 2 days before
	RainfallA3	Cumulative rainfall for today and the 3 days before
	RainfallA4	Cumulative rainfall for today and the 4 days before
	RainfallA5	Cumulative rainfall for today and the 5 days before
Target variable	**Drainage**	Today's amount of drained water, to be predicted

5 Methodology

Machine learning is one of the most active research areas in the field of artificial intelligence. It studies computer programs that automatically improve with experience (Mitchell, 1997). It also has numerous applications in the field of environmental and agricultural sciences (Debeljak and Džeroski, 2011).

In general, there are two types of learning: inductive and deductive learning (Machine learning, 2012). Inductive machine learning methods extract knowledge and patterns out of massive data sets. Deductive learning explains the rules and patterns and gives characteristic information about the data. Furthermore, inductive learning can be supervised or unsupervised, depending on the pattern that is presented as an outcome of the learning process.

Supervised inductive learning is the machine learning approach to learn a model from a set of data. It is also referred to as predictive modeling. The main assumption is that the future can be predicted only if the past or history is considered. The history is described with examples or instances which represent rows in a dataset. Each example is characterized with a set of features (attributes) that represent columns in a dataset. The features create the feature space, i.e. the space of independent properties of a given problem. Supervised learning assumes that each learning example includes some dependent property (attribute), and the goal is to learn a model that accurately predicts this property. Unsupervised learning aims to find hidden knowledge among data and examples, such as clusters, independently from a target attribute.

The data are usually given as a set of examples. An example represents one observation, object or measurement. Each example is described with a set of values of the attributes. The attributes can be continuous or discrete if they have numeric or nominal values, respectively.

A dependent variable is a property of interest that is associated with each example. Therefore, the typical machine learning task is to learn a model from a training dataset with the aim of predicting the value of the dependent variable for unseen examples.

In this study, the machine learning task is to learn a model that will be able to accurately predict the drained water flow from a field. The Machine Learning (ML) methodology includes regression and model trees, ensemble methods and learning polynomial equations. The data analysis is done by using these predictive modeling approaches implemented in the Waikato Environment for Knowledge Analysis (WEKA) (Witten and Frank, 2005), the Predictive Clustering System (CLUS) (Struyf and Džeroski, 2005), and the CIPER system for learning polynomial equations (CIPER) (Todorovski *et al*., 2004).

5.1 Decision trees

Decision tree learning is one of the most widely used and practical methods for inductive learning. As predictive model it uses decision trees, which map observations about an item to conclusions about the item's dependent value. Namely, it is a method for approximating target variables with discrete values and represents the learned function in form of a decision tree (Mitchell, 1997).

A decision tree is a classifier expressed as a recursive partition of the instance space. The decision tree consists of edges and nodes that form a rooted tree. This is a directed tree, with a node called a "root" that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is referred to as an "internal" or "test" node. All other nodes are called "leaves" (also known as "terminal" or "decision" nodes).A node is labeled by an attribute name and an arc by a valid value of the attribute associated with the node from which the arc originates. Each leaf is labeled by a class (value of the target attribute). In the decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attribute values. In the simplest and most frequent case, each test considers a single attribute, such that the instance space is partitioned according to the attributes value. In the case of numeric attributes, the condition refers to a range.

Decision tree induction algorithms (Table 10) are algorithms that automatically construct a decision tree from a given dataset. Typically the goal is to find the optimal decision tree by minimizing the generalization error. However, other target functions can be also defined, for instance, minimizing the number of nodes or minimizing the average depth of the tree.

Table 10: *Algorithm for Decision Trees.* Top-Down Algorithmic Framework for Decision Trees Induction (Rokach and Maimon, 2008)

TreeGrowing (*S,A,y,SplitCriterion,StoppingCriterion*)
Where:
S - Training Set
A - Input Feature Set
y - Target Feature
SplitCriterion - the method for evaluating a certain split
StoppingCriterion - the criteria to stop the growing process
Create a new tree T with a single root node.
    IF *StoppingCriterion(S)* THEN
        Mark *T* as a leaf with the most common value of *y* in *S* as a label.
    ELSE
        $\forall a_i \in A$ find *a* that finds the best *SplitCriterion*(a_i, *S*).
        Label *t* with *a*
        FOR each outcome v_i of *a*:
            Set $Subtree_i$= TreeGrowing ($\sigma_a=v_i$ *S, A, y*).
            Connect the root node of t_T to $Subtree_i$
            with an edge that is labelled with v_i
        END FOR
    END IF
RETURN TreePruning (*S,T ,y*)
TreePruning (*S,T ,y*)
Where:
S - Training Set
y - Target Feature
T - The tree to be pruned
DO
    Select a node *t* in *T* such that pruning it
    maximally improves some evaluation criteria
    IF *t* != Ø THEN *T* = *pruned(T, t)*
UNTIL *t* = Ø
RETURN *T*

Most algorithms that have been developed for learning decision trees are variants of a core algorithm that employs a top-down, greedy search thought the space of possible decision trees. This approach is exemplified by the ID3 algorithm (Quinlan, 1986) and its successor C4.5 (Quinlan, 1993). A simplified version of the top-down algorithm for learning decision trees by using growing and pruning is presented in Table 10.

5.2 Classification, Regression and Model trees

The problem of predictive learning can be viewed as a problem of finding a function that maps each point from the input/instance space to a point in output/target space. The construction of the function that will map the input values to output values requires a history in the form of example pairs of input/output values.

The tasks of classification and regression are the two most commonly addressed tasks in machine learning. They are concerned with predicting the value of the dependent variable (class or target) based on the values of explanatory variables (attributes). If the target is continuous, the task is called regression. If the target is discrete-valued, the task is called classification.

In both cases, a set of data is taken as input, and a predictive model is learned. The model can be in the form of decision trees that are referred to as classification and regression trees, if the problem solved is a classification or regression problem, respectively. Furthermore, the problem of regression can be addressed with linear regression model or with regression trees, the leaves of which can also be in the form of linear regression models. This type of a regression tree is called a model tree.

Model trees have leaves with linear regression functions. The regression models in the leaves represent a linear dependence between the descriptive variables and the target variable. Unlike regression trees, where the prediction of a leaf is a real value, model trees require an additional step to give result in real value predictions. This complexity of model trees is one of the disadvantages, and the reason to use regression tree instead of model trees. On the other hand, model trees have an advantage over regression trees in terms of predictive accuracy. Furthermore, the model trees are able to make predictions outside the range of the target attribute in the training examples, which is not the case with regression trees.

Several methods have been proposed for the construction of regression and model trees. Some of them have been implemented in some well-known tree induction systems such as M5 (Quinlan, 1992), RETIS (Karalic, 1992) and M5' (Wang and Witten, 1997). Regarding the problem of classification, the commonly used decision tree approach is the C4.5 algorithm (Quinlan, 1993). In this study, we use the J48 java implementation of the C4.5 algorithm provided within the WEKA (Witten and Frank, 2005) data mining suite. Furthermore, we also used the M5' regression and model trees implemented in the WEKA data mining suite (Witten and Frank, 2005), and the CLUS system (Struyf and Džeroski, 2005).

5.3 Ensembles

Ensemble methods are machine learning methods that construct a set of predictive models and combine their outputs into a single prediction. In the literature they are also referred to as multiple classifier systems, committees of classifiers, classifier fusion, combination or aggregation (Džeroski et al. 2008). The main idea is to follow the behavior of wise people when making critical decisions. Namely, they usually take into account the opinions of several experts rather than relying on their own judgment or that of a single trusted advisor. The same principle is followed by ensemble methods: learning the entire

set of models and then combining their predictions. This approach is computationally more expensive than learning just one simple model, but the predictions are usually more accurate.

There are several reasons why ensemble methods are useful and have more prominent predictions. From a statistical point of view, when we learn a model on the learning data it can result in, more or less, good predictive performance on the learning data. However, even if this performance is good, this does not guarantee good performance on unseen data. Therefore, when learning single models, we can easily end up with a bad model (although there are evaluation techniques that minimize this risk). By building ensembles, taking into account several models, and averaging their predictions, we can reduce the risk of selecting a very bad model.

Another reason why ensemble methods are useful is the lack of data. Very often, we are facing the need of learning a model from a very small dataset. As a result, the learned model can be unstable, i.e. can drastically change if we add or remove just one or two examples. A possible remedy to this problem is to draw several overlapping subsamples from the original data, learn one model for each subsample, and then combine their outputs.

In addition, complex data sources could be a problem, as well. Namely, in some cases, we have data sets from different sources where the same types of objects are described in terms of different attributes. It can be very difficult to learn a single model with all of these attributes. However, we can train a separate model for data from each source and then combine them. In this way, we can also emphasize the importance of a given data source, if we know, for example, that it is more reliable than the others.

An ensemble method constructs a set of predictive models (also referred to as ensemble) (Dietterich, 2000). It gives a prediction for a new instance by combining the predictions of its models for that instance. The outputs from the set of models can be provided by majority voting or by averaging in the case of regression.

The learning of ensembles consists of two steps (Džeroski et al. 2008). In the first step, we have to learn the base models that make up the ensemble. In the second step, we have to figure out how to combine these models (or their predictions) into a single prediction.

The base models need to be diverse, i.e. make errors on different learning examples. Combining identical or very similar models clearly does not improve the predictive accuracy of the base models. Moreover, it only increases the computational cost of the final model. By learning diverse models, their predictions can be combined in a smart way and the resulting prediction can be more accurate.

Once the set of diverse models is generated, their outputs need to be combined so that a single prediction can be obtained from the ensemble. The combination of the outputs can be performed in two different ways, by selection or by fusion. In model selection, the performance of the outputs of each base model is evaluated. The prediction of the base model that showed best performance is taken as the prediction of the ensemble. On the other hand, with model fusion, a real combination of the outputs of base models is considered.

The most commonly used technique for combining model outputs is voting. Namely, voting combines the predictions of the base models according to a static voting scheme, which does not depend on the learning data or on the base models. It corresponds to taking a linear combination of the models. The simplest type of voting is the plurality vote (also called majority vote), where each base model casts a single vote for its prediction. The output that collects most votes is the final prediction of the ensemble.

Several machine learning techniques have been developed to learn an ensemble of models and use them in combination. Most prominent among these are the schemes called

bagging (Breiman, 1996), boosting (Freund and Schapire, 1996), and random forests (Breiman, 2001).

Random forests (Breiman, 2001) is an ensemble method which generates a set of trees. The diversity between generated base models is obtained from two main sources: by using the bagging scheme and by changing the set of attributes during the learning process (Džeroski *et al*. 2008). Furthermore, during the stage of combining base models, it uses an aggregation scheme, i.e. majority voting and averaging for classification and regression, respectively.

During the learning process, at each node in the tree, a random subset of the input attributes is considered and the best split is selected from this subset. Namely, instead of considering all possible splits, the random forest method searches for an optimal attribute split within a small subset of randomly selected attributes. The best one is chosen from this subset.

Random forests also provide an estimate of which variables are important, at the same time generating an internal unbiased estimation of the generalization error as the forest building progresses. It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing. The generated forest can be saved for future use on other data and offers an experimental method for detecting variable interactions.

Random forests are a robust and typically very accurate ensemble method applicable to classification and regression problems. Furthermore, this method can be used for ranking the attributes in an attribute set. However, random forests can also suffer from the task of learning from an imbalanced training data set. As it is constructed to minimize the overall error rate, it will tend to focus more on the prediction accuracy for the majority class, which often results in poor accuracy for the minority class.

5.4 Polynomial equations

Polynomial equations are simple models that can be highly accurate on standard regression tasks. Despite the fact that piecewise regression models prevail over simple once, the simple polynomial equations can be induced efficiently and have competitive performance with piecewise models.

The method for inducing polynomial equation is called CIPER – Constrained Induction of Polynomial Equations for Regression (Todorovski *et al*., 2004). The method performs heuristic search through the set of candidate polynomial equations to find the one that has an optimal value of the heuristic function. The heuristic function is based on minimal description length principle and combines the degree of fit of the model to the data with model complexity. CIPER can be viewed as a kind of stepwise regression method for inducing polynomial equations (Todorovski *et al*., 2004).

The CIPER method allows constrained induction. The constraints that can be specified are related to the size of the equations. Furthermore, the right-hand side of the polynomial equation can be constrained with the maximum depth of a single term and maximum number of terms. In addition, the induced polynomial equations can be constrained with a sub-polynomial and super-polynomial. However, we do not use the constrained-learning capabilities of CIPER in our analyses.

The CIPER method is implemented in the C++ programming language.

5.5 ReliefF

In our study we used ReliefF for selection of the most prominent features among the previously defined feature set.

ReliefF (Relief-F) is a successful estimator of feature relevance. It is trying to solve the problem of estimating the quality of attributes (features), which is an important issue in the machine learning. It has commonly been viewed as feature subset selection method that is applied in a prepossessing step before the model is learned (Kira and Rendell, 1992). It is one of the most successful preprocessing algorithms (Dietterich, 1997).

ReliefF is actually a general feature estimator and has been used successfully in a variety of circumstances. It can be used to select splits in the building phase of decision tree learning (Kononenko *et al*., 1997) or as an attribute weighting method (Wettschereck *et al*., 1997). The basic idea of the ReliefF algorithm is to estimate the quality of given attributes according to how well their values distinguish between instances that are near each other. The algorithm is very robust and can deal with incomplete and noisy data.

The ReliefF algorithm randomly selects an instance and then searches for k nearest neighbors from the same class. These k nearest neighbors are called nearest hits. The algorithm then searches for k nearest neighbors from each of the other classes. These neighbors are called nearest misses. The last step in an iteration of the process is to update the quality estimation array for all attributes depending on their values for the selected instance. The quality estimation for one attribute is calculated as the difference between hits and misses during the given iteration. Furthermore, the whole process is repeated m times, where m is a user-defined parameter.

5.6 Evaluation

Given a set of data, only a part of it is typically used to learn a predictive model. This part is referred to as the training set. The remaining part is reserved for evaluating the predictive performance of the learned model and is called the testing set. The testing set is used to estimate the performance of the model on unseen data.

More reliable estimates of the performance on unseen data are obtained by using cross-validation, which partitions the entire data available into n (in this study n is set to 10) subsets of approximately equal size. Each of these subsets is in turn used as a testing set, with all of the remaining data are used as training set. The performance figures for each of the testing sets are averaged to obtain an overall estimate of the performance on unseen data.

For assessing the performance of the predictive models we used the measures of Root Mean Squared Error (RMSE), Root Relative Squared Error (RRSE) and the correlation coefficient. Namely, RMSE is a frequently-used measure of the differences between the values predicted by the model and the values actually observed from the experiments being modeled. The RRSE is an accuracy measure of the differences between the predicted and measured values relative to the standard deviation of the target variable. Finally, in statistics the correlation coefficient indicates the strength and direction of a linear relationship between two random variables, while in machine learning it indicates the linear relationship between the predicted and measured values of the target variable.

6 Experimental setup

Our task consists of building models for predicting the drained water flow from a field by using machine learning algorithms, and validating the models by using standard validation techniques. The problem requires the use of techniques that can learn an accurate model that is simple and understandable. Furthermore, the available expert knowledge gives us a possibility of model validation. Therefore, the solution proposed in the thesis generates outputs that are adjusted to local specifics (like soil properties, surface properties and agricultural practices) and validated by expert knowledge.

The general scheme for our experimental setup is shown in Figure 14. It presents all phases and steps of performing machine learning experiments that which will be described in the following sub-sections.

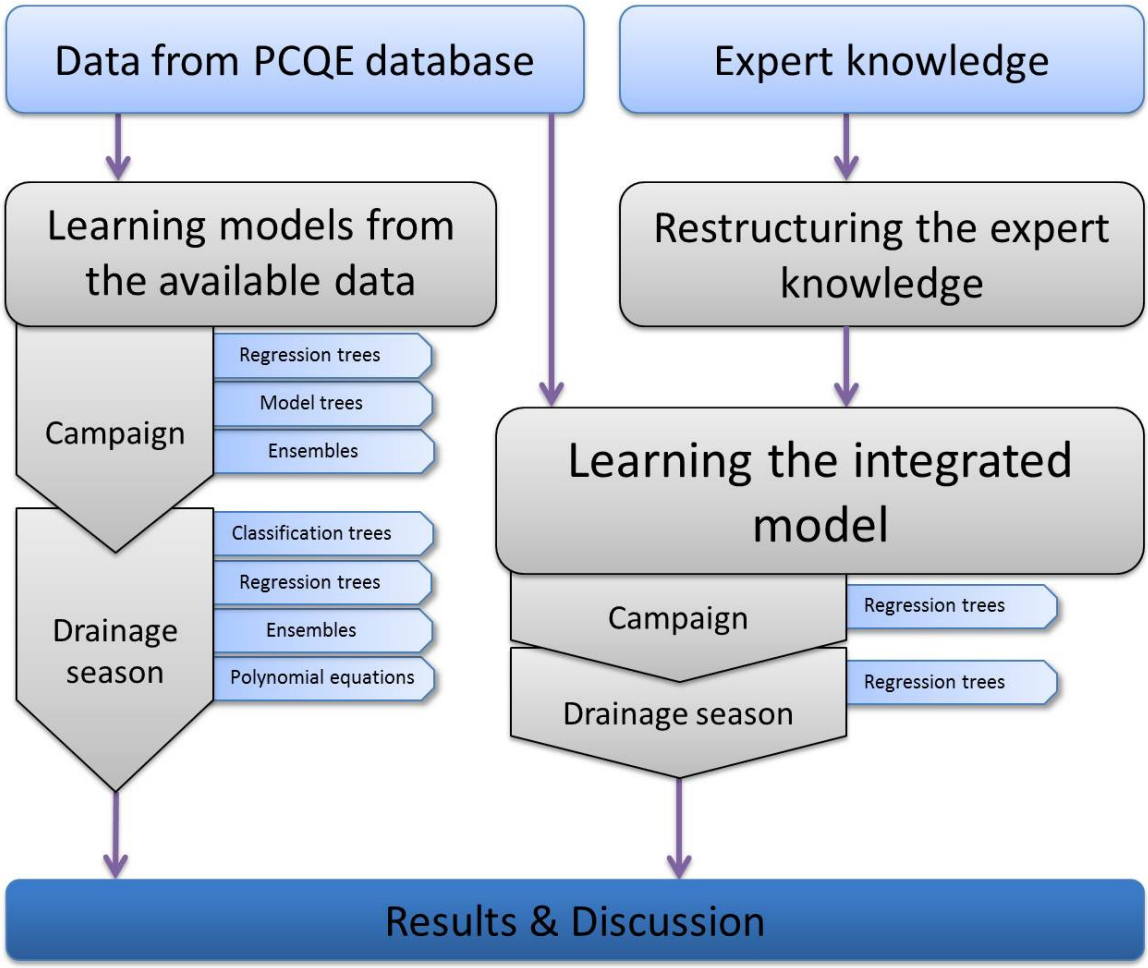


Figure 14: *Experimental setup*. The overall scheme of the machine learning experiments performed in our study.

For the purposes of integration of the available expert knowledge, in the first stage we consider its restructuring in a format suitable for further integration. Next, we present the process of feature selection and feature set creation from the available dataset and the defined features. Next, the learning of predictive models from data is explained. Finally, the integration of the expert knowledge with predictive modeling from data is described.

6.1 Restructuring the expert knowledge

The first step in the integration of the expert knowledge with the learned predictive models for water flow is the restructuring of the available expert knowledge into decision trees. The available expert knowledge is structured in the form of decision rules stored in tables. This format is inappropriate for the methodology that we are trying to develop, which is based on integration of the learned data mining models with the available expert knowledge. Therefore, we restructure the expert knowledge in the form of decision trees. We chose decision trees because they are the most appropriate type of model for realizing the integration with the CLUS tool. The integration will be described in Section 7.

First of all, the decision rules from all tables were stored in one single table. The complete size of the created dataset is shown in Table 6. Next, we used the WEKA suite to build a model that covers all the existing decision rules. The coverage space is estimated to 100 % accuracy over training dataset. This means that for each possible rule there is a path through the decision tree that gives the correct value of the target attribute, in this case the intensity of drainage water flow.

6.2 Learning predictive models

The next task was to predict the amount of drained water in a field per day. For this task, we were able to either use the whole dataset or choose a field and learn models only from the data of the selected field. Moreover, according to the experts, the prediction of drained water flow is more suitable when only the drainage seasons are considered.

A drainage season is a period of intensive drainage events that roughly lasts until the first week that is registered without a drainage event. Visually, drainage seasons cover the extremes on a chart of drainage events and can appear during winter and spring. Therefore, we have two main scenarios within this phase: (1) learning a predictive model for a campaign and (2) learning a predictive model for a drainage season.

The phase of learning a predictive model for a drainage season includes the learning of predictive models for predicting the start and the end of the winter drainage seasons. Namely, there is a simple expert rule for determining the start of the winter drainage period, which states that the winter drainage period starts when the cumulative drainage from the start of the agricultural season (September 1st – August 31st) reaches 5 mm. However, the quantity of drained water is measured in an experimental setting, and the analyses on the field are expensive so that the farmers usually do not have any measured data for the drained water. Therefore, we need to predict the start of the winter drainage period from data that are easily obtainable, such as meteorological data.

The meteorological data (rainfall, temperature, evapotranspiration) have been already introduced and were provided for each field and each day for the period 1987–2011. For the purpose of this study, we used only rainfall and temperature data. For explanatory purposes we introduced two new attributes based on the existing ones: “Rainfall_Cumul” which describes the cumulative amount of rainfall for each day since the beginning of a campaign (September 1st); and “Temp_weekly_avg” which keeps information about the average temperature in the previous 7 days.

The target variable has been constructed with two possible values: *no_drainage* – in

case when a winter drainage season has not started yet; and *start_drainage* for the days within a winter drainage season.

For the second study, prediction of the end of a winter drainage season, there are no clearly defined conditions. Therefore, we used data mining and machine learning techniques to find some regularity in the data and obtain a model for predicting the end of winter drainage season. The predictive models were learned with the J48 algorithm for induction of classification trees. The target variable has been defined as nominal with two possible values: *drainage* – which covers the period of the drainage season; and *end_drainage* which emphasizes the ending of a drainage season. The full attribute set used in the data analyses consists of meteorological data for the previous and next 14 days from the observed day:

- *"Avg_temp_past_1-7_days"* — the average temperature in the previous 7 days
- *"Avg_temp_past_8-14_days"* — the average temperature in the 7 days before last week
- *"Avg_temp_past_1-14_day"* — the average temperature in the previous 14 days
- *"Tot_rainfall_past_1-7_days"* — the cumulative rainfall in the previous 7 days
- *"Tot_rainfall_past_8-14_days"* — the cumulative rainfall in 7 days before last week
- *"Tot_rainfall_past_1-14_days"* — the cumulative rainfall in the previous 14 days
- *"Avg_temp_next_1-7_days"* — the average temperature in the following 7 days
- *"Avg_temp_next_8-14_days"* — the average temperature in the 7 days after that
- *"Avg_temp_next_1-14_day"* — the average temperature in the following 14 days
- *"Tot_rainfall_next_1-7_days"* — the cumulative rainfall in the following 7 days
- *"Tot_rainfall_next_8-14_days"* — the cumulative rainfall in the 7 days after that
- *"Tot_rainfall_next_1-14_days"* — the cumulative rainfall in the following 14 days

For the task of learning predictive model for a campaign, we used machine learning and data mining methods for learning regression trees, model trees and ensembles implemented in the CLUS system (Blockeel, 1998). For the second scenario we considered regression trees, ensembles implemented in the CLUS system and polynomial equations (CIPER (Todorovski, 2004)) as prominent patterns for knowledge representation. Moreover, before we attempted to learn the predictive model for predicting water flow as part of the second scenario, the model for estimating the start and end of a winter drainage season has been learned.

In the previous section we have described the feature (attribute) set which consists of about 85 attributes. This is a massive feature set, which needs to be reduced and only the most important features (attributes) need to be selected. For the task of attribute selection,

we consider a few scenarios which include feature ranking and a wide range of analysis with different combination of the attributes. The results are presented in Section 7.

The next sub-task covers the selection of the most suitable fields used for learning the predictive model on predicting drainage water flow in both scenarios: in a campaign and in a drainage season. We learned models on data from one field, because the data are with lower variance and sometimes give better results compared with models learned on data from a whole region. As mentioned before, the experimental site La Jaillière has 11 fields. Some of the fields are drained, but there are also two of them (Fields: T1 and T2) which do not have a drainage system (Figure 6), and later are excluded from the "field selection" task. The final selection concludes that fields T3 and T6 are the most suitable fields for learning a predictive model.

Finally, we created 3 sub-scenarios for learning predictive models for predicting drainage water flow. First, we took all the available data including soil properties data and built a predictive model. Second, a model has been learned from data for field T3. Finally, we used data for field T6 (Figure 3) to build a predictive model.

Estimations of performances on unseen data are obtained by using cross-validation. We use 10-fold cross-validation, which partitions the entire data available into 10 subsets of roughly equal size. The results are presented and discussed in the next section.

6.3 Learning a predictive integrated model

The last step in building a model that will achieve better performance on the task of prediction of drainage water flow from a field is to integrate the already learned data mining models with the available restructured expert knowledge. Namely, the integrated predictive model is learned in two stages: (1) support the learning predictive integrated model for a campaign and (2) support the learning predictive integrated model for a drainage season. The integration of these two concepts can be described as the usage of an available expert knowledge to define the most important attribute(s) and then learning the rest of the model from data. From the data available in the PCQE database, the only attribute that can be considered as the most important in the new data mining model is "Season". The available expert knowledge recognizes the attribute "Season" as the most important and the most prominent in the root of the new model. In the second stage, instead of "Season" we used the "Drainage Season" attribute which defines the existence of drainage season on the reviewed example (day) and the type of drainage season: winter or spring drainage season. Therefore, we "supervised" the process of learning a model from the data by defining a constraint which is in the form of a partial decision tree. The rest of the decision tree has been learned by the algorithm for learning regression trees implemented within the CLUS tool.

In the next section, we will discuss and compare the results from all the defined stages in the experimental design and the learned predictive models that predict the daily amount of drained water or intensity of drainage water flow from a field.

7 Results

We are interested in predicting the target variable which represents the amount of drained water from a field. To achieve this goal, we proceed as follows. We first select the attributes and the fields that will be used for further analyses (Section 7.1). We then build predictive models from data only, either from entire campaigns or from drainage seasons only (Section 7.2). We next restructure the expert knowledge provided by the domain expert and test its predictive power (Section 7.3). Finally, we build integrated models that take as input both the data and parts of the expert knowledge (Section 7.4).

Hereafter, we present the results from the data analyses and discuss how they support our hypothesis. We present the results through several dimensions of evaluation: estimating the best combination of explanatory attributes that explain the drainage water flow in the most accurate way; estimating the quality of the prediction of the target variable by applying several data mining and machine learning methods; and visual comparison of the predictions from the learned predictive models and the measured data.

Furthermore, we present a set of tables of the obtained results from the data analyses using three types of accuracy and performance measurements: RMSE (Root Mean Squared Error) which presents the square root of the mean squared difference between the predicted and actual values; RRSE is the RMSE relative to the standard deviation of the target variable; and the correlation coefficient, which measures the correlation (linear dependence) between the predicted and actual values.

Bellow, we give a brief discussion for each of the learned models regarding their accuracy and possible usage, while a general discussion of the results achieved by this thesis is given in Section 8.

7.1 Attributes and fields selection

At the beginning, we need to find the best combination of explanatory attributes (that explain the drained water flow in the most accurate way) and fields from the experimental site. We first define attribute set by attribute/feature selection. We then choose the most prominent fields for learning predictive models on reduced data, i.e. data from only one field.

For the task of attribute selection, we used several scenarios which include feature ranking and a wide range of exploratory analysis with different combinations of the attributes. For feature ranking, we used the ReliefF algorithm from the WEKA data mining suite. The results of the preliminary analyses of the data from each individual field are shown in Table 12.

Table 11: *The relevant features*. The list of most relevant attributes chosen by feature selection.

Attribute	Description
Day	The consecutive day in the campaign, 1st of September being the first day.
Season	Autumn-Winter, Spring or Summer
DrainageSeason	Existence and type of drainage season (possible values: ND – No drainage season, WD – Winter drainage season, and SD – Spring drainage season)
Crop	Crop that is on the field at the time of observation.
CDCoef	Crop development coefficient
Slope	Slope of the observed field
RainfallA1	Cumulative rainfall for today and yesterday
Temp	Today's average temperature
Runoff	Today's measured runoff
DrainageN1	Yesterday's measured drained water

The next sub-task covers the selection of the most suitable fields to be used for learning the predictive models for predicting drainage water flow in both scenarios: in a campaign and in drainage season. As mentioned before, the experimental site La Jaillière has 11 fields. Some of the fields are drained, but there are also two of them (fields T1 & T2) which do not have a drainage system (Figure 6), so they are excluded from the "field selection" task.

Furthermore, we analyzed each field by learning models from data of the selected field. It is worth mentioning here that the preliminary analyses were performed without soil properties data, unlike the final learning models. The results were compared and two the most two prominent fields were selected for further model learning (Table 12). The fields T3 and T6 were best suited for learning predictive models for drained water flow.

Table 12: *Preliminary analyses of individual fields*. Results of the preliminary analyses performed for each field, separately. The fields T3 and T6 will be considered in further analyses.

Field	Regression/Model tree	Correlation	RRSE
T3	Regression tree	**0.8124**	**58.37 %**
T3	Model tree	**0.8926**	**45.09 %**
T4	Regression tree	0.7926	61.02 %
T4	Model tree	0.8698	49.36 %
T5	Regression tree	0.6985	71.56 %
T5	Model tree	0.7519	65.95 %
T6	Regression tree	**0.8383**	**54.66 %**
T6	Model tree	**0.8978**	**44.08 %**
T7	Regression tree	0.7917	61.48 %
T7	Model tree	0.8555	51.80 %
T8	Regression tree	0.711	70.31 %
T8	Model tree	0.811	58.54 %
T9	Regression tree	0.7901	61.34 %
T9	Model tree	0.845	53.51 %
T10	Regression tree	0.7913	61.43 %
T10	Model tree	0.8571	51.51 %
T11	Regression tree	0.8147	58.13 %
T11	Model tree	0.8758	48.31 %

7.2 Evaluation of the predictive models

As described before, in the phase of learning predictive models from data, we learn predictive models based on data for whole campaign. We also learn predictive models that predict the amount of drained water within drainage seasons. The next two sub-sections discuss the predictive performance for each of these two types of models.

7.2.1 Campaign based predictive models

The models have been learned within two different environments (data mining suites): WEKA, and CLUS. Three machine learning algorithms were used for building predictive models to predict the amount of drained water from a field: regression trees, model trees, and ensembles (random forests). The ensembles from the random forest algorithm implemented in the CLUS system were learned in 10 iterations, i.e. included 10 regression trees. The decision trees (including regression trees and random forests) were constrained to a maximal depth of 4 levels.

The attribute set used in this phase is the same as defined in the previous sub-section (Table 11). The performance of the induced models on unseen data was estimated by 10-fold cross validation. The results from the data analyses are presented in Table 13 and Table 14 with performance measured by training data and by 10-fold cross validation, respectively. The performance measure: RMSE, RRSE and correlation coefficient (r) are presented. The models were built from data of all fields, field T3, and field T6.

Table 13: *The accuracy of campaign based models on the training data*. *Std. Dev.* stands for standard deviation of the target variable on the training data set.

Fields	Model	Std. Dev.	RMSE	RRSE	Corr. coeff. (r)
All	Regression tree	1.918	1.0457	54.51 %	0.8384
All	Model tree	1.918	0.9547	46.76 %	0.8674
All	Random forest	1.918	1.0860	56.61 %	0.8596
T3	Regression tree	1.976	0.9765	49.42 %	0.8694
T3	Model tree	1.976	1.0313	52.19 %	0.8530
T3	Random forest	1.976	0.9869	49.95 %	0.8734
T6	Regression tree	2.279	0.9359	41.07 %	0.9118
T6	Model tree	2.279	0.9510	41.73 %	0.9088
T6	Random forest	2.279	0.9633	42.28 %	0.9112

Table 14: *The predictive performance of campaign based models estimated on unseen data by 10-fold cross validation*. *Std. Dev.* stands for standard deviation of the target variable on the training data set.

Fields	Model	Std. Dev.	RMSE	RRSE	Corr. coeff. (r)
All	Regression tree	1.918	1.0792	56.25 %	0.8268
All	Model tree	1.918	0.9700	50.55 %	0.8628
All	Random forest	1.918	1.1094	57.83 %	0.8440
T3	Regression tree	1.976	1.1193	56.64 %	0.8256
T3	Model tree	1.976	1.0347	52.35 %	0.8520
T3	Random forest	1.976	1.0907	55.20 %	0.8491
T6	Regression tree	2.279	1.0688	46.91 %	0.8832
T6	Model tree	2.279	1.0194	44.73 %	0.8945
T6	Random forest	2.279	1.2004	52.68 %	0.8610

The results showed that field T6 was the most prominent for learning predictive models for drained water predictions. The algorithm M5P for building model trees, implemented in WEKA suite, gave the best results. The most accurately learned model (Model tree for field T6) is shown in Figure 15.

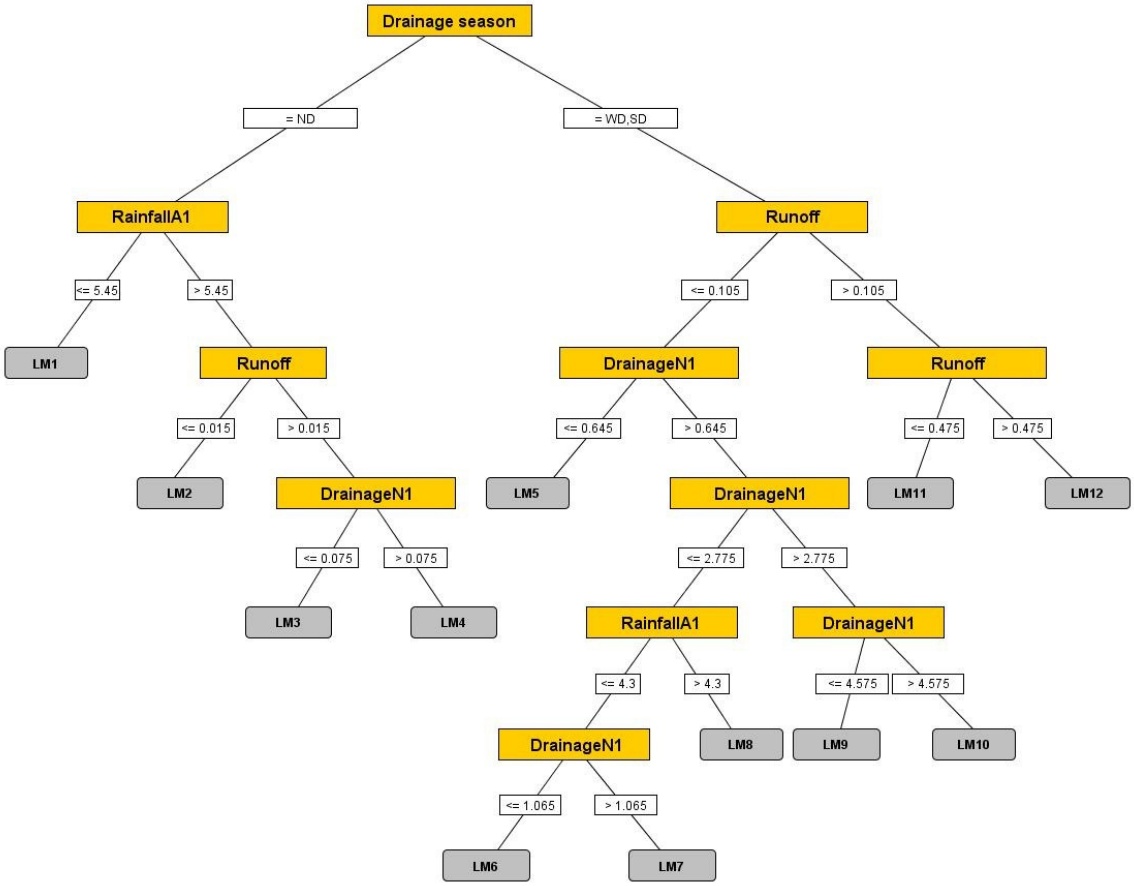


Figure 15: *Model tree*. The model tree for predicting drained water flow learned from whole campaign data of field T6.

In the learned model (Figure 15), the attribute "Drainage season" appears as the most important attribute at the root of the model tree. Paramount importance of this attribute has been confirmed by the domain experts. Furthermore, "Runoff" is considered as the next most important attribute (in the case of winter or spring drainage season). This is due to the fact that the soil is very often saturated during these drainage seasons, when extreme drainage events and rainfalls are registered. Otherwise, out of the drainage seasons, the drainage water flow is highly dependent on the cumulative amount of rainfall in the last two days. This is a logical dependence under no drainage season's circumstances, when the soil is not saturated.

The complexity of the model is acceptable because the model is easily understandable. Namely, the decision tree has 12 leaves and 6 levels in depth. At each leaf of the model, a linear regression model is included. The full list of linear regression models is presented in Table 15, followed by a visualization of the predicted values (Figure 16). Moreover, visualizations of the predictions from the other learned models (regression trees and random forests) are shown in Figure 17 & 18.

Table 15: *Set of linear regression models*. The given set of linear regression models contains the models located in the leaves of the model tree shown above (Figure 15).

Leaf	Model
LM1	*Drainage* = 0.0001 * *CDCoef* + 0.0001 * *RainfallA1* + 0.0017 * *DrainageSeason*=SD,WD + 0.0255 * *Runoff* + 0.0021 * *DrainageN1* + 0.0134
LM2	*Drainage* = 0.0021 * *CDCoef* + 0.0001 * *RainfallA1* + 0.0017 * *DrainageSeason*=SD,WD + 0.154 * *Runoff* + 0.0139 * *DrainageN1* + 0.0198
LM3	*Drainage* = 0.0267 * *CDCoef* + 0 * *Temp* + 0.0034 * *RainfallA1* + 0.0017 * *DrainageSeason*=SD,WD + 2.0575 * *Runoff* + 3.4986 * *DrainageN1* – 0.1439
LM4	*Drainage* = 0.0628 * *CDCoef* + 0.0017 * *DrainageSeason*=SD,WD + 8.4902 * *Runoff* + 0.116 * *DrainageN1* – 0.1567
LM5	*Drainage* = 0.0458 * *Crop*=Winter_peas,Spring_peas, Winter_horse_bean, Rapeseed, Rgi, Barley(spring), Wheat, CIPAN + 0.0058 * *CDCoef* + 0.0001 * *Temp* + 0.0105 * *RainfallA1* + 0.0034 * *DrainageSeason*=SD,WD + 4.464 * *Runoff* + 0.8678 * *DrainageN1* – 0.0399
LM6	*Drainage* = – 0.0001 * *Day* – 0.0038 * *Crop*=Winter_peas, Spring_peas,Winter_horse_bean, Rapeseed, Rgi,Barley(spring),Wheat,CIPAN + 0.1218 * *CDCoef* + 0.0001 * *Temp* + 0.0072 * *RainfallA1* + 0.0034 * *DrainageSeason*=SD,WD

Table 15 *(continued)*: *Set of linear regression models.*

		+ 0.3544 * *Runoff* + 0.6398 * *DrainageN1* + 0.0445
LM7	*Drainage* =	– 0.0001 * *Day* – 0.0038 * *Crop*=Winter_peas, Spring_peas, Winter_horse_bean, Rapeseed, Rgi, Barley(spring), Wheat, CIPAN + 0.0277 * *CDCoef* + 0.0001 * *Temp* + 0.0537 * *RainfallA1* + 0.0034 * *DrainageSeason*=SD,WD + 2.9046 * *Runoff* + 0.4525 * *DrainageN1* + 0.3282
LM8	*Drainage* =	– 0.0001 * *Day* – 0.2997 * *Crop*=Winter_peas, Spring_peas,Winter_horse_bean, Rapeseed, Rgi, Barley(spring), Wheat, CIPAN + 0.0342 * *CDCoef* + 0.0001 * *Temp* + 0.0851 * *RainfallA1* + 0.0034 * *DrainageSeason*=SD,WD + 4.2892 * *Runoff* + 0.4291 * *DrainageN1* + 0.2593
LM9	*Drainage* =	– 0.0006 * *Day* – 0.1156 * *Crop*=Winter_peas, Spring_peas,Winter_horse_bean, Rapeseed, Rgi, Barley(spring), Wheat, CIPAN + 0.0413 * *CDCoef* + 0.0001 * *Temp* + 0.0568 * *RainfallA1* + 0.0034 * *DrainageSeason*=SD,WD + 5.8283 * *Runoff* + 0.232 * *DrainageN1* + 0.9682
LM10	*Drainage* =	– 0.0053 * *Day* – 1.5115 * *Crop*=Winter_peas, Spring_peas,Winter_horse_bean, Rapeseed, Rgi, Barley(spring), Wheat, CIPAN + 0.0413 * *CDCoef* + 0.0001 * *Temp* + 0.1745 * *RainfallA1* + 0.0034 * *DrainageSeason*=SD,WD + 0.243 * *Runoff* + 0.0349 * *DrainageN1* + 3.975

Table 15 *(continued)*: *Set of linear regression models.*

LM11	*Drainage* = – 0.0058 * *Day* + 0.5052 * *Crop*=Winter_peas, Spring_peas,Winter_horse_bean, Rapeseed, Rgi, Barley(spring), Wheat, CIPAN + 0.0707 * *RainfallA1* + 0.0034 * *DrainageSeason*=SD,WD + 8.5731 * *Runoff* + 0.163 * *DrainageN1* + 0.6737
LM12	*Drainage* = – 0.0322 * *Day* + 0.0618 * *Crop*=[Winter_peas, Spring_peas, Winter_horse_bean, Rapeseed, Rgi, Barley(spring), Wheat, CIPAN] + 0.2281 * *RainfallA1* + 0.0034 * *DrainageSeason*=SD,WD + 1.8235 * *Runoff* + 0.2364 * *DrainageN1* + 6.378

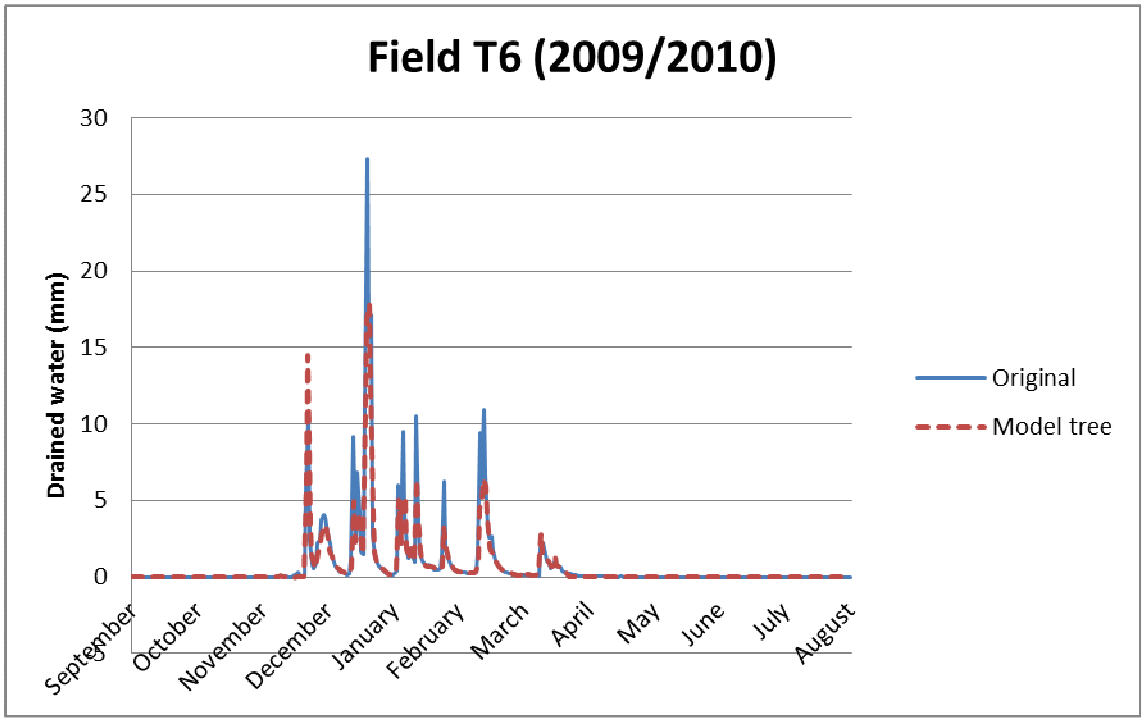
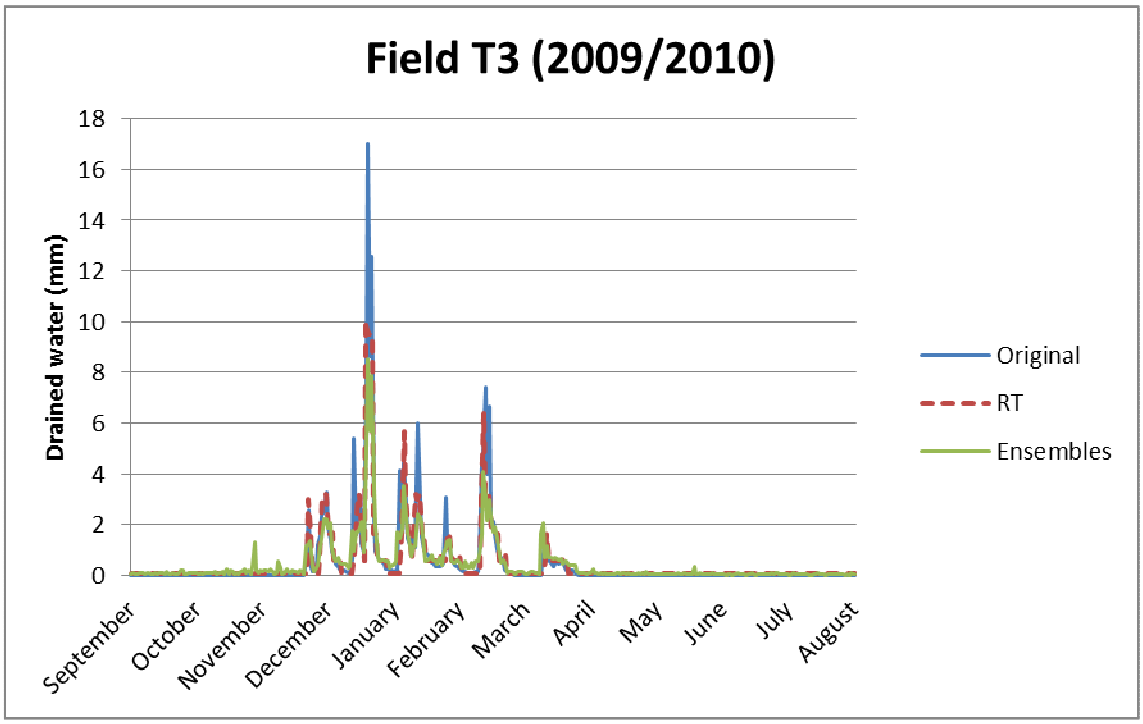
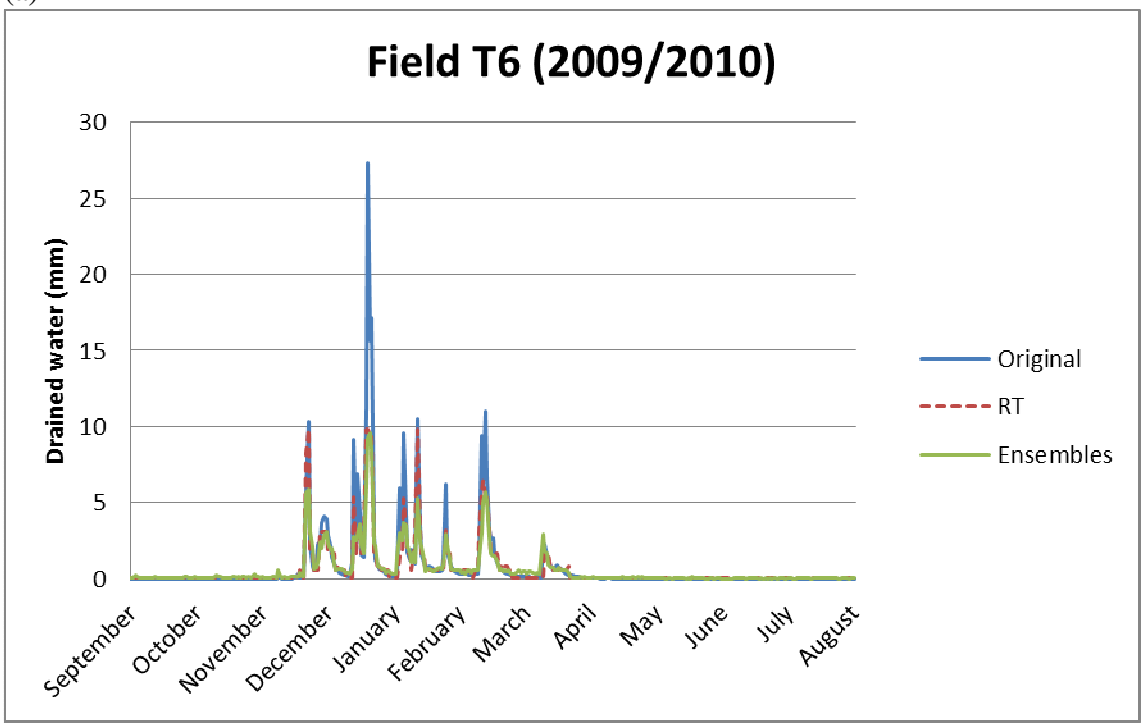


Figure 16: *Amount of drained water (mm) predicted by the model tree learned on field T6.* Visualization of the amount of drained water predicted for filed T6 by model tree learned on data from field T6, compared with the measured (original) values. The period of September 1st, 2009 – August 31st, 2010 was considered.

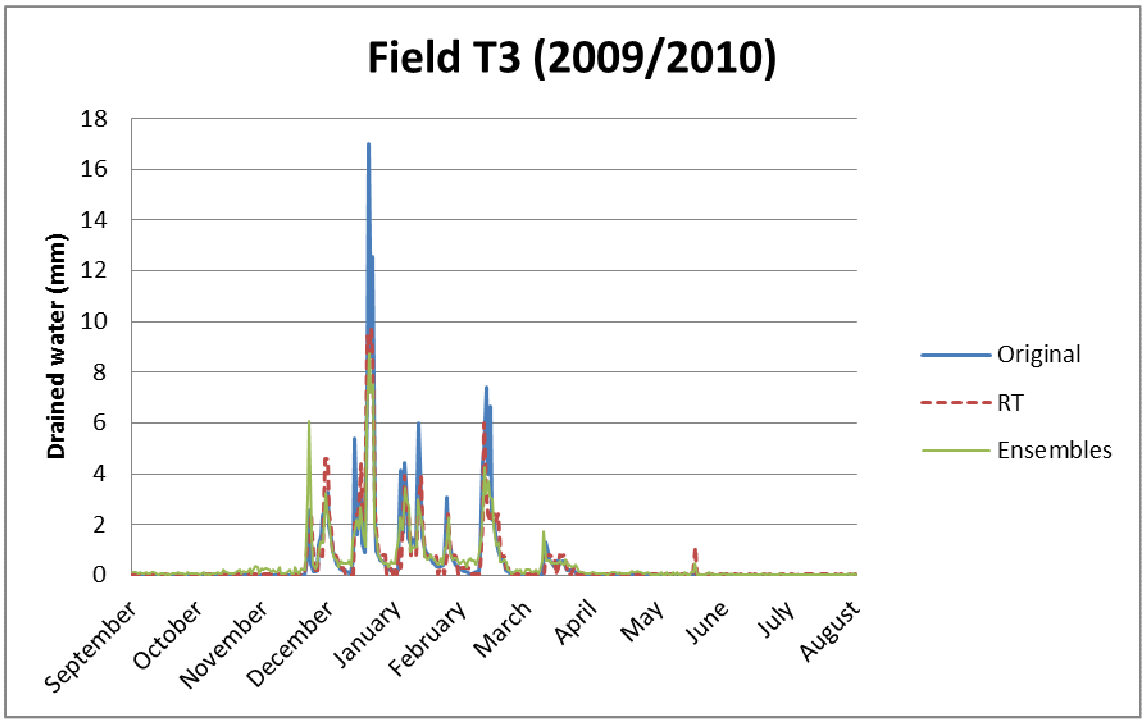


(a)

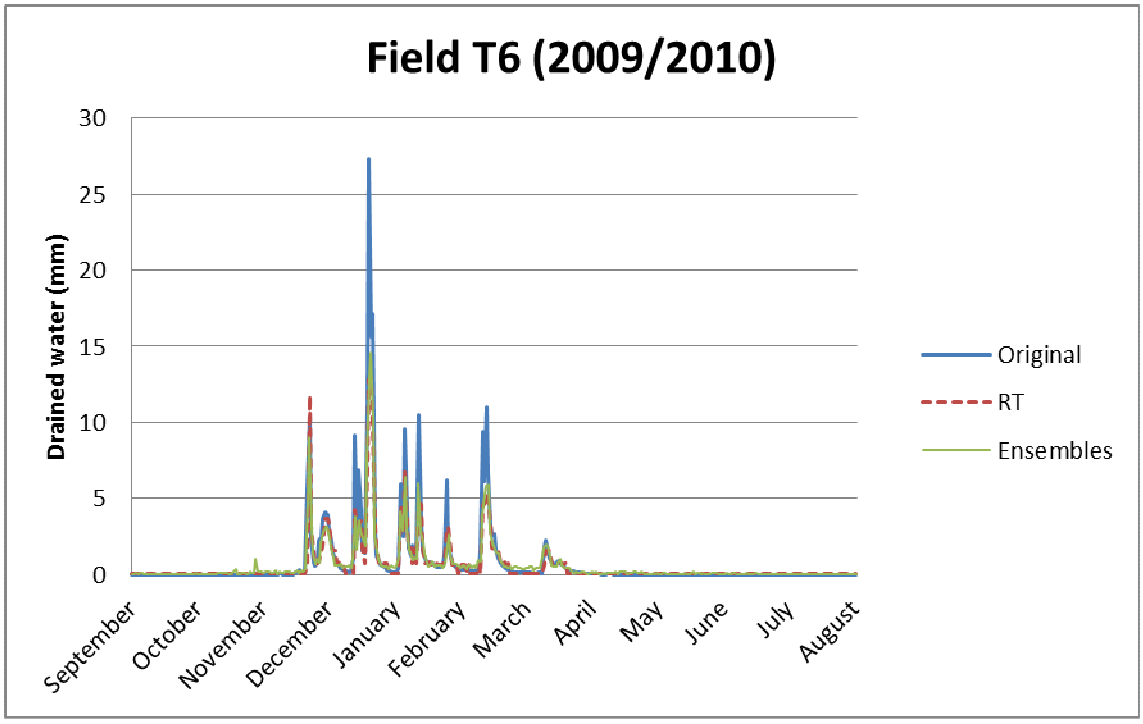


(b)

Figure 17: *Amount of drained water (mm) predicted by the regression tree and ensemble models learned on all fields*. Visualization of the amount of drained water predicted (for fields T3 (a) and T6 (b)) by RT (regression tree model) and ensembles (random forests) learned on data from all fields, compared with the measured (original) values. The period of September 1st, 2009 – August 31st, 2010 was considered.



(a)



(b)

Figure 18: *Amount of drained water (mm) predicted by the regression tree and ensemble models learned on field T3 & T6.* Visualization of the amount of drained water predicted by RT (regression tree models) and ensembles (random forests) learned on data from field T3 – given in (a) and T6 given in (b), compared with the measured (original) values: The period of September 1st, 2009 – August 31st, 2010 was considered.

The visualization of the predicted values from the learned models is in accordance with the performance measures given above (Table 13). Namely, all models have difficulties when it comes to the prediction of the amount of drained water during the winter, when extreme drainage events are registered. It is the case when the amount of

drained water is higher than usual. Otherwise, predictions successfully follow the trend of the measured (original) values.

Although the learned models do not cover the extremes well, they still can be used and eventually improved with additional knowledge. Hence, we used the available expert knowledge in order to improve this predictive performance. The results are presented in the Section 7.3.

7.2.2 Drainage season based predictive models

As mentioned before, two different tasks have been considered for drainage season based predictive models. For the first task we used classification trees in order to predict the start and the end of the winter drainage season. Furthermore, these models explain the conditions under which a day can be considered as a part of the drainage season.

For the second task we explored different machine learning techniques for prediction of the amount of drained water flow from a field during the drainage seasons only. With this, we are trying to get more accurate predictions of the amount of drained water during the most critical period within a campaign. Previously, we have seen that the learned models based on the data from the whole campaign do not have ability to closely predict the amount of drained water.

7.2.2.1 Predicting the start and the end of the winter drainage season

Here we address the task of predicting the days when a winter drainage period is beginning and ending. The predictive models that we build are based on easily accessible meteorological data, including rainfall and temperature.

For the task of predicting the beginning of the winter drainage season, a classification tree model has been obtained. The model has been learned with the J48 algorithm for induction of classification trees, implemented in the WEKA data mining suite. The model itself is presented in Figure 19.

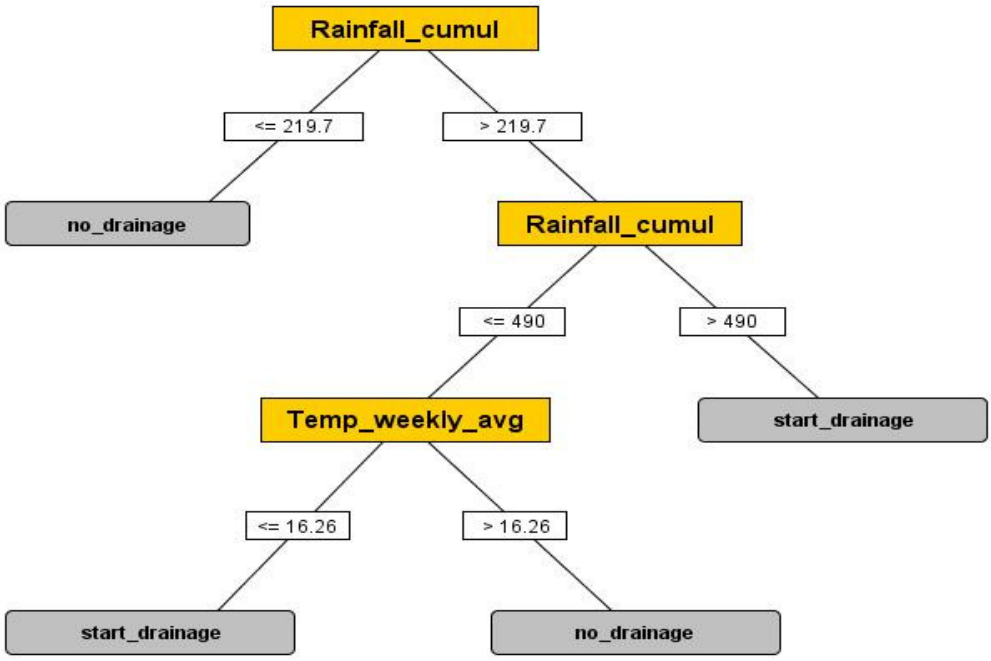


Figure 19: *Classification tree*. The predictive model for predicting the start of the winter drainage season

The obtained predictive model is simple, with a total of 5 leaves. Moreover, the predictive accuracy of the learned model on unseen data is 93.5 %, estimated by 10-fold cross validation. Although the domain experts cannot clearly define the rules for the beginning of a winter drainage season, the accuracy of the learned model shows that the beginning of the winter drainage season can be accurately predicted from easily accessible data.

On the other hand, the task of predicting the end of a winter drainage season requires more attention, especially because the defining conditions are not clearly defined. Therefore, with machine learning techniques, we explored the possible conditions which need to be fulfilled in order to determine the end of a winter drainage season. Different scenarios were taken into account by constructing different attribute sets for learning the predictive model for the end of a winter drainage season. Here, we present the two most accurate models.

The first predictive model (Figure 20) uses only meteorological data from the previous days. The second one (Figure 21) uses meteorological data both from the previous days and the following 7 days. The historical data for the following 7 days are given in the data set (Appendix A, Table A.1) that contains the dates for the start and end of the drainage seasons. On the other hand, in real situations, the weather forecast for the next 7 days should be considered.

Table 16: *The accuracy of the models for predicting the end of a drainage season*. The accuracy of learned predictive models is estimated by 10-fold cross validation

Model	Attributes	Accuracy
Model 1	Avg_temp_past_1-7_days Avg_temp_past_8-14_days Tot_rainfall_past_1-7_days Tot_rainfall_past_8-14_days	85.9383 %
Model 2	Avg_temp_past_1-7_days Avg_temp_past_8-14_days Avg_temp_next_1-7_days Tot_rainfall_past_1-7_days Tot_rainfall_past_8-14_days Tot_rainfall_next_1-7_days	88.0911 %

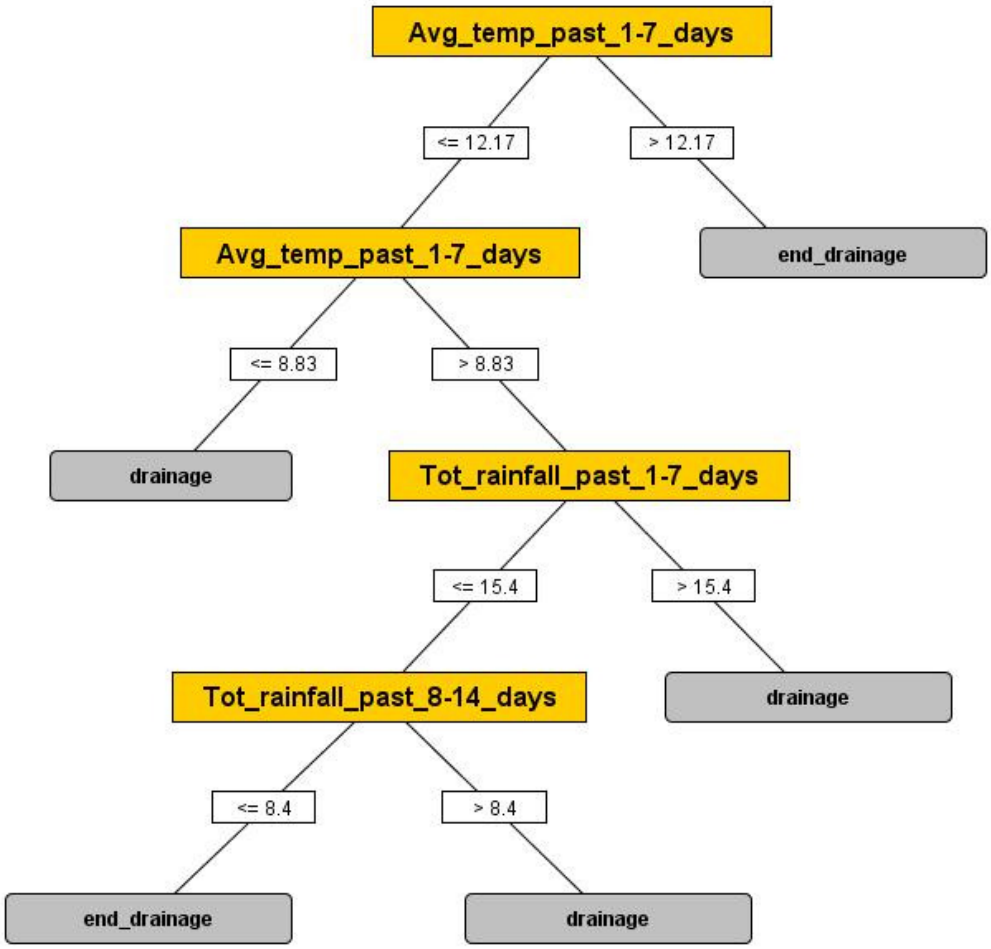


Figure 20: *Classification tree for predicting the end of the winter drainage season learned from past meteorological data*. The predictive model for predicting the end of the winter drainage season

The accuracy of the learned models is given in Table 16. The learned models are simple/small, with a size of 4 levels and 5 leaves. This is an additional advantage of these models. Namely, their accuracy and complexity makes them applicable in real situations.

The learned models show high accuracy in the prediction of the start and end of a winter drainage season. Therefore, they can be used in order to give a reliable estimation for a drainage season and extreme drainage events for the La Jaillière region. Furthermore, the data input of the learned models are based on meteorological data for the past and the near future, which can be easily obtained from the nearest meteorological station. Thus, the learned models and their predictions can be used as practical information for planning of those agricultural practices, whose application depends on the drainage seasons and possibly extreme drainage events.

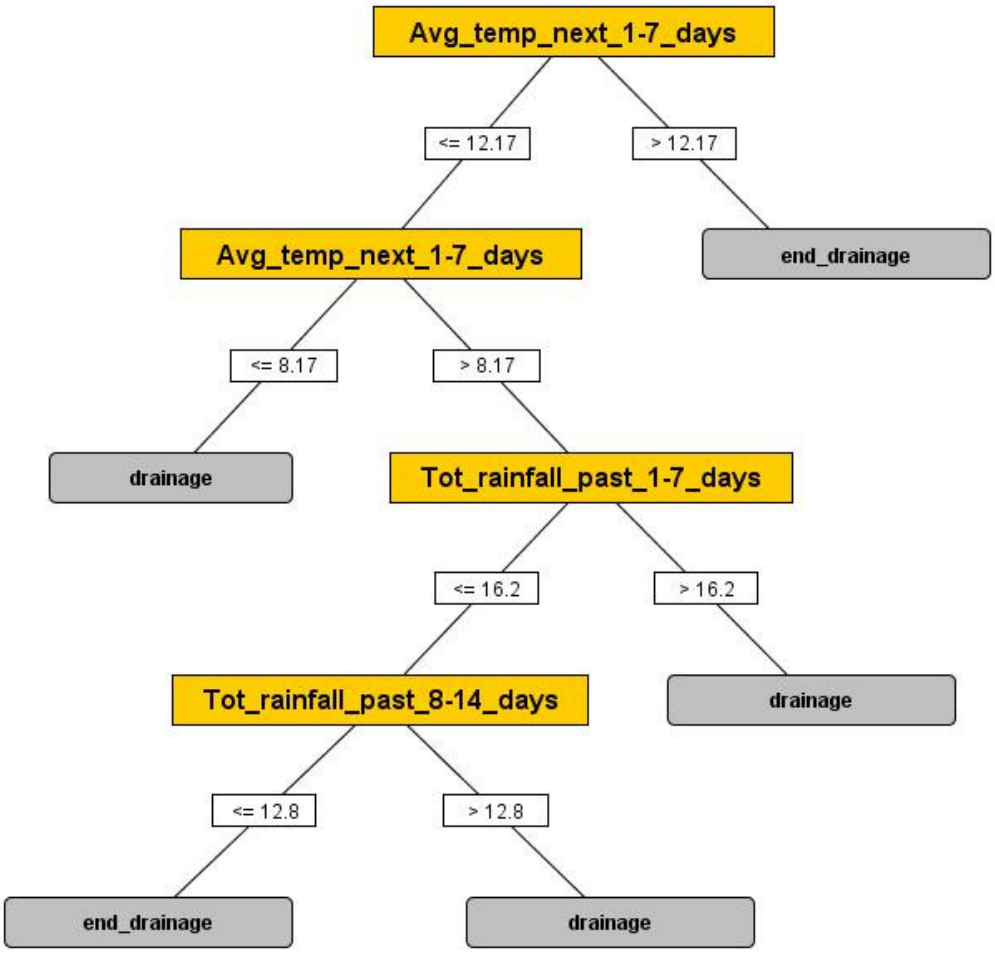


Figure 21: *Classification tree for predicting the end of the winter drainage season with combination of past and future meteorological data.*

7.2.2.2 Predicting the amount of drained water during drainage seasons

We next considered the learning of predictive models only from data during the drainage seasons (Appendix A, Table A.1), because these periods are most important and most critical for very intensive drainage water flow. Furthermore, these periods are critical for the leaching of phytochemicals used in agriculture. Therefore, we consider only these periods of a campaign in order to predict the amount of drained water flow more accurately.

The models have been learned by 3 different machine learning algorithms: regression trees, ensembles, and constrained induction of polynomial equations for regression. The regression trees have been constrained to a maximal depth of 4 levels, while the ensembles were constructed in 10 iterations, i.e. contains 10 trees.

The attribute set used in this study is the same as defined previously, except in the case of polynomial equation induction, where only the real valued attributes were used. The complete set of attributes used in the polynomial equation induction is listed in Table 17. As compared to Table 11, the attributes *Day*, *Season*, *Drainage Season* and *Crop* are excluded.

Table 17: *Attributes used for polynomial equation induction*. All of the selected attributes are numeric.

Attributes	
CDCoef	Slope
Temp	Runoff
RainfallA1	DrainageN1

The results from the data analyses with ensembles and regression trees are presented in Table 18 and Table 19, while Table 20 presents the results from polynomial equation discovery. The models learned with ensembles and regression trees are tested on the training data set and with 10-fold cross validation, respectively. On the other hand, the polynomial equation induction has been evaluated with test sets. Namely, the polynomial equations were induced from data of 8 fields, while one field has been used for the test set. This has been done in 9 iterations, where data from each field has been used as a test set in exactly one iteration. The ensembles and regression trees were built from data of all fields, field T3, and field T6.

Table 18: *The accuracy of the drainage season based predictive model over the training data*. *Std. Dev.* stands for standard deviation of the target variable on the training data set.

Fields	Model	Std. Dev.	RMSE	RRSE	Corr. coeff. (r)
All	Regression tree	3.17	1.7023	53.71 %	0.8437
All	Random forest	3.17	2.0717	65.36 %	0.8209
T3	Regression tree	3.056	1.5927	52.12 %	0.8534
T3	Random forest	3.056	1.6550	54.16 %	0.8782
T6	Regression tree	3.6	1.5709	43.64 %	0.8997
T6	Random forest	3.6	1.7084	47.46 %	0.8961

Table 19: *The predictive performance of the learned drainage season predictive model on unseen data, estimated by 10-fold cross validation*. *Std. Dev.* stands for standard deviation of the target variable from the training data set.

Fields	Model	Std. Dev.	RMSE	RRSE	Corr. coeff. (r)
All	Regression tree	3.17	1.7262	54.46 %	0.8388
All	Random forest	3.17	1.8865	59.52 %	0.8305
T3	Regression tree	3.056	1.9880	65.06 %	0.7637
T3	Random forest	3.056	1.8177	59.49 %	0.8210
T6	Regression tree	3.6	1.8076	50.22 %	0.8659
T6	Random forest	3.6	1.9639	54.56 %	0.8500

Table 20: *The predictive performance of the models learned by polynomial equation induction*. Models learned from data on 8 fields are tested on the data from the remaining one field. The *Std. Dev.* stands for standard deviation of the target variable on the training data set.

Fields	Test field	Std. Dev.	RMSE	RRSE	Corr. coeff. (r)
All	T3	3.187	2.1119	66.26 %	0.7855
All	T4	3.188	1.7220	54.01 %	0.8273
All	T5	3.163	2.2478	71.06 %	0.7467
All	T6	3.096	2.1784	70.36 %	0.8096
All	T7	3.229	1.3286	41.15 %	0.7812
All	T8	3.210	1.3813	43.03 %	0.7839
All	T9	3.210	1.5927	49.62 %	0.7434
All	T10	3.130	1.6672	53.26 %	0.7766
All	T11	3.108	1.6274	52.36 %	0.7841
T3	T6	3.056	2.1251	69.54 %	0.8125
T6	T3	3.600	2.0961	58.22 %	0.7745

The regression tree learned on data from field T6 was the most accurate. In this case, “Runoff” appears as most important attribute. Second most important attribute is the amount of drained water from a day before, “DrainageN1”. Apparently, the model (Figure 22) has a similar structure as right sub-tree of the model (Figure 15) built on data for a whole campaign. Namely, the same attributes are recognized as most important for predicting amount of drained water during the drainage seasons.

The size of the model (Figure 22) is 4 levels in depth and it has 16 leaves. Unlike the model tree (Figure 15) learned on a data from while campaign, this model (Figure 22) contains real values in the leaves. These values are predictions of amount of drained water under circumstances defined in upper part of the model.

The regression tree model is shown in Figure 22, followed by the visualization of its predictions for the 2009/2010 campaign. Furthermore, the best polynomial equation model is presented in Table 21. The remaining models are given in Appendix B.

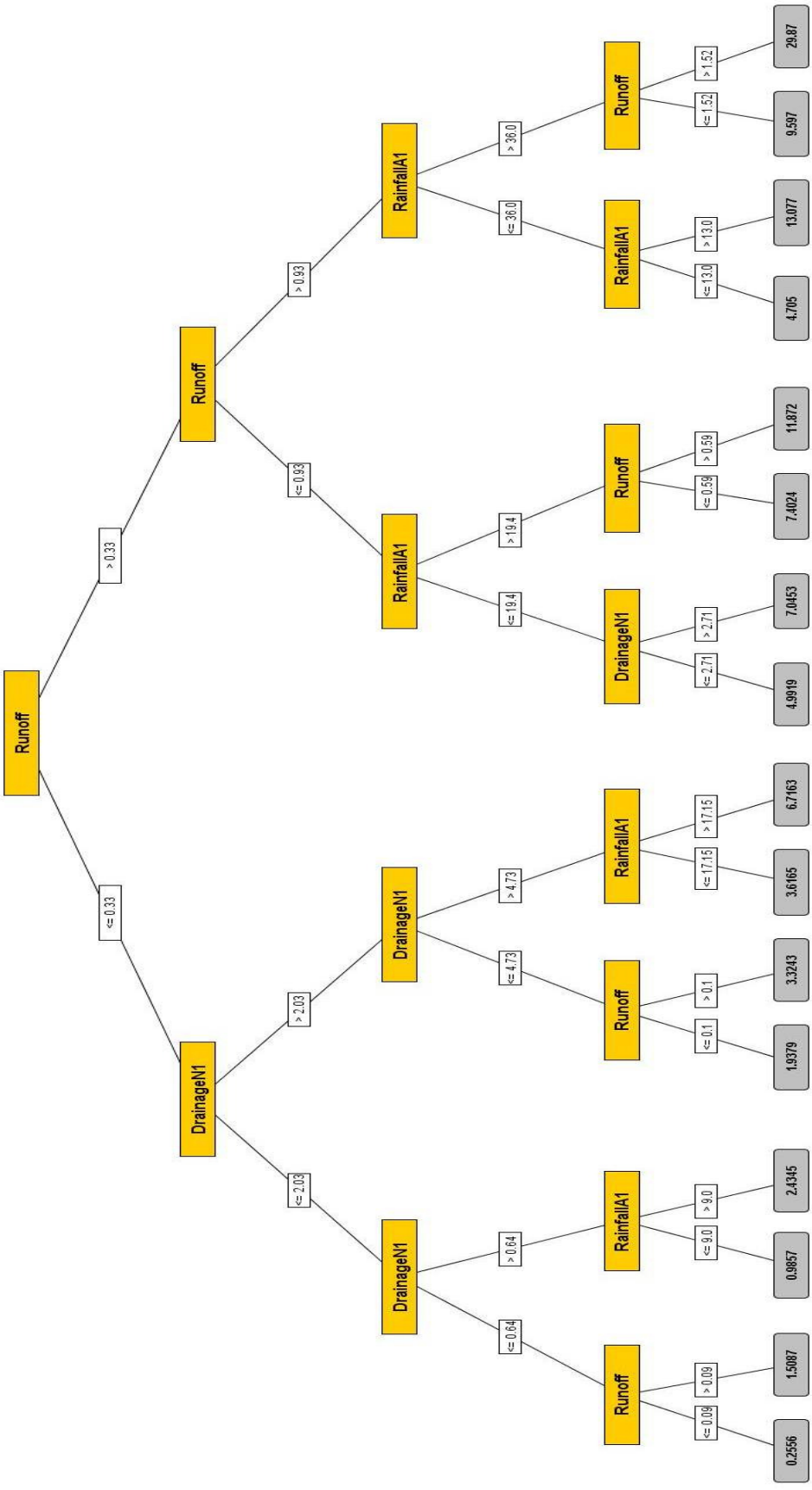


Figure 22: *Regression tree model for predicting the daily amount of drained water during a drainage season*. This was the most accurately predictive model for the task of predicting daily amount of drained water flow.

Table 21: *Most accurate polynomial equation model for predicting drained water flow within a drainage season*. The polynomial equation model has been learned from data of all fields, except field T4, which has been used as test set.

Model (All/T4)
$Drainage = 0.0196445 * RainfallA1 * Temp$ $+ 0.33246 * DrainageN1$ $+ 0.000662861 * CDCoef^2 * RainfallA1^2 * Slope$ $+ 0.00000107253 * Runoff * DrainageN1 * Temp^2 * RainfallA1^2$ $- 0.00115983 * Runoff^2 * DrainageN1 * Slope^3$ $- 0.00114057 * Temp^2 * RainfallA1$ $+ 0.00153725 * RainfallA1^2$ $+ 1.63563 * Runoff * Slope$ $- 1.90622 * Runoff$ $- 0.0231748 * Slope^2 * RainfallA1$ $+ 0.0654042 * RainfallA1 * Slope$ $- 0.00755737 * Slope^3 * CDCoef^3 * Runoff^2$ $+ 0.0675951 * Slope$ $- 0.146702$

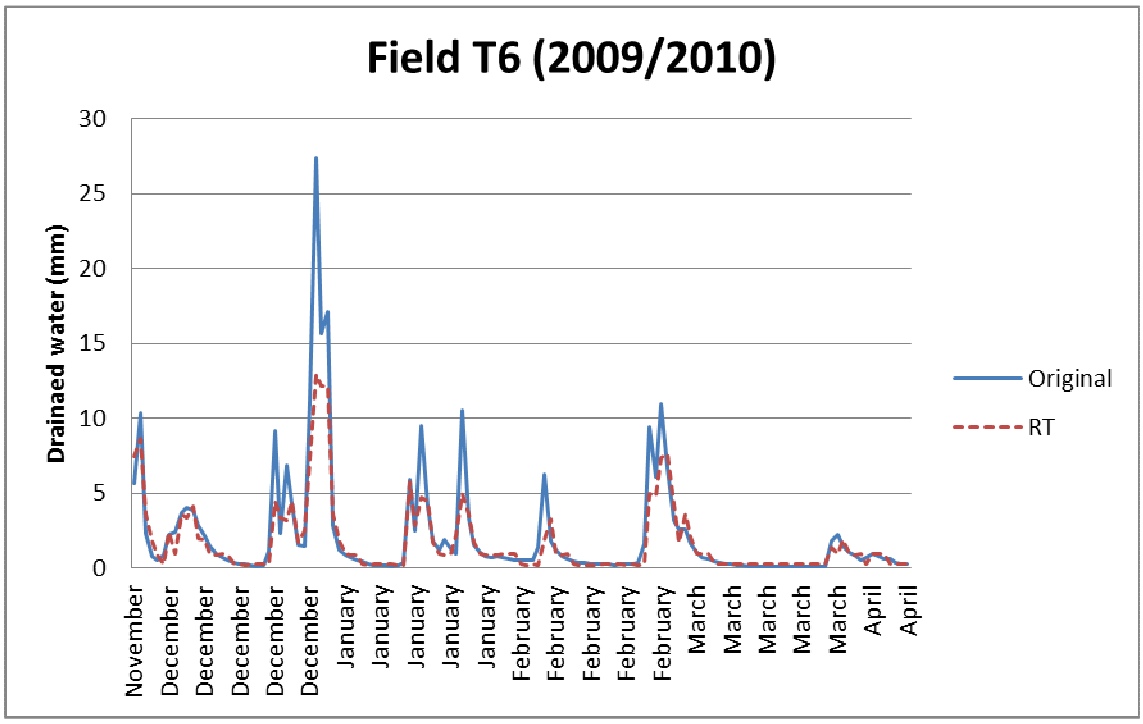


Figure 23: *Predicted amount of drained water (mm) for field T6*. Visualization of the predicted values from a RT – regression tree model, learned on data from field T6, compared with the measured (original) values. The winter drainage season period of November 28th, 2009 – April 9th, 2010, was considered.

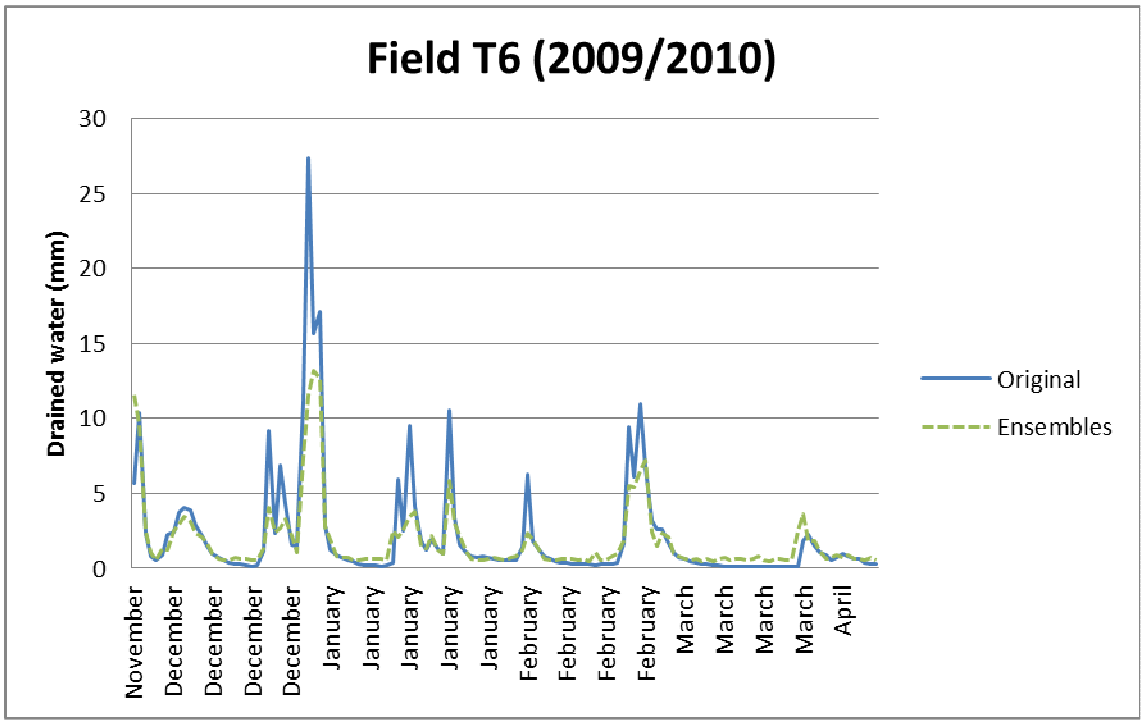


Figure 24: *Predicted amount of drained water (mm) for field T6*. Visualization of the predicted values from an ensemble model, learned on data from field T6, compared with the measured (original) values. The winter drainage season period of November 28th, 2009 – April 9th, 2010, was considered.

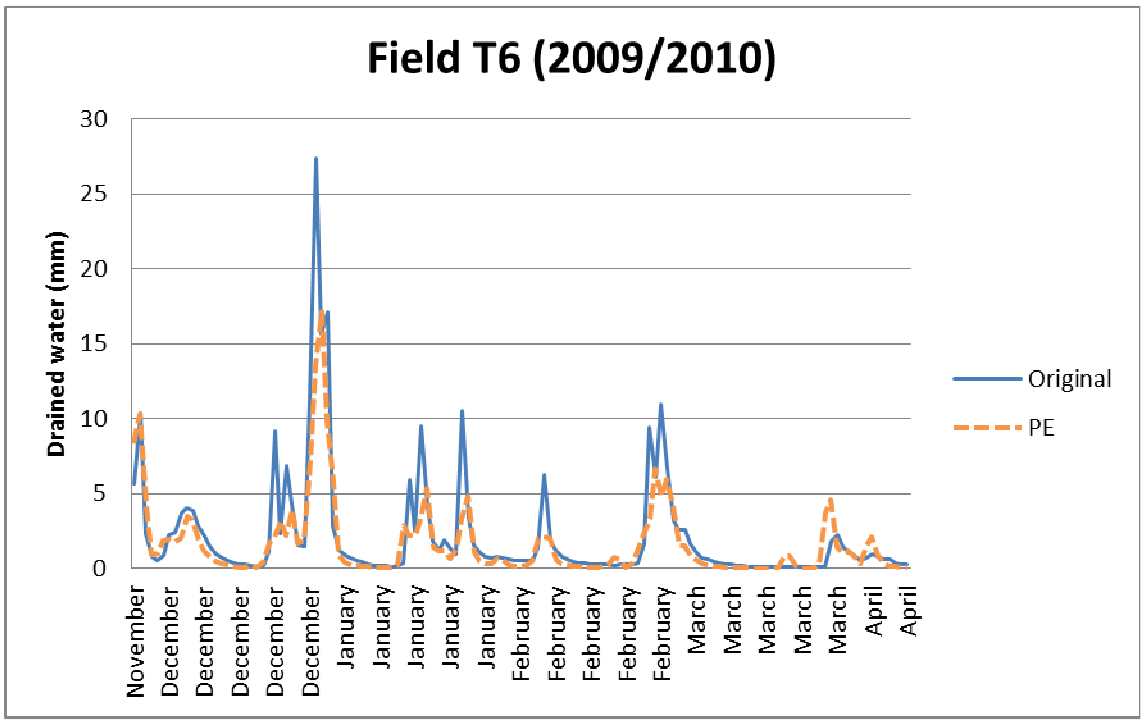


Figure 25: *Predicted amount of drained water (mm) for field T6*. Visualization of the predicted values from a PE - polynomial equation, learned on data from field T6, compared with the measured (original) values. The winter drainage season period of November 28th, 2009 – April 9th, 2010, was considered.

The most accurate regression tree model (Figure 22) successfully follows the trend of extreme drainage events with lower amplitude. Moreover, the learned predictive model is stable and almost never overestimates the drained water flow. The other two learned models, however, overestimate the drained water flow. Furthermore, the complexity of the regression tree model is small, which in a positive way supports the hypothesis.

Therefore, the learned regression tree model can be used in further practices in order to predict the amount of the drained water. Namely, the learned model used in combination with the models presented previously for predicting the start and the end of a drainage season gives us the possibility to make on field analysis, which could present practical information for planning agricultural practices, whose application depends on the drainage season and the intensive drainage events that appear at that time.

7.3 Evaluation of the integrated predictive model

The last part of our study of building models for predicting the amount of drained water from a field focuses on integration of the available expert knowledge with the models learned from data. The integration can be described as the support from an expert in the form of defining the most important attribute(s). We then use this knowledge to define the root of the tree that should be then learned from data.

However, the available expert knowledge provided by ARVALIS was in format inappropriate for the integration that we proposed. Therefore, we first used data mining and machine learning methods for restructuring the available expert knowledge. We then integrated these two approaches by using the regression tree algorithm implemented in the CLUS system.

7.3.1 Structuring the expert knowledge

First, in order to use data mining and machine learning methods for restructuring the available expert knowledge, we transform the tables from the expert knowledge and merge them into a single table. Then, the additional information in the form of text was extracted and appended to the created table. This process is described in detail in Section 4.1. Next, the created table was saved as a data set that was used for training decision trees.

Finally, the algorithm J48 for learning decision (classification) trees, implemented in the WEKA data mining suite was applied to the data set. The outcome is a set of learned decision trees, one for each of the 12 targets defined in the data set. For the purpose of this study, we consider only the learned model for predicting the intensity of the drained water flow (Figure 26).

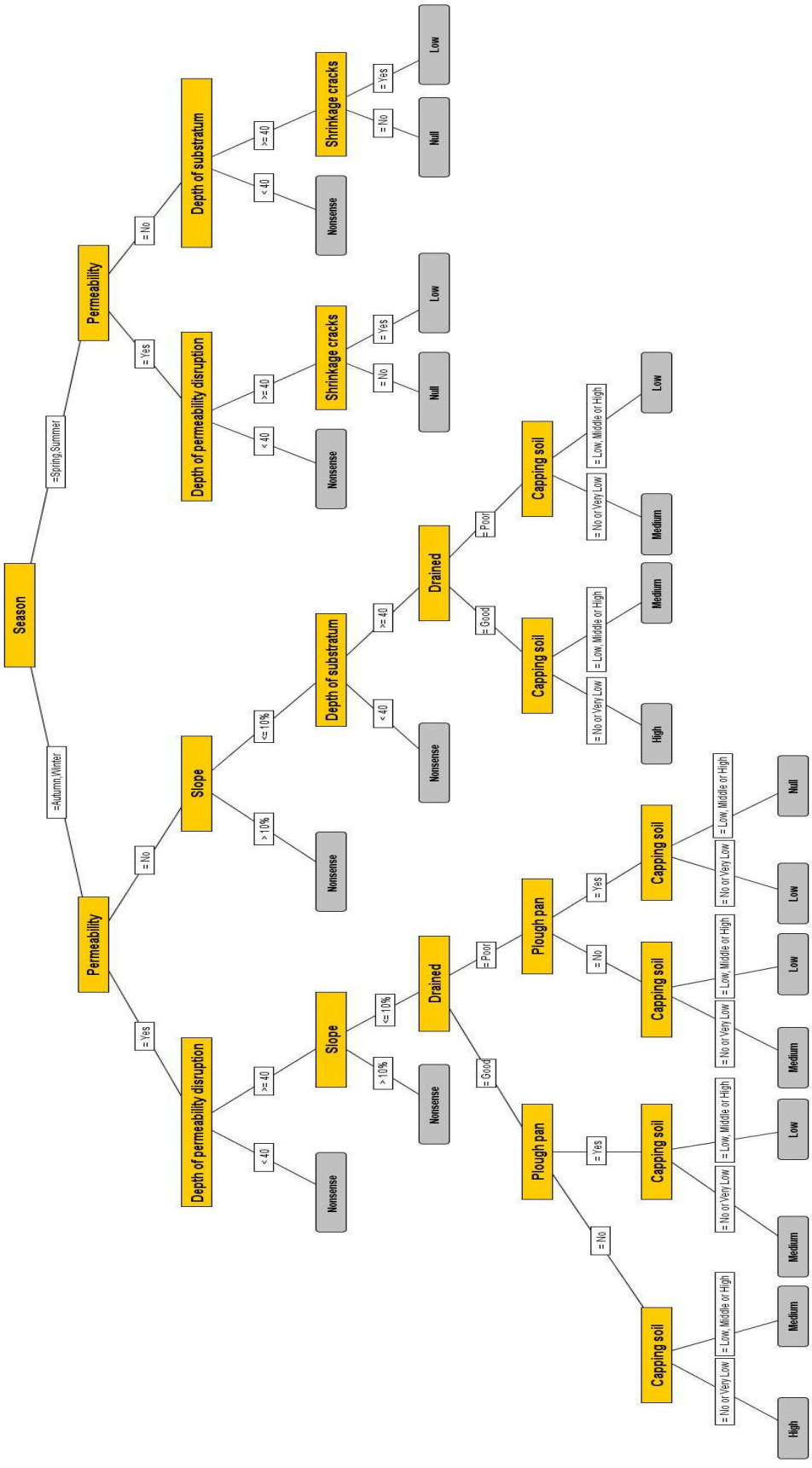


Figure 26: *Restructured expert knowledge for classification of intensity of drainage water flow*. The learned decision tree contains a path from the root of the tree to a leaf for each existing rule defined in the expert knowledge.

The learned model was tested on the training data set, which resulted in 100 % of correctly classified examples (rules). The accuracy of 100 % proves the "coverage" of each existing rule in the expert knowledge. This means that each existing rule has a path from the root of the decision tree to a leaf that predicts the correct value for the intensity of drained water flow.

Furthermore, the restructured expert knowledge can be validated with existing measured data. Therefore, we use the data from the La Jaillière experimental site to validate the model. For this purpose, the data of drained water flow were discretized with thresholds shown in Table 22.

Table 22: *Defined intervals for discretization of drained water flow*. Range of real values of drained water (mm). The corresponding to each discrete value is given.

Discrete value	Range of real values
Null	0
Low	(0–1]
Medium	(1–10]
High	10 +

The model was validated with the discretized measured data, resulting in 40 % of correctly classified examples. The reason for the low accuracy of the restructured EK is due to the fact that it is constructed in a general way and aimed to be applicable on different fields with different local specifics with additional support from experts for applying to local circumstances. Furthermore, the available expert knowledge does not take in account the meteorological properties.

The resulting predictive model defines the "Season" attribute as most important attribute. Hence, it must be included in the model for predicting the amount of drained water flow. Also, the other two attributes ("Permeability" and "Depth of permeability disruption") that follow the top most one are important in the process of predicting the amount of drained water flow. However, they depend on the soil properties. Therefore, we did not consider them in a process of integration due to the lack of data from different sites. Namely, the data from La Jaillière experimental site have uniform soil properties.

7.3.2 Integration of expert knowledge in model learning

The last step in learning a model that achieves better performance for predicting the amount of drained water flow from a field is to integrate the existing expert knowledge in the process of model learning. The integration is achieved by a process of constrained learning of models from data where the constraints are in the form of partial decision trees. Therefore, we "supervised" the process of learning a model from data by defining constraints in this form. The remainder of the decision tree has been learned by the algorithm for learning regression trees, implemented in the CLUS system.

We proceed as follows. First, we use the whole data set defined from data for all fields over whole campaigns. The enforced expert knowledge to be integrated contains information for the actual season. Therefore, the partial decision tree is consists of one root node representing the "Season" attribute.

Second, only the data from the defined drainage seasons were considered in the learning process. Furthermore, as improvements in seeking the most "active" periods when extreme drainage events are registered, we use the information from the "Drainage Season" attribute instead of "Season". Therefore, the partial decision tree consists of one root node representing the "Drainage Season" attribute.

Hereafter, we present the results and performance of the integrated predictive model. The models have been evaluated on the training data and by 10-fold cross validation. Moreover, the results are compared using RMSE, RRSE and correlation coefficient as performance measures. The results are shown in Tables 23–24 and Tables 25–26 for campaign based and drainage season based predictive models. The campaign based predictive model that has been learned from data of field T6 is shown in Figure 27 and its predictions are visualized for fields T3 and T6 with data from the campaign 2009/2010 (Figure 28–30).

Table 23: *The accuracy of constrained campaign based predictive models on training data*. *Std. Dev.* stands for standard deviation of the target variable from the training data set.

Fields	Model	Std. Dev.	RMSE	RRSE	Corr. coeff. (r)
All	Regression tree	1.918	0.9738	50.76 %	0.8617
T3	Regression tree	1.976	0.9174	46.43 %	0.8857
T6	Regression tree	2.279	0.8712	38.24 %	0.9240

Table 24: *The predictive performance of constrained campaign based predictive model*. The accuracy on unseen data is estimated by 10-fold cross validation. *Std. Dev.* stands for standard deviation of the target variable on the training data set.

Fields	Model	Std. Dev.	RMSE	RRSE	Corr. coeff. (r)
All	Regression tree	1.918	1.0137	52.84 %	0.8490
T3	Regression tree	1.976	1.1526	58.33 %	0.8155
T6	Regression tree	2.279	1.1274	49.48 %	0.8705

Table 25: *The accuracy of constrained drainage season based predictive models on training data*. *Std. Dev.* stands for standard deviation of the target variable on the training data set.

Fields	Model	Std. Dev.	RMSE	RRSE	Corr. coeff. (r)
All	Regression tree	3.17	0.95	49.52 %	0.8689
T3	Regression tree	3.056	0.9328	47.21 %	0.8816
T6	Regression tree	3.6	0.9162	40.21 %	0.9156

Table 26: *The predictive performance of constrained drainage season based predictive model*. The accuracy on unseen data is estimated by 10-fold cross validation. *Std. Dev.* stands for standard deviation of the target variable on the training data set.

Fields	Model	Std. Dev.	RMSE	RRSE	Corr. coeff. (r)
All	Regression tree	3.17	0.9724	50.69 %	0.8621
T3	Regression tree	3.056	1.1384	57.61 %	0.8206
T6	Regression tree	3.6	1.0845	47.59 %	0.8802

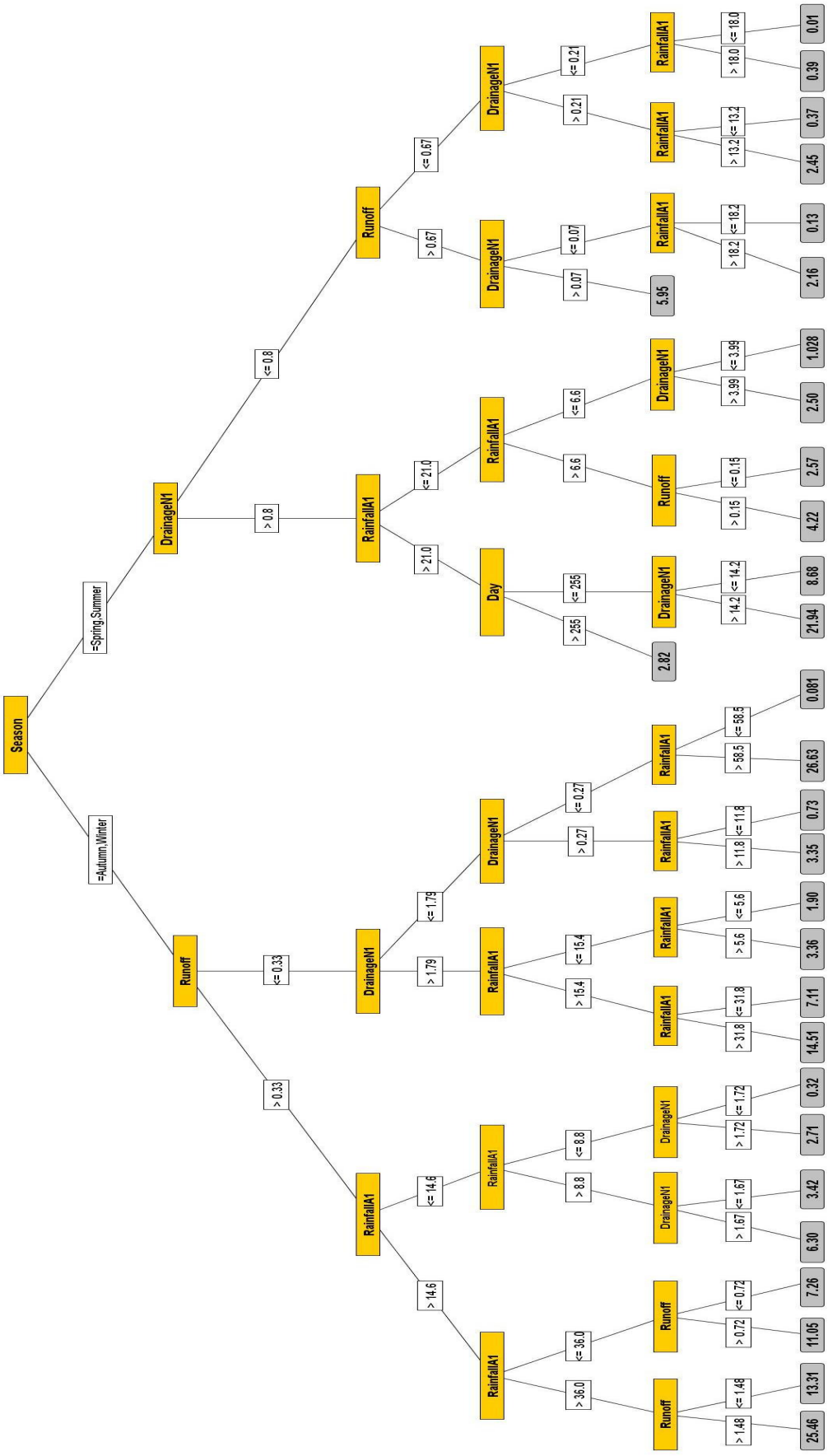


Figure 27: *Integrated predictive model for predicting the amount of drained water flow*. The campaign based regression tree has been built on data from field T6.

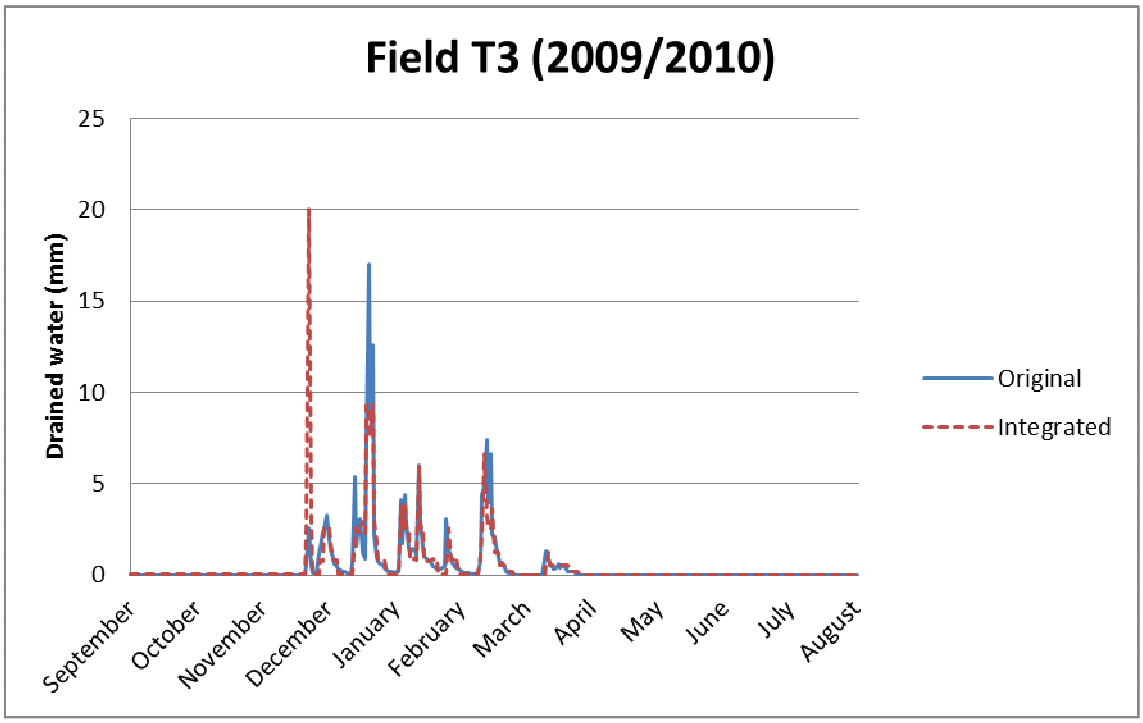


Figure 28: *Predicted amount of drained water (mm) for Field T3*. Visualization of the predicted values from an integrated campaign based predictive model, learned on data from field T6, compared with the measured (original) values. The period of September 1st, 2009 – August 31st, 2010 was considered.

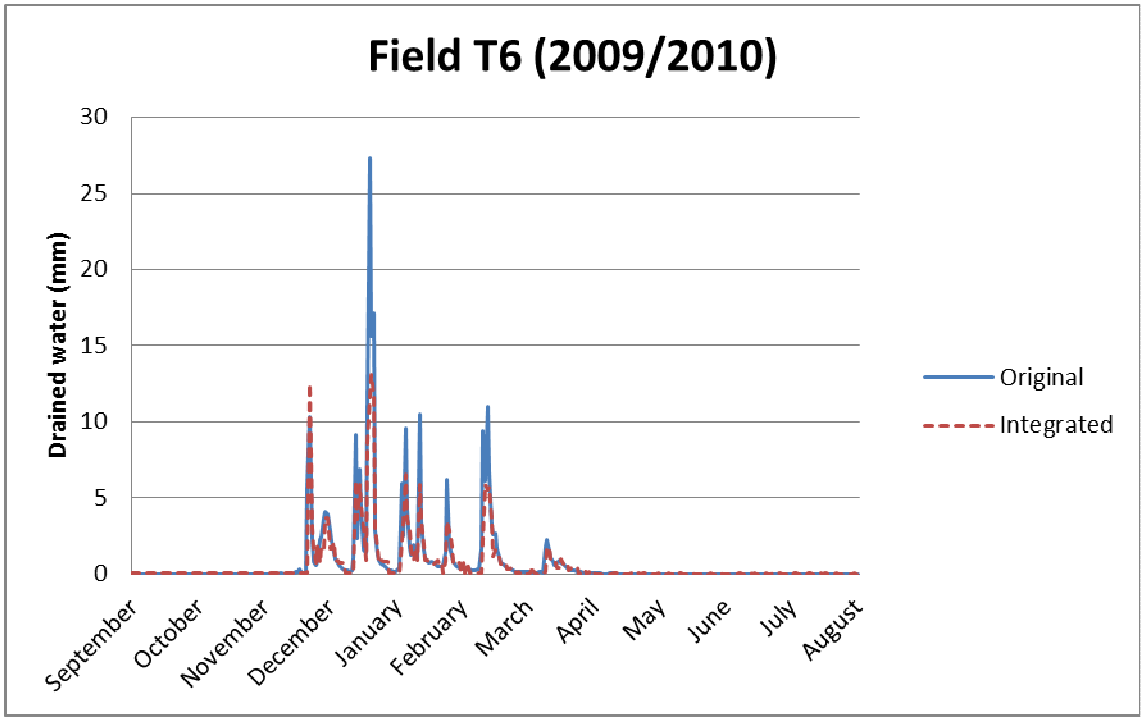


Figure 29: *Predicted amount of drained water (mm) for Field T6*. Visualization of the predicted values from the integrated campaign based predictive model, learned on data from field T6, compared with the measured (original) values. The period of September 1st, 2009 – August 31st, 2010 was considered.

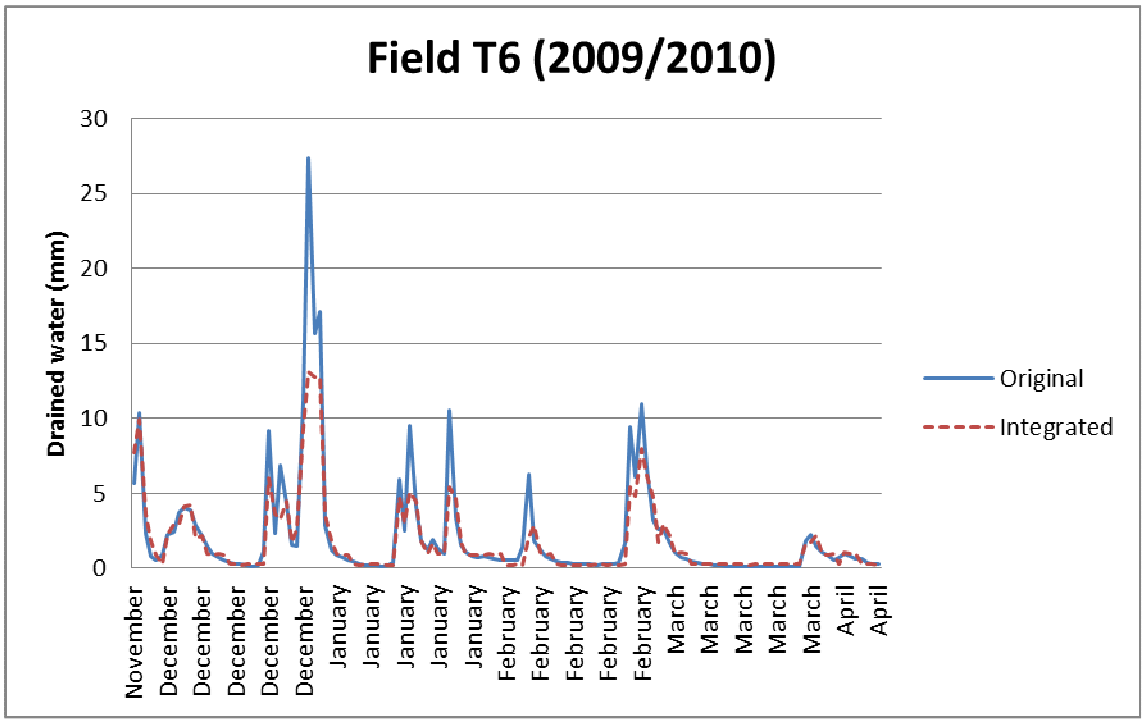


Figure 30: *Predicted amount of drained water (mm) – Field T6*. Visualization of the predicted values from the integrated drainage season based predictive model, learned on data from field T6, compared with the original values. The winter drainage period of November 28th, 2009 – April 29th, 2010, was considered.

The integrated models, which are based on both expert knowledge and the available data performed better than the models learned from data only. Although, in the case of campaign based predictive model learned from the data of field T6 (Figure 27) the model overestimates the values in some periods of the campaign, the model is still enough accurate and can be considered for further usage as it is or with other some improvements. On the other hand, the drainage season based model (Figure 30) shows that these overestimations are corrected. Therefore, this model is the most accurate in the prediction of the amount of drained water flow from a field. Overall, the integrated models are better than either the models built from data only or the restructured expert knowledge.

They improve the general recommendations of the expert knowledge by adjusting the existing expert knowledge based on the data collected and the predictions of the integrated predictive models.

The integrated models give a better explanation and the possibility of additional understanding of the interactions between the features (attributes) involved in the process of water flow from a field.

8 Conclusion

The thesis presents a study where data mining and machine learning methods have been used to predict the intensity drained water flow. The predictive models have been learned from data collected at the La Jaillière experimental site, considering expert knowledge at some stages. The data have been analyzed along several dimensions of feature combination and data mining and machine learning method selection.

The feature combination includes feature generation and feature selection. First, features based on existing knowledge were generated from the available data. We defined a feature set that can well describe the process of drainage. Second, various techniques for feature selection and feature ranking were applied in order to define the final feature set that can describe the dependences between explanatory and target variables in a best way.

A successful study relies on good methodological design. The methodology used in this study is based on data mining/machine learning models that predict the amount of drained water flow from a field. This kind of methodology is a powerful way of integrating the available expert knowledge and existing data from an experimental site. Namely, the machine learning approach includes predictive models that represent the obtained knowledge in patterns such as regression and model trees, ensembles, and polynomial induced equations. We select regression and model trees patterns, because they express the gained knowledge in the most understandable way. Furthermore, the possibility of model trees to include within a rule a linear regression model additionally improves the accuracy of the learned model. In addition, with ensembles we investigated the range of accuracy, since they aim to improve the predictive performance of their base classifier. The algorithms that have been applied in this study are implemented in the WEKA data mining suite, the CLUS system, and the CIPER tool.

In addition, the experimental design of our study has been created in order to solve the previously defined problem and support the hypothesis in the most accurate way. The study has been organized in phases and step by step the obtained results have been improved.

First, we used all the data from the La Jaillière experimental site and learned three different types of predictive models: regression trees, model trees and ensembles. These were used for learning predictive models from data of all fields, as well as particular fields (T3 and T6). The most accurate predictive model was the model learned on data from field T6. This model performed well during the whole campaign, except for the extreme drainage events. Hence, the model accurately predicts the amount of drained water when no drainage or a low amount of drained water is registered. This is due to the fact that the target variable (Drainage) has an exponential distribution over the test set. This means that further splitting is required for particular periods during a campaign.

Therefore, in the second step, we used the dates estimated from experts that defined the periods in a campaign with intensive drainage events. These periods are defined as drainage seasons. Thus, in this phase we consider only these periods and not a whole campaign. Also, three machine learning approaches have been used: regression trees, ensembles and polynomial equations. Again, the regression tree learned on data from field T6 performed in a most accurate way compared with the other two models

(ensembles and polynomial equations). Since we learned these models only for prediction of the amount of drained water during the drainage season, the most accurate models (regression tree) have reduced the errors that were high in the previous phase and more accurately "followed" the trend of real data. More importantly, in this phase, the best model overestimates the predicted (as compared to actual value) in very few cases, unlike the other two models (learned by ensembles and polynomial equations) that have this disadvantage.

In addition, in this phase, we tried to predict the start and the end of the winter drainage period based on easily accessible meteorological data (rainfall and temperature). This goal was set due to the availability of the data (expert's estimated dates), which are estimated under not very well defined conditions. Namely, the start of a drainage season is constrained with 5 mm of cumulative drainage amount since the beginning of a campaign. On the other hand, the winter drainage season ends when the weekly cumulative drainage is below 1 mm, which is not very informative. Therefore, we learned classification trees that preformed very accurately with 10-fold cross validation. Namely, the models consist of data, which could be obtained from weather forecasts, so that they can give predictions whether the drainage will begin soon (some day in the next week) or whether it will stop in the following week. While the data from the experimental station can be used only for ex post analysis, the models we induced give us the possibility to make also ex ante analysis, which could present a practical tool for planning agricultural practices, whose application depends on the drainage period.

In the last phase, the study focuses on the integration of the available expert knowledge with standard data mining and machine learning methods in order to achieve better performance of the learned models. We consider the available expert knowledge written in the form of decision rules in tables. First, we restructured the knowledge in a suitable format – decision trees. Then, we constrained the standard learning process with a partial decision tree generated from the expert knowledge. This phase covers both learning predictive models for entire campaigns and learning predictive models for drainage seasons only. The results present improvements for the integrated models which are based on both expert knowledge and the available data. Consequently, we can say that if we consult the expert knowledge during the learning of the data mining models it will lead us to improvements in performance of the models. The topmost attributes of the restructured expert knowledge were selected to form the partial decision trees that we used in constrained learning process. But, due to the lack of data from different regions where the soil has different properties, we were not able to support the learning process with a wider "expert's recommendation" from the expert knowledge. Hence, only the most important attribute "Season" was selected to take the root position in the integrated predictive model. The rest of the decision tree was learned from data.

Compared with other previously learned models, the integrated predictive models performed better. In addition, the built models improve the general recommendations from expert knowledge by adjusting the existing expert knowledge based on the data collected and the predictions of the models learned by data mining. Finally, the integrated model allowed us to better understand the interactions between the features (attributes) involved in the process of water flow in a field and the amount of drained water as the target variable.

To summarize, the study presented in this thesis justifies our expectation in general. The results from this study are better than the results or predictions from the expert knowledge, which is general and not accurately applicable to local circumstances. Hence, higher correlation coefficients and lower error rates are obtained. The study results in highly accurate predictive models for prediction of the amount of drained water from a field. Furthermore, we successfully learned a model for estimating the beginning and

ending of a winter drainage season. The existence of spring drainage seasons and their estimation of start and end of a period appeared as an additional task that needs to be addressed and was not considered in this study. Finally, we expand the approach of integration of an expert knowledge within the process of learning predictive models.

8.1 Contributions

Because of the interdisciplinary of the thesis, it contributes to two scientific areas: information technologies and agriculture.

From the information technologies point of view, it improves and extends the application of different machine learning techniques to a new area and raises interesting application issues. The main contributions in this scientific area are:

- A new methodology, based on machine learning, to predict the amount of drained water from a field. The methodology integrates existing expert knowledge and data from experimental fields. The main improvement is that the methodology is general and re-useable across different geographic areas, with the ability of downscaling to respect local characteristics (features).
- The evaluation of the methodology, from different relevant aspects. We evaluate the methodology with data from different data sources (fields in the experimental area) and compare the results of different machine learning algorithms (regression and model trees, ensembles and polynomial equations).

From the agricultural point of view, the thesis makes the following main contributions:

- Structuring of the expert knowledge improves its interpretability and the understanding of water flows in the fields. This will make knowledge easy for distribution among agricultural experts.
- Comparisons of the existing expert knowledge with the knowledge learned from data, which can lead to new expert knowledge.

8.2 Further work

In future work, we plan to extend the role of the expert knowledge in the process of learning models for predicting water flows. This is highly related to the availability of data from experimental sites at different locations. We will try to apply the approach of integration of expert knowledge and data to the learning models for other types of water flows such as runoff.

Modelling water flows is the first step towards our overall goal. Having this step completed, we will move towards modeling pesticide leaching in agriculture. Finally, the developed and upgraded models will be used for building a decision support system for proposing appropriate mitigation measures for water protection from phytochemicals.

Due to the complexity of the overall problem, we will develop our approach in several stages. The present thesis covers the stage of modeling drainage water flow, while the stages of modeling pesticide leaching in agriculture and developing a decision support system for proposing mitigation measures will be completed during my doctoral studies.

9 Acknowledgements

I would like to acknowledge and thank my supervisor Prof. Dr. Marko Debeljak and co-supervisor Prof. Dr. Sašo Džeroski, who accepted me as a Master of Science student. I am grateful for all the support and guidance that they have given me during my studies. Also, many thanks to Dr. Florence Leprince, Julie Maillet-Mezeray and Benoît Real from ARVALIS Institute and Dr. Aneta Trajanov from Jožef Stefan Institute for their help, collaboration and assistance with the thesis.

Research presented in this thesis was conducted as a part of the project EVADIFF - "Evaluation et développement de modèles et outils d'aide à la décision utilisés pour la prévention des pollutions diffuses par les produits phytopharmaceutiques" ("Evaluation of existing models and development of new decision-making tools to prevent diffuse pollution caused by plant protection products"), which was financially supported by ARVALIS Institute – France.

Finally, I would like to express my special gratitude to Zorica Angjelovska for supporting me and proofreading the thesis, and everyone who has helped me and supported me during my master studies at Jožef Stefan Institute – International Post-graduate School.

10 References

Ahuja, L. R.; Rojas, K. W.; Hanson, J. D. Shaffer, J. J.; Ma L. (ed.) *The Root Zone Water Quality Model* (Water Resources Publications LLC, Highlands Ranch, CO, 2000).

Arlot, M. *Nitrate in water: actor drainage, drainage witness? Lessons from a hydrological approach and hydraulic* (Ph. D. thesis, University Pierre et Marie Curie, Paris, France, 1999).

Beard, G. Reports on soil profile characteristics and topsoil structure at the ARVALIS experimental site. *Technical Report* (La Jailliere, France, 2005).

Beven, K.; Germann, P. Macropores and water flow in soils. *Water Resource Research* **18**, 1311–1325 (1982).

Blockeel, H.; Struyf, J. Efficient algorithms for decision tree cross-validation. *Journal of Machine Learning Research* **3**, 621–650 (2002).

Branger, F.; Debionne, S.; Viallet, P.; Braud, I.; Vauclin, M. Using the LIQUID framework to build an agricultural subsurface drainage model. In: *Proceedings of the 7th Hydroinformatics International Conference.* 2024–2031 (Nice, France, 2006).

Breiman, L. Bagging predictors. *Machine Learning* **24**, 123–140 (1996).

Breiman, L. Random forests. *Machine Learning* **45**, 5–32 (2001).

Brederhoeft, J. The conceptualization model problem – surprise. *Hydrogeology Journal* **13**, 37–46 (2005).

Clothier, B. E.; Green, S. R.; Deurer, M. Preferential flow and transport in soil: progress and prognosis. *European Journal of Soil Science*, volume 2–13 (2008).

Clothier, B. E.; Green, S. R. Roots: the big movers of water and chemical in soil. *Soil Science* **162**, 534–543 (1997).

Debeljak, M.; Džeroski, S. Decision trees in ecological modelling. In: *Jopp, F.;Reuter, H.; Breckling, B.(ed.) Modelling Complex Ecological Dynamics*. 1–15 (Springer, Berlin, Germany, 2011).

Dietterich, T. G. Machine Learning Research: Four Current Directions *AI Magazine* **18**, 97–136 (1997).

Dietterich, T. Ensemble Methods in Machine Learning. In: *Proceedings of the 1st international workshop on multiple classifier systems (MCS '00).* 1-15 (Springer, Berlin, 2000).

Dubus, I.; Beulke, S.; Brown, C. Calibration of pesticide leaching models: critical review and guidence for reporting. *Pest Managemenet Science* **58**, 745–758 (2002).

Džeroski, S.; Panov, P.; Ženko, B. Ensemble Methods in Machine Learning. In: *Encyclopedia of Complexity and Systems Science* (Springer, New York, USA, 2008).

Ersahin, S.; Papendick, R. I.; Smith, J. L.; Keller, C. K.; Manoranjan, V. S. Macropore transport of bromide as influenced by soil structure differences. *Geoderma* **108**, 207–223 (2002).

Feigenbaum, E. A.; McCorduck, P. *The fifth generation - 1st edition* (Addison-Wesley, Reading, MA, USA, 1983).

Fenech, A.; Foster, J.; Hamilton, K.; Hansell, R. Natural capital in ecology and economics: an overview. *Environmental Monitoring and Assessment* **86**, 3–17 (2003).

FOCUS-WG FOCUS: surface water scenarios in the EU evaluation process under 91/414/EEC. In: *SANCO/4802/2001-rev2* (Forum for the Coordination of Pesticide Fate Models and their Use, 2001).

FOOTWAYS 2009. Web application. See also: http://www.footways.eu (accessed July, 2012)

Freund, Y.; Schapire, R. E. Experiments with a new boosting algorithm. In: *Proceedings of the 13th International Conference on Machine Learning (ICML '96).* 148–156 (Morgan Kaufmann, San Francisco, 1996).

Gerke, H. H.; Köhne, J. M. Dual-permeability modelling of preferential bromide leaching from a tile drained glacial till agricultural field. *Journal of Hydrology* **289**, 239–257 (2004).

Gerke, H. H.; van Genuchten, M. Th. A dual-porosity model for simulating the preferential movement of water and solutes in structured porous media. *Water Resource Research* **29**, 305–319 (1993).

Germann, P. F. Kinematic wave approximation to infiltration and drainage into and from soil macropores. *Transactions ASAE* **28**, 745–749 (1985).

Hawken, P.; Lovins, A.; Lovins, L. H. *Natural Capitalism* (Back Bay Books, New York, USA, 1999).

Jarvis, N. J. The MACRO model (Version 3.1): Technical Description and Sample Simulations. In: *19th Reports and Dissert* (Department of Soil Sciences, Swedish University of Agricultural Sciences, Uppsala, Sweden, 1994).

Karalic, A. Linear regression in regression tree leaves. In: *Proceedings of 10th*

European conference on Artificial intelligence 440-441 (John Wiley & Sons, Ljubljana, Slovenia, 1992).

Kira, K.; Rendell, L. A. A practical approach to feature selection. In: *Proceedings of the 9th International Conference on Machine Learning.* 249–256 (Morgan Kaufmann, San Mateo, CA, USA, 1992).

Kreuger, J.; Nilsson, E. Catchment scale risk-mitigation experiences — key issues for reducing pesticide transport to surface waters. In: *Proceedings of BCPC 78th Symposium - Pesticide behavior in soil and water.* 319–324 (British Crop Protection Council, UK, 2001).

Kononenko, I.; Šimec, E.; Robnik - Šikonja, M. Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF. *Applied Intelligence* **7**, 39–55 (1997).

Kuzmanovski, V.; Džeroski, S.; Debeljak, M. *Integration of structured expert knowledge*. In: Proceeding of 4th Jožef Stefan International Postgraduate School – Student Conference (IPSSC). 137–143 (International Postgraduate School Jožef Stefan, Ljubljana, Slovenia, 2012).

Köhne, J. M.; Köhne, S.; Šimůnek, J. A review of model applications for structured soils: Water flow and tracer transport. *Journal of Contaminant Hydrology* **104**, 4–35 (2009).

Larsbo, M.; Jarvis, N. J. MACRO5.0: A model of water flow and solute transport in macroporous soil. *Technical Description*. 1–48 (Department of Soil Sciences, Swedish University of Agricultural Sciences, Uppsala, Sweden, 2003).

Larsbo, M.; Jarvis, N. J. Simulating solute transport in a structured field soil: uncertainty in parameter identification and predictions. *Journal of Environmental Quality* **34**, 621–634 (2005).

Lighthill, M. J.; Whitham, G. B. On kinematic waves — I: Flood movement in long rivers. II: A theory of traffic flow on long crowded roads. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* (Royal Society Publishing, London, UK, 1955).

Madrigal, I. *Retention of pesticides in soils buffer devices, grassed and wooded - the role of the organic content* (Ph. D. thesis, INA-PG, Paris, France, 2004).

Machine learning. Web resources for the term Machine Learning. See also: http://en.wikipedia.org/wiki/Machine_learning (accessed July, 2012).

Mitchell, T. *Machine Learning* (McGraw-Hill, Columbos, OH, USA, 1997).

Moore, E. C. S. *Sanitary Engineering* (B. T. Batsford, London, 1898).

Quinlan, J. *C4.5: Programs for Machine Learning* (Morgan Kaufmann, San Mateo, CA, USA, 1993).

Quinlan, J. Induction of decision trees. *Machine Learning* **1**, 81–106 (1986).

Quinlan, J. Learning with continuous classes. In: *Proceedings of 5th Australian Joint Conference on Arti cial Intelligence 92*. 343–348 (World Scientific, 1992).

Reichenberger, S,; Bach, M.; Skitschak, A.; Frede, H. G. Mitigation strategies to reduce pesticide inputs into ground- and surface water and their effectiveness: A review. *Science of the Total Environment* **384**, 1–35 (2007).

Richard L. A. Capillary conduction of liquids through porous mediums. *Physics* **1**, 318–333 (1931).

Struyf, J.; Džeroski, S. Constraint based induction of multi-objective regression trees. In: *Proceedings of the 4th International Workshop on Knowledge Discovery in Inductive Databases.* 223-233 (Springer, Port, Portugal, 2005).

Todorovski, L.; Ljubič, P.; Džeroski, S. Inducting polynomial equations for regression. *Lecture notes in computer science* **3201**, 441–452 (2004).

Wang, Y.; Witten, I.H. Inducing Model Trees for Continuous Classes. In: *Poster Papers of the 9th European Conference on Machine Learning* 128–137 (University of Economics, Prague, Czech Republic, 1997).

Wettschereck, D.; Aha, D. W.; Mohri, T. A Review and Empirical Evaluation of Feature Weighting Methods for a Class of Lazy Learning Algorithms. *Artificial Intelligence Review* **11**, 273–314 (1997).

Wikipedia — Evapotranspiration. http://en.wikipedia.org/wiki/Evapotranspiration (accessed July, 2012)

Witten, I. H.; Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques - 2nd edition* (Morgan Kaufmann, San Francisco, CA, USA, 2005).

Šimůnek, J.; Jarvis, N. J.; van Genuchten, M. Th.; Gärdenäs, A. Review and comparison of models for describing non-equilibrium and preferential flow and transport in the vadose zone. *Journal of Hydrology* **272**, 14–35 (2003).

Šimůnek, J.; Köhne, J. M.; Kodešová, R.; Šejna, M. Simulating non-equilibrium movement of water, solutes and particlesusing HYDRUS — a reviewof recent applications. *Soil & Water Resources: Special Issue* **3**, S42–S51 (2008c).

Šimůnek, J.; van Genuchten, M. Th. Modelling nonequilibrium flowand transport with HYDRUS. *Vadose Zone Journal: Vadose Zone Modelling* **7**, 782–797 (2008).

Šimůnek, J.; van Gencuhten, M. Th.; Šejna, M. Development and applications of the HYDRUS and STANMOD Software Packages and Related Codes. *Vadose Zone Journal* **7**, 587–600 (2008b).

Šimůnek, J.; van Genuchten, M. Th.; Šejna, M. The HYDRUS-1D software package for simulating the one-dimensional movement of water, heat, and multiple solutes in variably-saturated media (Version 3.0). In: *HYDRUS Software Series 1*. 1–270 (Department of Environmental Sciences, University of California Riverside, Riverside, CA, 2005).

Šimůnek, J.; van Genuchten, M. Th.; Šejna, M. The HYDRUS Software Package for Simulating Two- and Three-Dimensional Movement of Water, Heat, and Multiple Solutes in Variably-Saturated Media. *Technical Manual (Version 1.0)*. 1–241 (PC Progress, Prague, Czech Republic, 2006).

Šimůnek, J.; Wendroth, O.; Wypler, N.; van Genuchten, M. T. Non-equilibrium water flow characterized by means of upward infiltration experiments. *European Journal of Soil Sciences* **52**, 13–24 (2001).

Šimůnek, J.; Šejna, M.; Saito, H.; Sakai, M.; van Genuchten, M. Th. The HYDRUS-1D software package for simulating the one-dimensional movement of water, heat, and multiple solutes in variably-saturated media (Version 4.0). In: *HYDRUS Software Series 4*. 1–315 (Department of Environmental Sciences, University of California Riverside, Riverside, CA, USA, 2008a).

Šimůnek, J.; Šejna, M.; van Genuchten, M. Th. The HYDRUS-1D software package for simulating the movement of water, heat, and multiple solutes in variably-saturated media (Version 2.0). In: *IGWMC-TPS-70* (International Ground Water Modelling Center, Colorado School of Mines, Golden, CO, 1998).

Šimůnek, J.; Šejna, M.; van Genuchten, M. Th. The HYDRUS-2D software package for simulating two-dimensional movement of water, heat, and multiple solutes in variably-saturated media (Version 2.0). In: *IGWMC-TPS-53*(International Ground Water Modelling Center, Colorado School of Mines, Golden, CO, 1999).

Index of Figures

Index of Tables

Appendix A. Additional information on data

In this section we present some additional information and explanation of the data used in our study. First, all of the dates of the start and end of the drainage seasons are presented. Then, the amount of drained water from field T6 is visualized for each campaign in the period 1987–2011. The field T6 has been selected as most prominent field.

A1. Start and end of drainage seasons

Table A.1: *Start and end of drainage seasons*. Complete table of dates when drainage seasons starts and ends. The dates are for each field for all campaigns in period 1987–2011

Season	Field	Provided dates for start of winter drainage season	Provided dates for end of winter drainage season
1987-1988	T01	16.12.87	01.03.88
	T03	16.12.87	07.04.88
	T04	16.12.87	07.04.88
	T05	16.12.87	07.04.88
	T06	16.12.87	07.04.88
	T07	N/A	N/A
	T08	N/A	N/A
	T09	N/A	N/A
	T10	N/A	N/A
	T11	N/A	N/A
1988-1989	T01	21.02.89	28.03.89
	T03	21.02.89	30.03.89
	T04	21.02.89	27.03.89
	T05	21.02.89	27.03.89
	T06	21.02.89	28.03.89
	T07	N/A	N/A
	T08	N/A	N/A
	T09	N/A	N/A
	T10	N/A	N/A
	T11	N/A	N/A
1989-1990	T01	23.01.90	30.01.90
	T03	20.12.89	02.03.90
	T04	N/A	N/A
	T05	15.12.89	02.03.90
	T06	20.12.89	08.03.90
	T07	N/A	N/A

	T08	N/A	N/A
	T09	N/A	N/A
	T10	N/A	N/A
	T11	N/A	N/A
	T01	03.01.91	29.03.91
	T03	29.12.90	05.04.91
	T04	N/A	N/A
	T05	29.12.90	05.04.91
1990-1991	T06	01.01.91	05.04.91
	T07	30.12.90	27.03.91
	T08	30.12.90	27.03.91
	T09	30.12.90	23.03.91
	T10	N/A	N/A
	T11	N/A	N/A
	T01	N/A	N/A
	T03	N/A	N/A
	T04	N/A	N/A
	T05	N/A	N/A
1991-1992	T06	N/A	N/A
	T07	N/A	N/A
	T08	N/A	N/A
	T09	N/A	N/A
	T10	N/A	N/A
	T11	N/A	N/A
	T01	20.11.92	16.02.93
	T03	19.11.92	30.01.93
	T04	21.11.92	30.01.93
	T05	17.11.92	04.02.93
1992-1993	T06	21.11.92	10.02.93
	T07	18.11.92	31.01.93
	T08	20.11.92	10.02.93
	T09	19.11.92	15.02.93
	T10	19.11.92	08.02.93
	T11	21.11.92	07.02.93
	T01	16.10.93	04.03.94
	T03	13.12.93	25.04.94
	T04	22.12.93	05.03.94
	T05	14.12.93	09.03.94
1993-1994	T06	26.12.93	16.03.94
	T07	15.10.93	02.03.94
	T08	15.10.93	02.03.94
	T09	12.10.93	06.03.94
	T10	11.12.93	16.03.94

	T11	15.12.93	14.03.94
1994-1995	T01	30.09.94	19.03.95
	T03	29.10.94	27.03.95
	T04	29.09.94	02.04.95
	T05	22.10.94	22.03.95
	T06	04.11.94	05.04.95
	T07	04.11.94	26.03.95
	T08	04.11.94	22.03.95
	T09	30.10.94	24.03.95
	T10	09.10.94	27.03.95
	T11	24.09.94	02.04.95
1995-1996	T01	02.01.96	05.03.96
	T03	24.12.95	06.03.96
	T04	02.01.96	06.03.96
	T05	23.12.95	07.03.96
	T06	08.01.96	09.03.96
	T07	02.01.96	28.02.96
	T08	24.01.96	28.02.96
	T09	02.01.96	01.03.96
	T10	23.12.95	08.03.96
	T11	08.01.96	07.03.96
1996-1997	T01	21.12.96	22.03.97
	T03	01.12.96	14.03.97
	T04	06.12.96	11.03.97
	T05	27.11.96	06.03.97
	T06	30.11.96	15.03.97
	T07	02.12.96	05.03.97
	T08	29.11.96	01.03.97
	T09	29.11.96	05.03.97
	T10	01.12.96	08.03.97
	T11	19.12.96	13.03.97
1997-1998	T01	25.12.97	11.03.98
	T03	12.12.97	01.02.98
	T04	11.12.97	01.02.98
	T05	11.12.97	27.01.98
	T06	12.12.97	04.02.98
	T07	11.12.97	24.01.98
	T08	11.12.97	21.01.98
	T09	11.12.97	23.01.98
	T10	18.12.97	25.01.98
	T11	24.12.97	25.01.98
1998-1999	T01	14.11.98	15.03.99

	T03	25.10.98	17.03.99
	T04	10.12.98	22.03.99
	T05	24.10.98	13.03.99
	T06	25.10.98	25.03.99
	T07	26.10.98	14.03.99
	T08	25.10.98	12.03.99
	T09	25.10.98	13.03.99
	T10	14.11.98	15.03.99
	T11	13.12.98	14.03.99
1999-2000	T01	20.09.99	04.03.00
	T03	21.09.99	10.03.00
	T04	22.09.99	11.03.00
	T05	20.09.99	05.03.00
	T06	23.09.99	13.03.00
	T07	29.09.99	08.03.00
	T08	29.09.99	07.03.00
	T09	21.09.99	08.03.00
	T10	20.09.99	03.03.00
	T11	20.09.99	15.05.00
2000-2001	T01	21.10.00	05.05.01
	T03	16.10.00	12.05.01
	T04	17.10.00	10.05.01
	T05	19.10.00	04.05.01
	T06	14.10.00	15.05.01
	T07	21.10.00	09.05.01
	T08	30.10.00	11.05.01
	T09	18.10.00	10.05.01
	T10	30.10.00	05.05.01
	T11	17.10.00	11.05.01
2001-2002	T01	29.12.01	25.03.02
	T03	20.10.01	06.04.02
	T04	29.12.01	25.03.02
	T05	20.10.01	22.03.02
	T06	29.12.01	29.03.02
	T07	29.12.01	25.03.02
	T08	29.12.01	22.03.02
	T09	29.12.01	22.03.02
	T10	20.10.01	23.03.02
	T11	29.12.01	29.03.02
2002-2003	T01	02.11.02	11.03.03
	T03	25.10.02	23.03.03
	T04	03.11.02	10.03.03
	T05	25.10.02	05.03.03

	T06	03.11.02	20.03.03
	T07	02.11.02	07.03.03
	T08	26.10.02	05.03.03
	T09	25.10.02	07.03.03
	T10	09.11.02	07.02.03
	T11	04.11.02	11.03.03
2003-2004	T01	01.12.03	22.03.04
	T03	17.11.03	28.04.04
	T04	26.11.03	30.03.04
	T05	01.12.03	05.03.04
	T06	18.11.03	19.04.04
	T07	27.11.03	04.02.04
	T08	01.12.03	30.01.04
	T09	17.11.03	05.03.04
	T10	29.11.03	05.03.04
	T11	01.12.03	24.03.04
2004-2005	T01	N/A	N/A
	T03	11.01.05	05.02.05
	T04	23.01.05	28.01.05
	T05	23.01.05	24.01.05
	T06	23.01.05	22.02.05
	T07	N/A	N/A
	T08	N/A	N/A
	T09	N/A	N/A
	T10	N/A	N/A
	T11	N/A	N/A
2005-2006	T01	18.02.06	05.04.06
	T03	10.01.06	11.04.06
	T04	30.12.05	07.04.06
	T05	17.01.06	06.04.06
	T06	05.12.05	15.04.06
	T07	19.01.06	04.04.06
	T08	19.02.06	02.04.06
	T09	02.01.06	03.04.06
	T10	18.02.06	05.04.06
	T11	12.01.06	12.04.06
2006-2007	T01	19.11.06	30.03.07
	T03	23.10.06	09.04.07
	T04	24.10.06	04.04.07
	T05	21.10.06	09.03.07
	T06	25.10.06	06.04.07
	T07	24.10.06	16.03.07
	T08	23.11.06	11.03.07

	T09	21.10.06	11.03.07
	T10	20.11.06	09.03.07
	T11	20.11.06	01.04.07
	T01	09.12.07	08.04.08
	T03	04.12.07	15.јун.08
	T04	07.01.08	12.05.08
	T05	08.12.07	13.03.08
2007-2008	T06	05.01.08	14.05.08
	T07	07.01.08	01.04.08
	T08	07.01.08	24.03.08
	T09	10.12.07	11.05.08
	T10	03.12.07	14.03.08
	T11	07.12.07	08.05.08
	T01	18.01.09	24.03.09
	T03	12.11.08	21.03.09
	T04	19.12.08	19.03.09
	T05	12.11.08	09.02.09
2008-2009	T06	12.11.08	27.03.09
	T07	19.01.09	09.03.09
	T08	19.01.09	11.02.09
	T09	11.11.08	09.03.09
	T10	18.01.09	12.02.09
	T11	06.12.08	15.03.09
	T01	07.12.09	08.03.10
	T03	04.12.09	08.03.10
	T04	29.11.09	12.04.10
	T05	29.11.09	02.03.10
2009-2010	T06	28.11.09	09.04.10
	T07	29.11.09	07.03.10
	T08	29.11.09	06.03.10
	T09	28.11.09	08.03.10
	T10	29.11.09	05.03.10
	T11	05.12.09	11.03.10
	T01	19.12.10	20.03.11
	T03	07.12.10	09.03.11
	T04	10.12.10	07.04.11
	T05	14.11.10	02.03.11
2010-2011	T06	08.12.10	16.05.11
	T07	08.12.10	07.03.11
	T08	07.12.10	07.03.11
	T09	14.11.10	07.03.11
	T10	08.12.10	05.03.11
	T11	19.12.10	17.03.11

A2. Drained water flow field T6

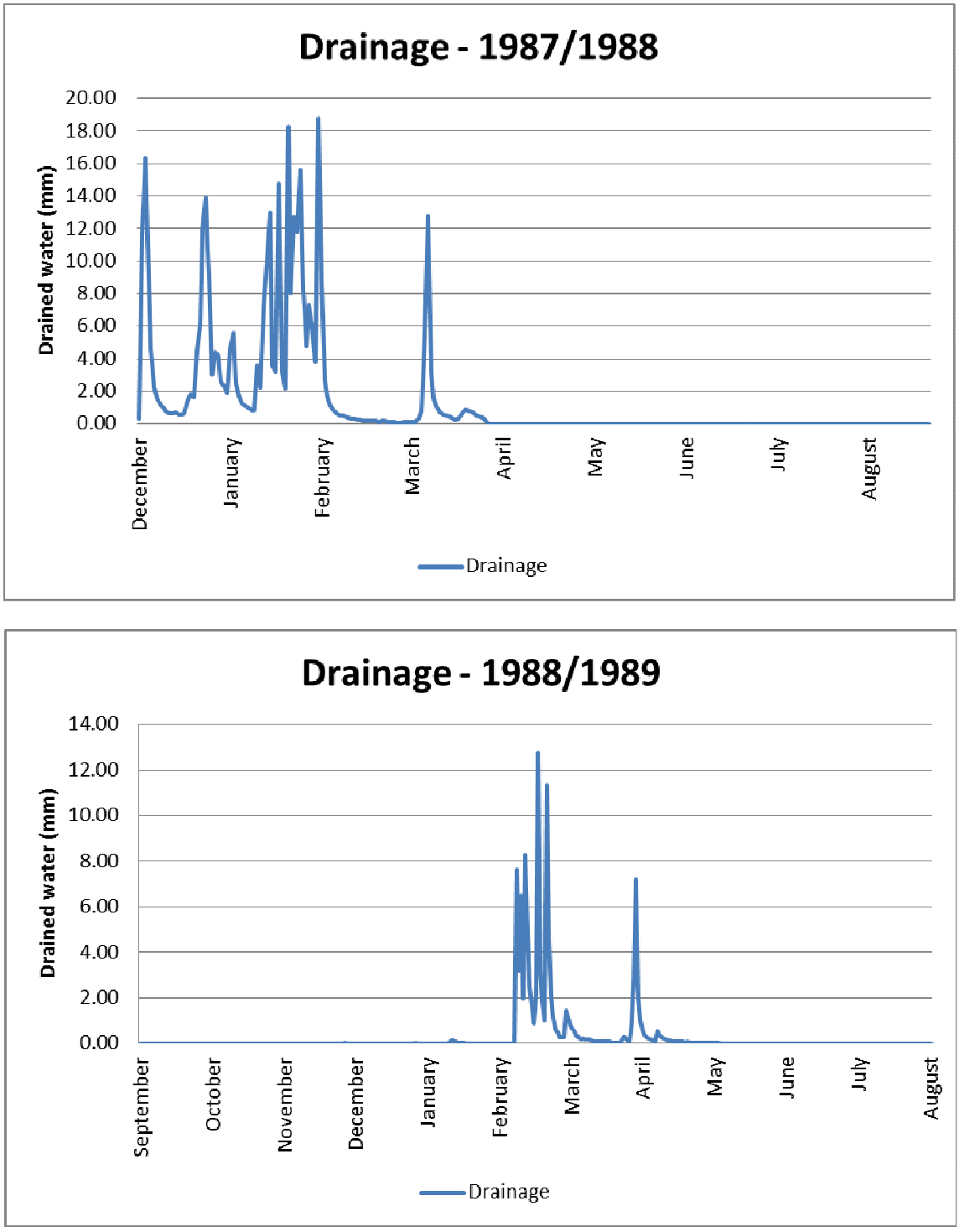


Figure A.2: *Amount of drained water (mm) – Field T6*. Visualization of the actual amount of drained water from field T6 in period 1987–2011

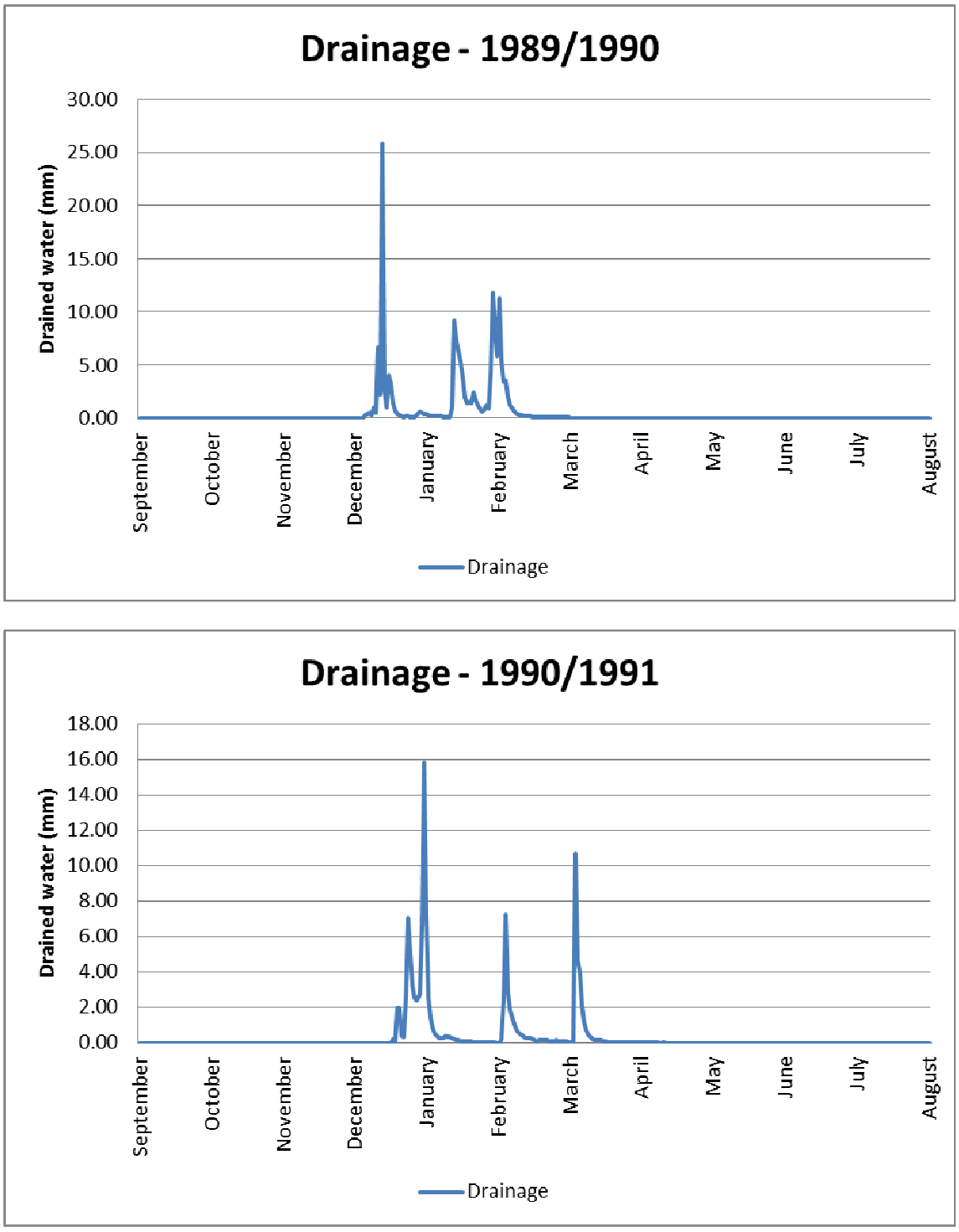


Figure A.2 *(continued)*: *Amount of drained water (mm) – Field T6*

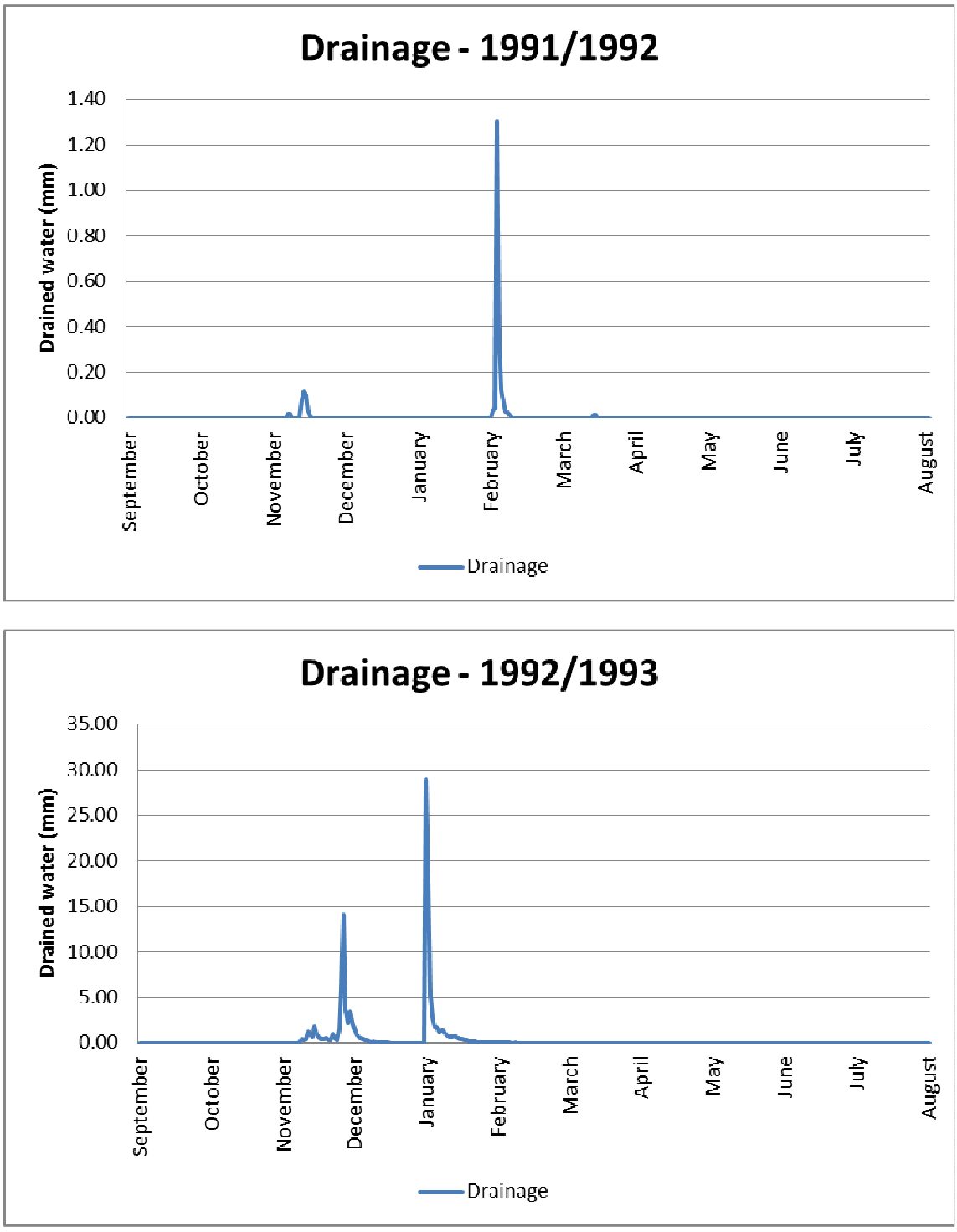


Figure A.2 *(continued)*: *Amount of drained water (mm) – Field T6*

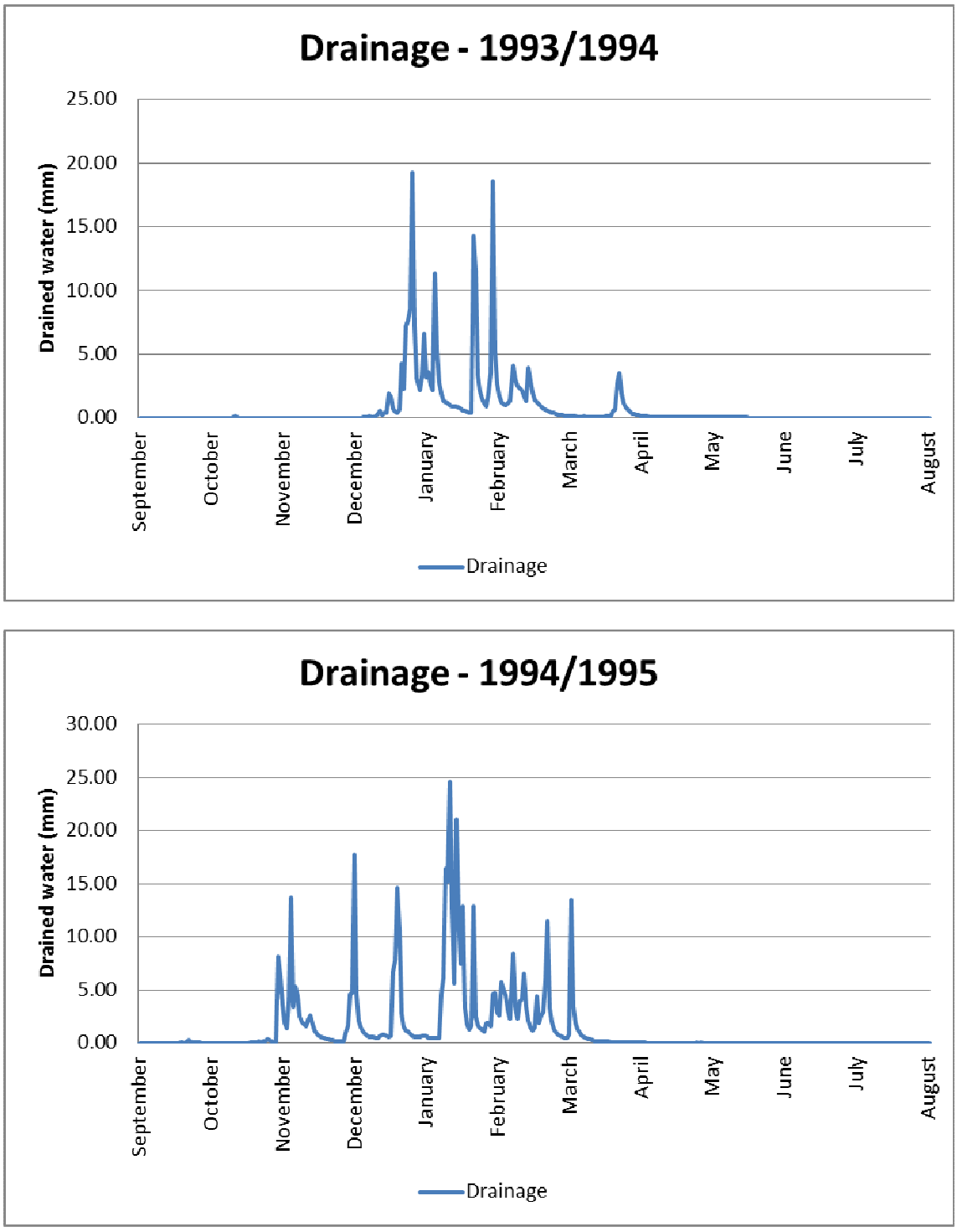
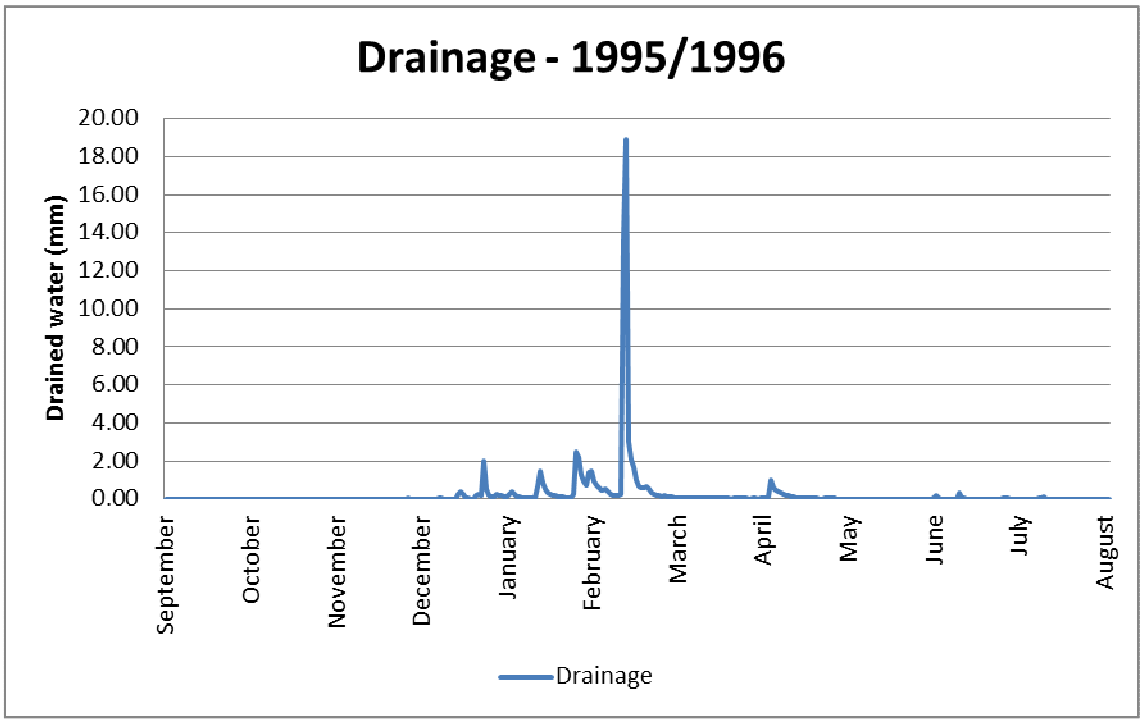


Figure A.2 *(continued)*: *Amount of drained water (mm) – Field T6*



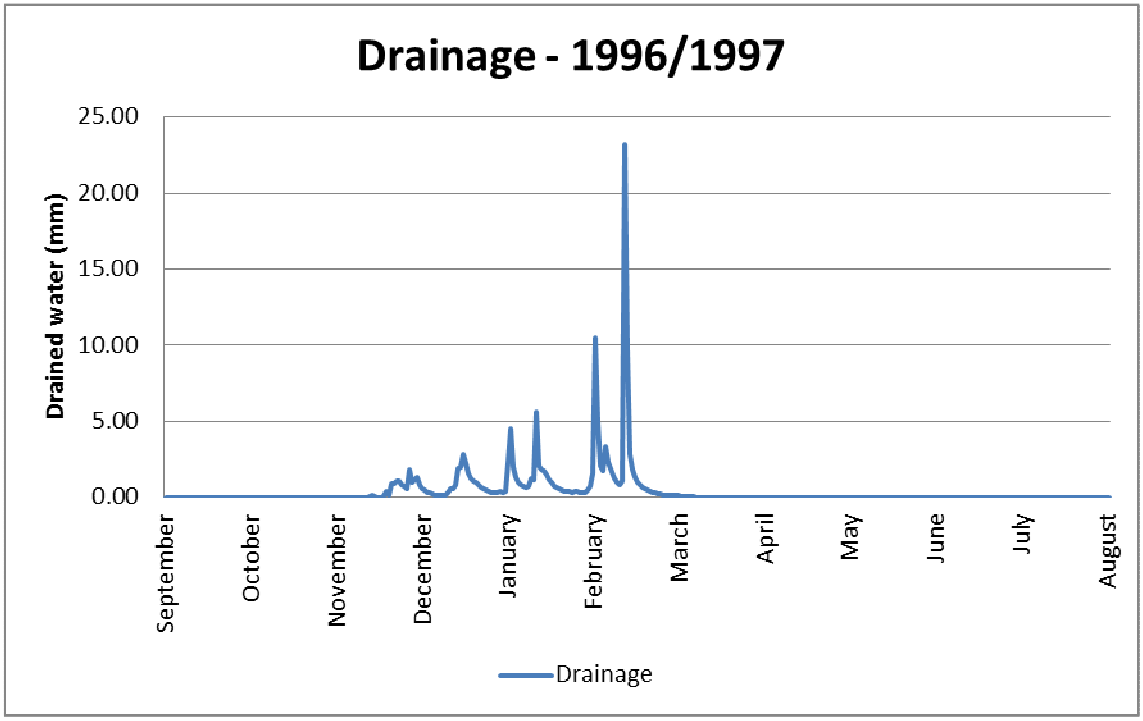


Figure A.2 *(continued)*: *Amount of drained water (mm) – Field T6*

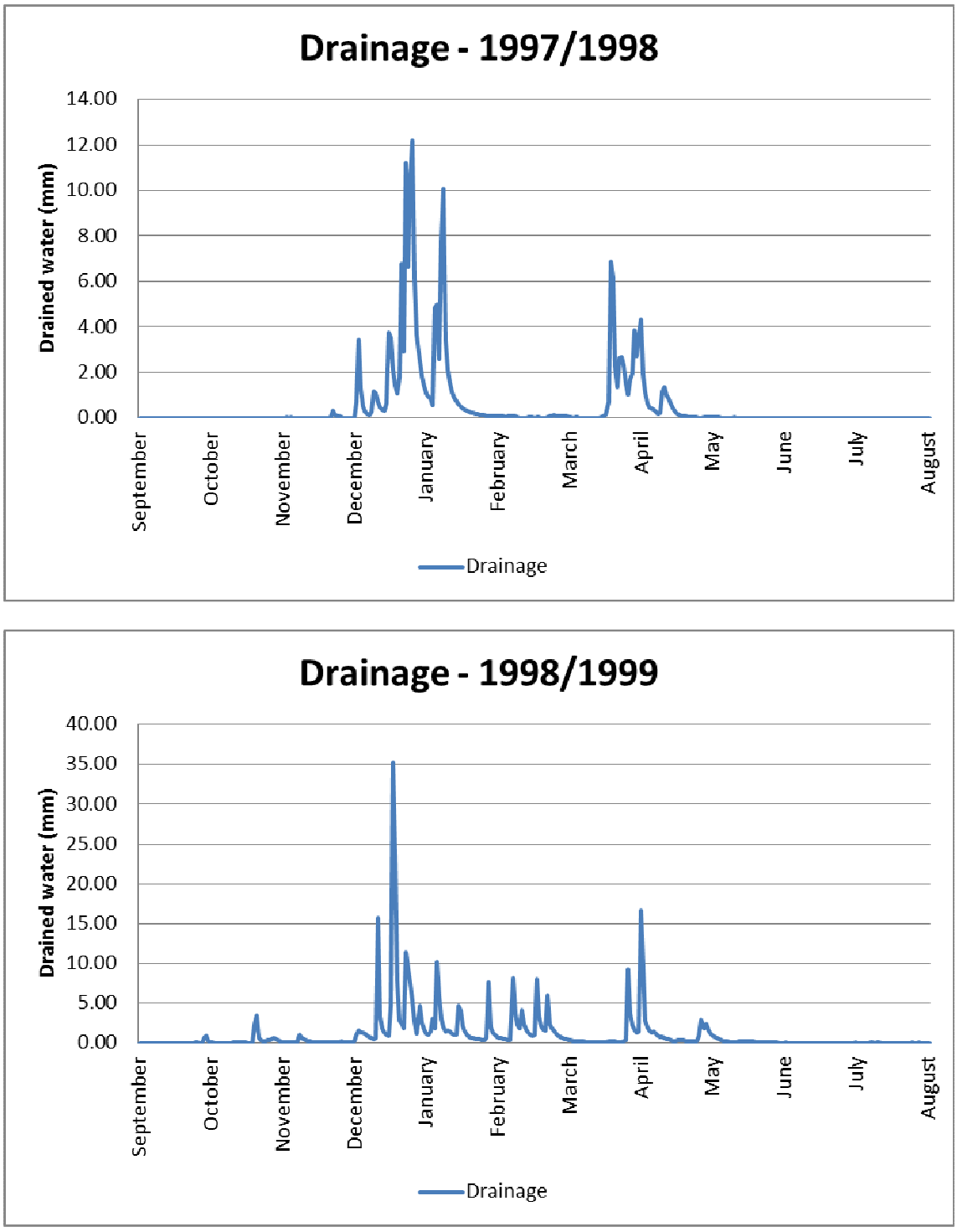
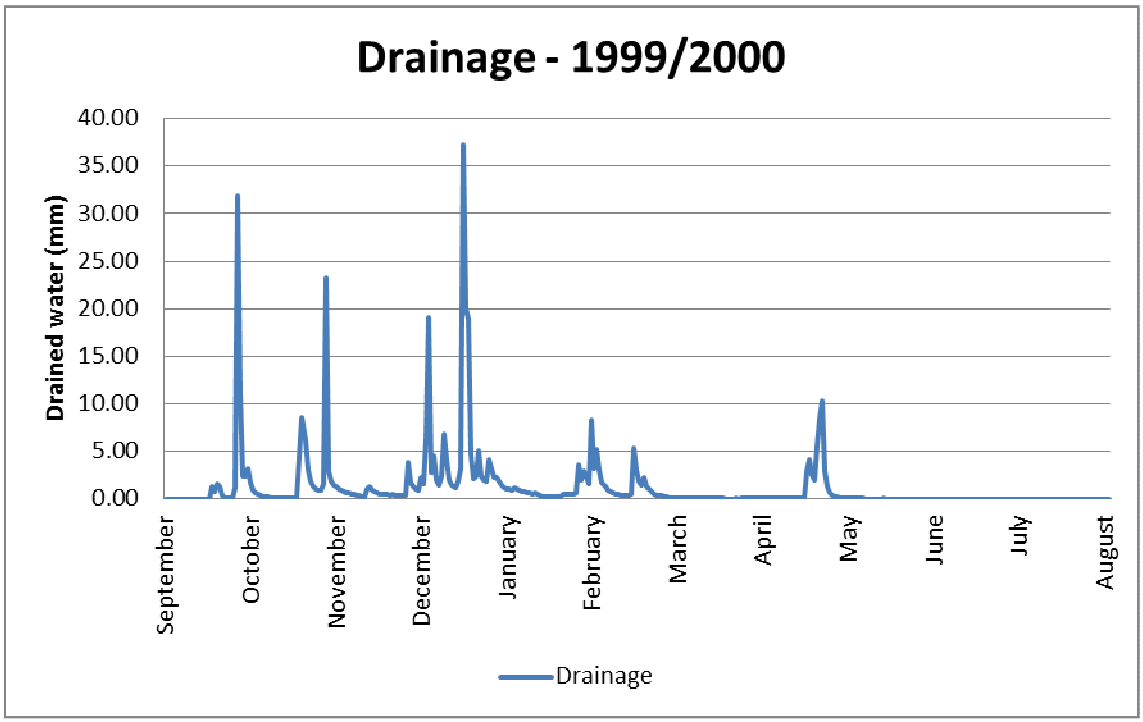


Figure A.2 *(continued)*: *Amount of drained water (mm) – Field T6*



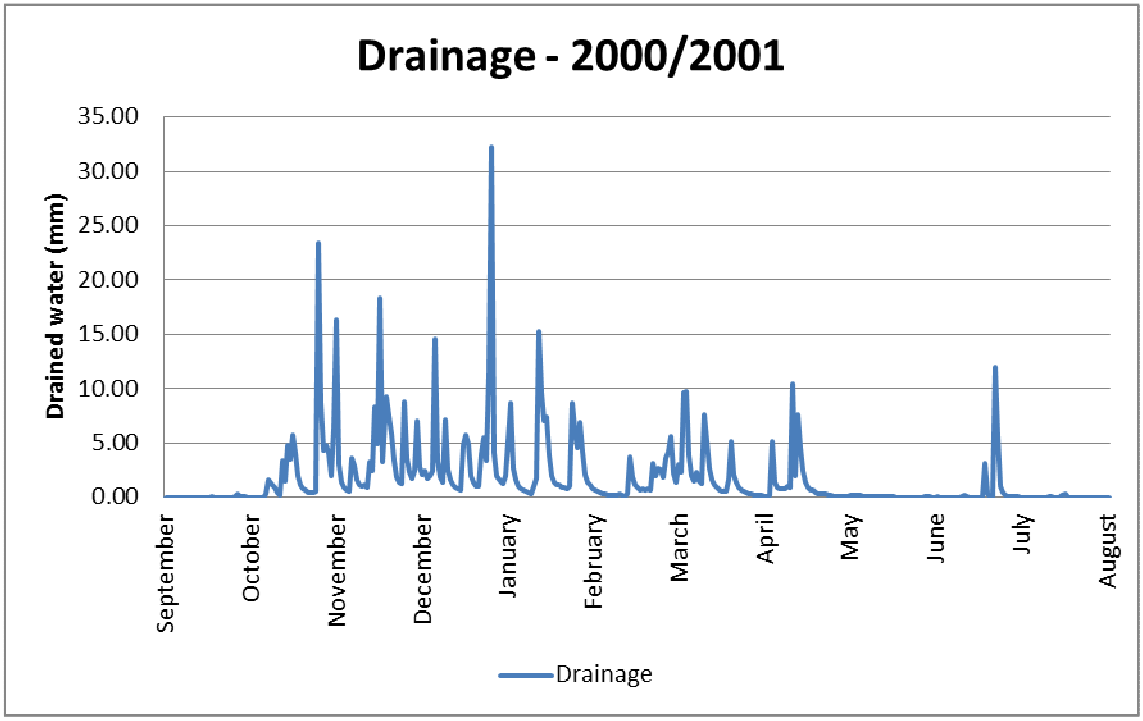


Figure A.2 *(continued)*: *Amount of drained water (mm) – Field T6*

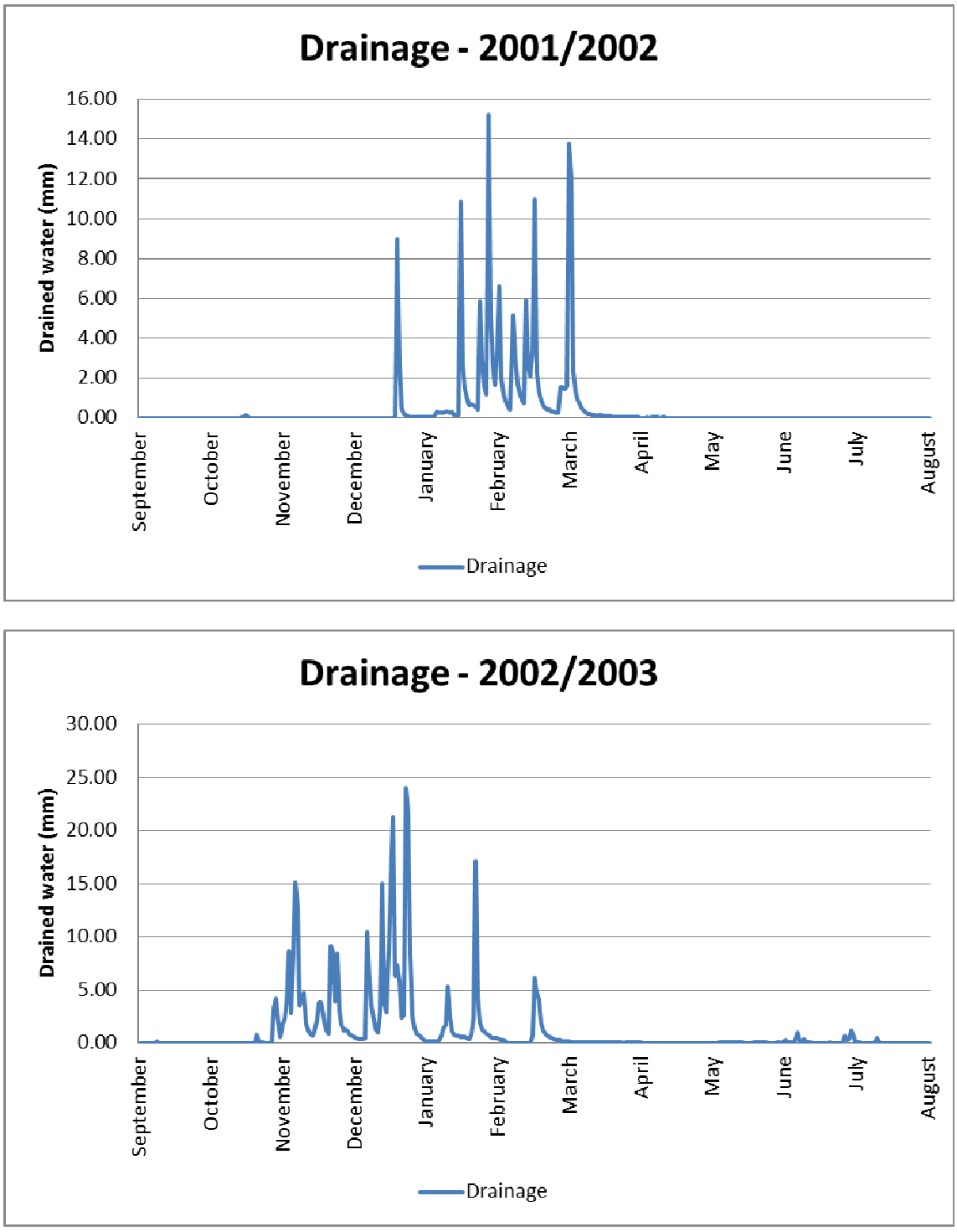


Figure A.2 *(continued)*: *Amount of drained water (mm) – Field T6*

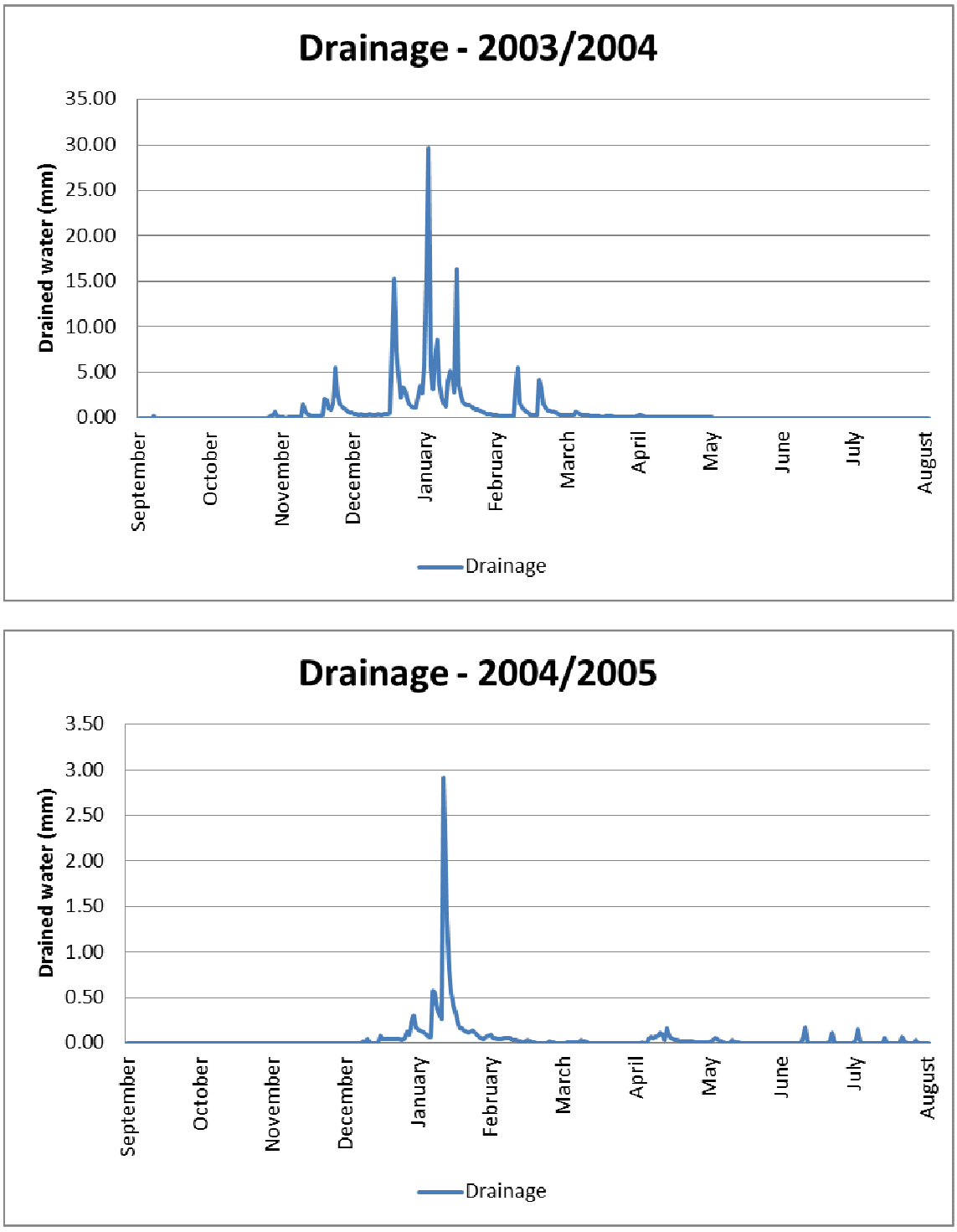


Figure A.2 *(continued)*: *Amount of drained water (mm) – Field T6*

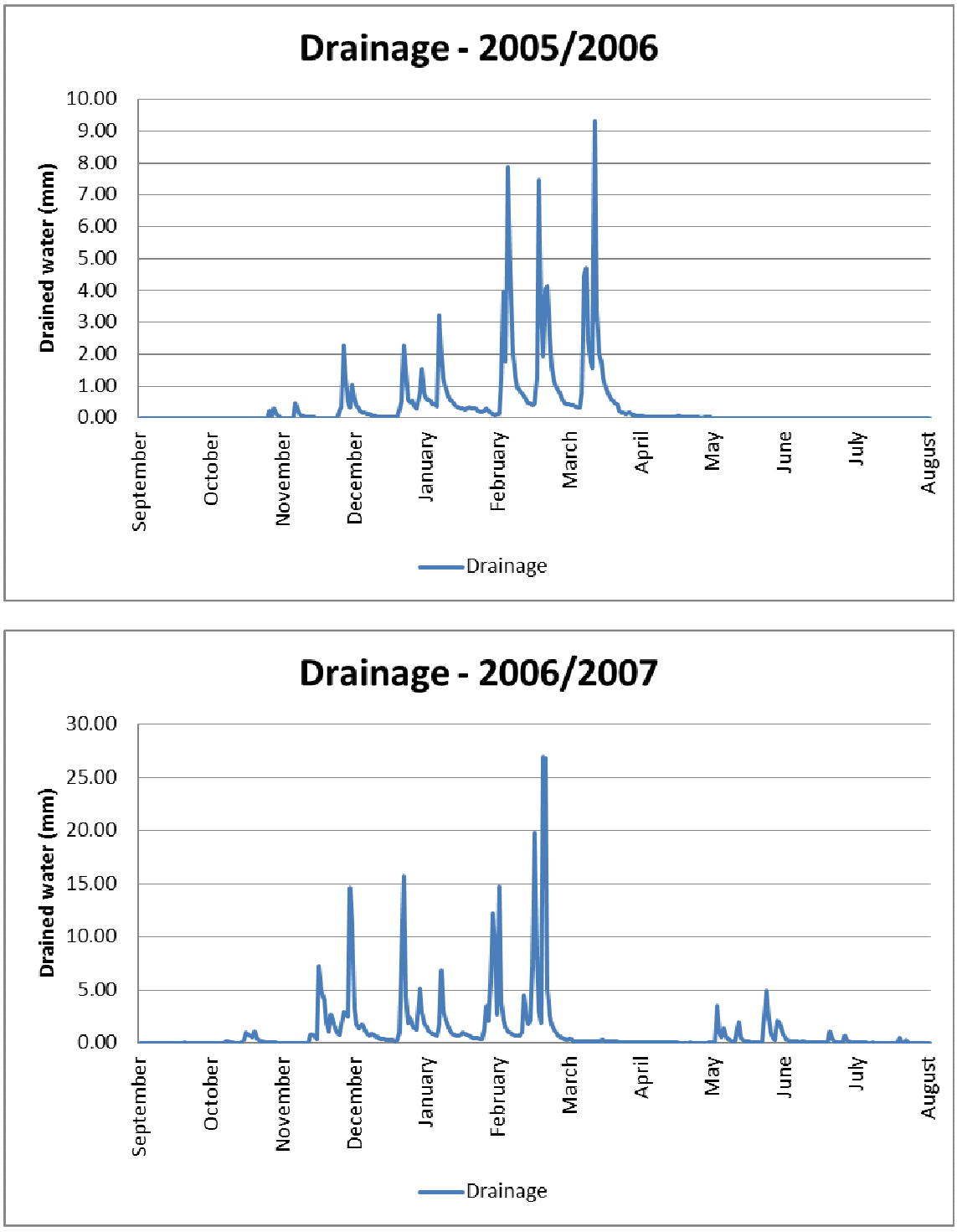


Figure A.2 *(continued)*: *Amount of drained water (mm) – Field T6*

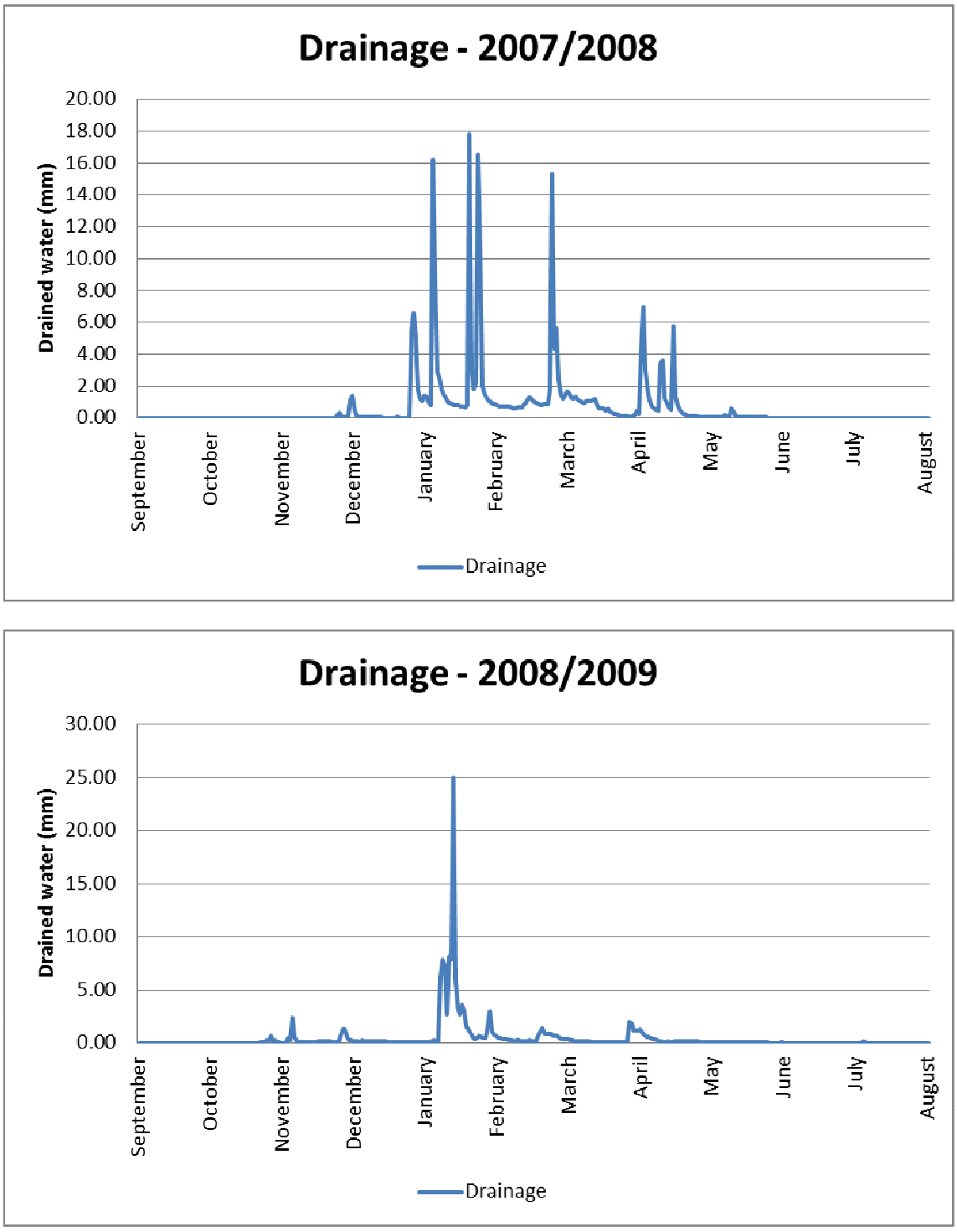


Figure A.2 *(continued)*: *Amount of drained water (mm) – Field T6*

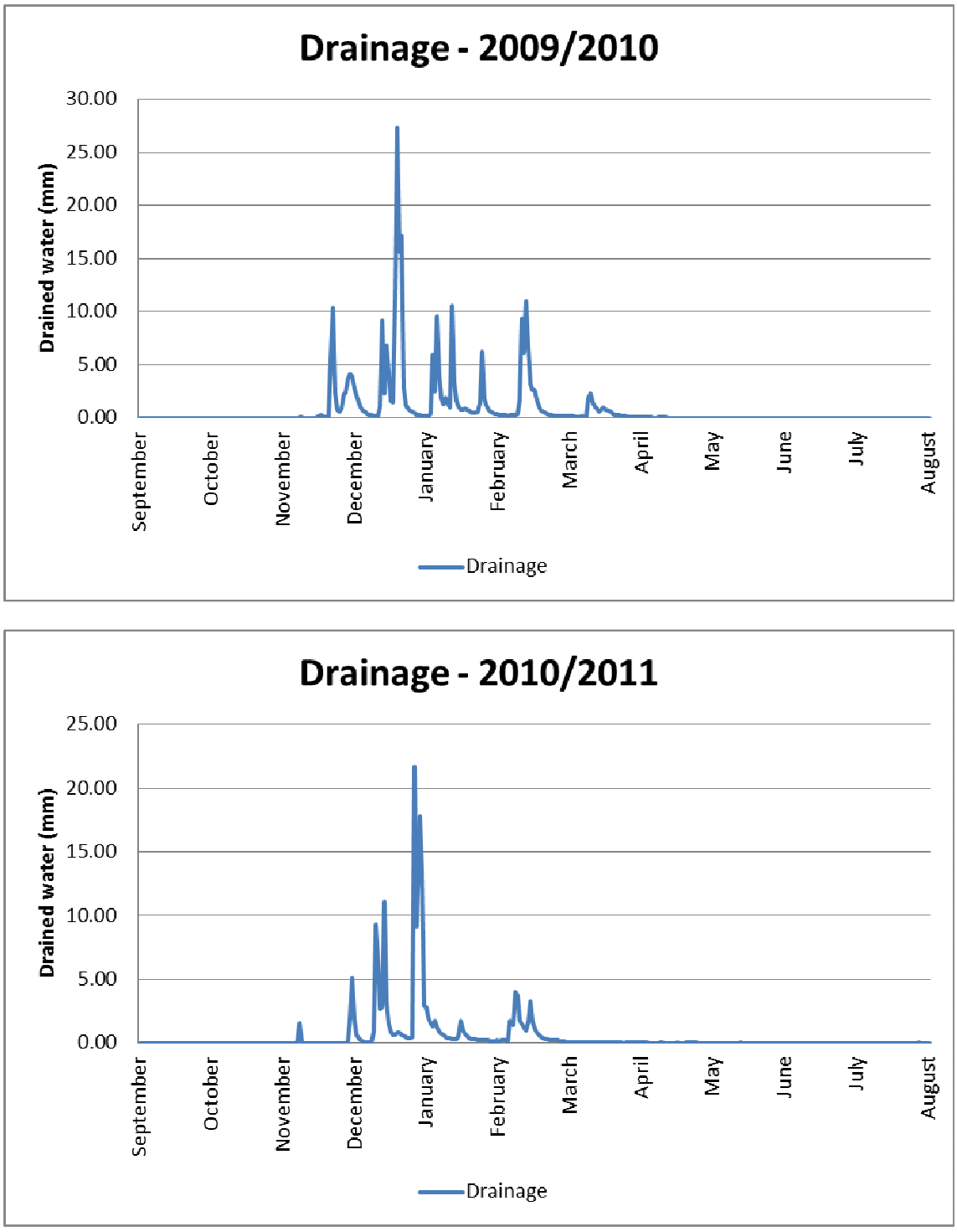
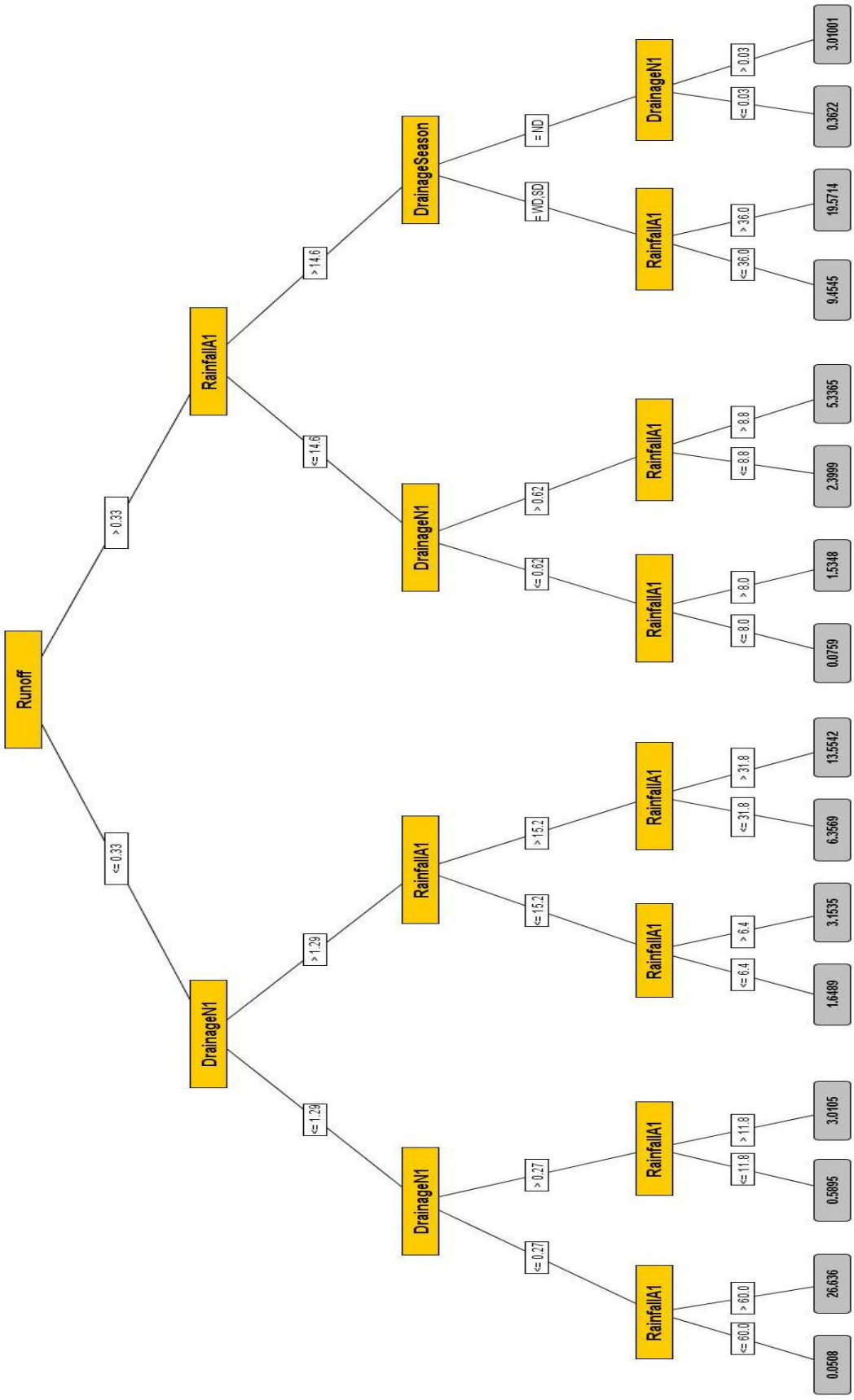


Figure A.2 *(continued)*: *Amount of drained water (mm) – Field T6*

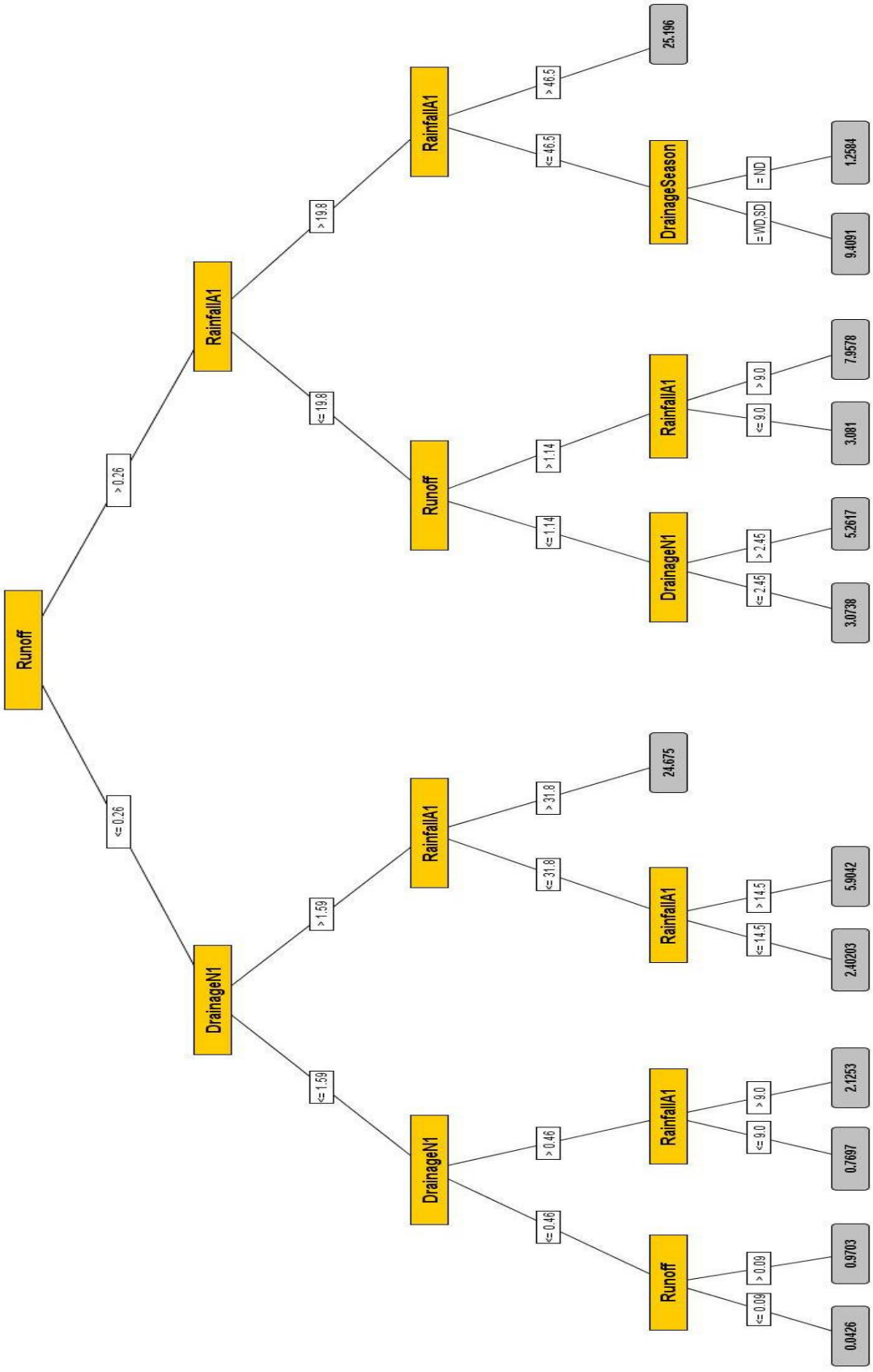
Appendix B. The models learned by machine learning

In this section we present the models learned with machine learning and data mining methods. In following figures are presented learned regression and model trees for All fields, only field T3, and only field T6. The models listed in previous sections are not included in this appendix.



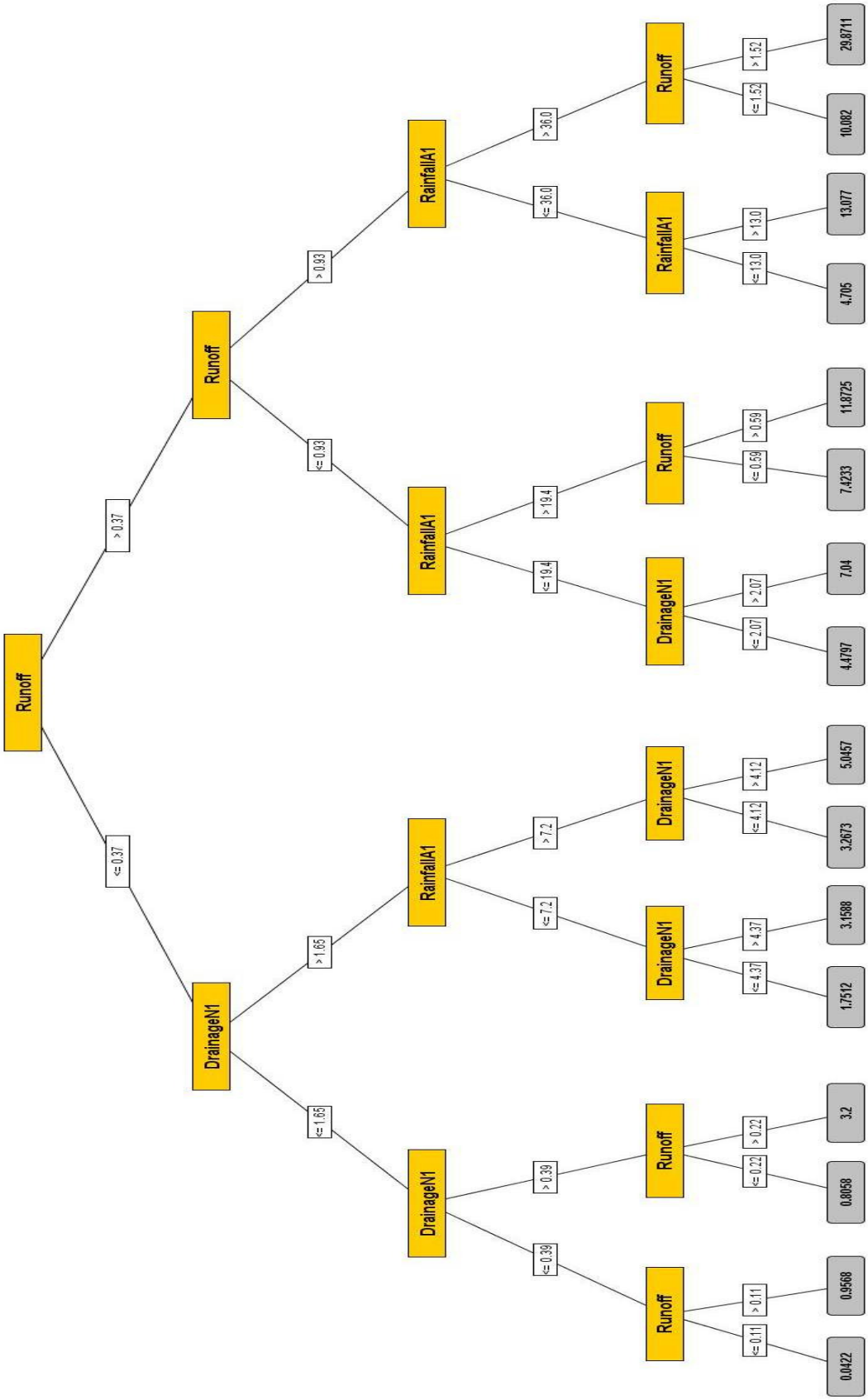
(a)

Figure B.1: *Campaign based predictive models – Regression trees*. The models have been learned on data from all fields (a), only field T3 (b), and only field T6 (c), respectively.



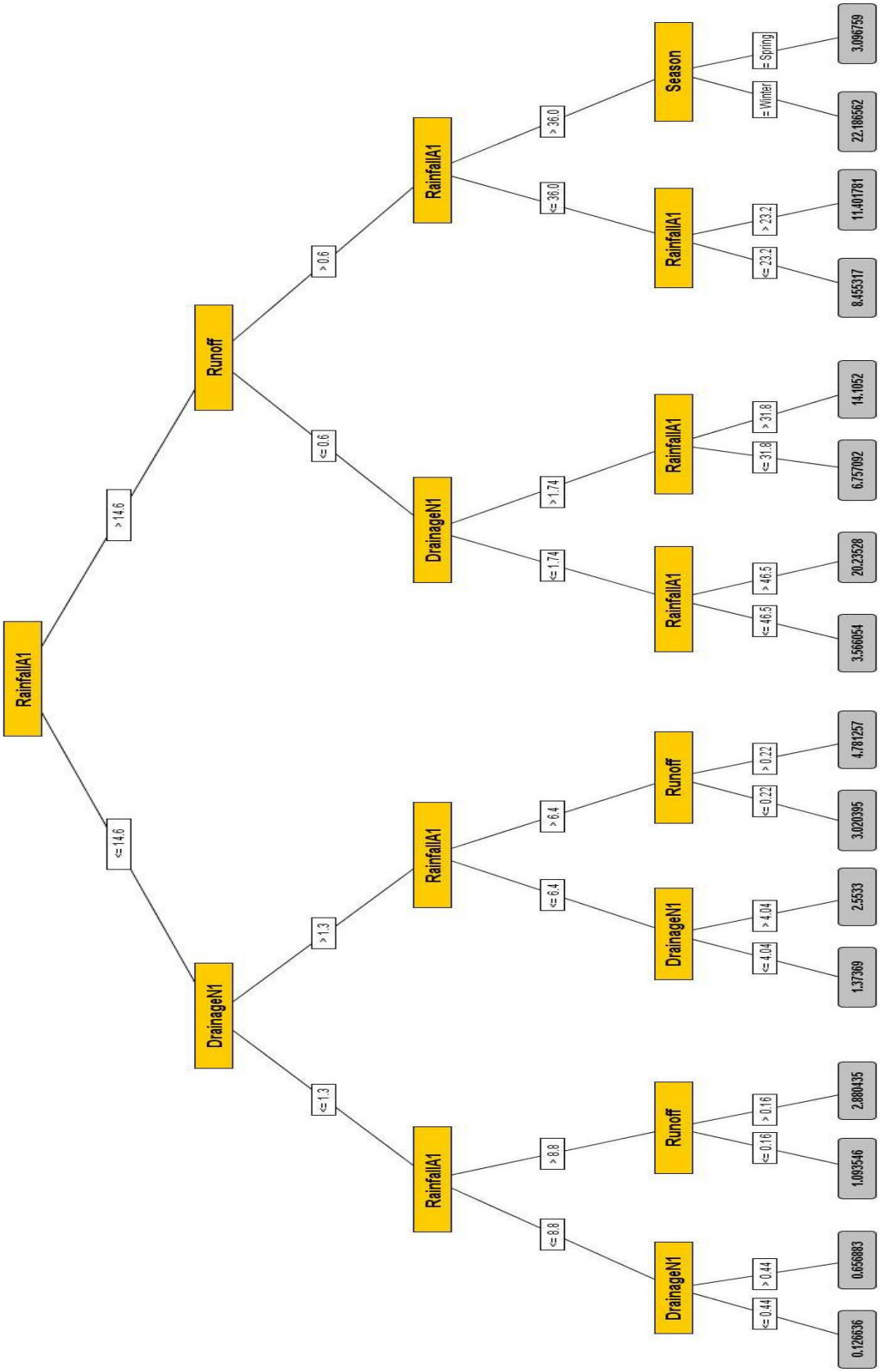
(b)

Figure B.1 *(continued)*: *Campaign based predictive models – Regression trees*.



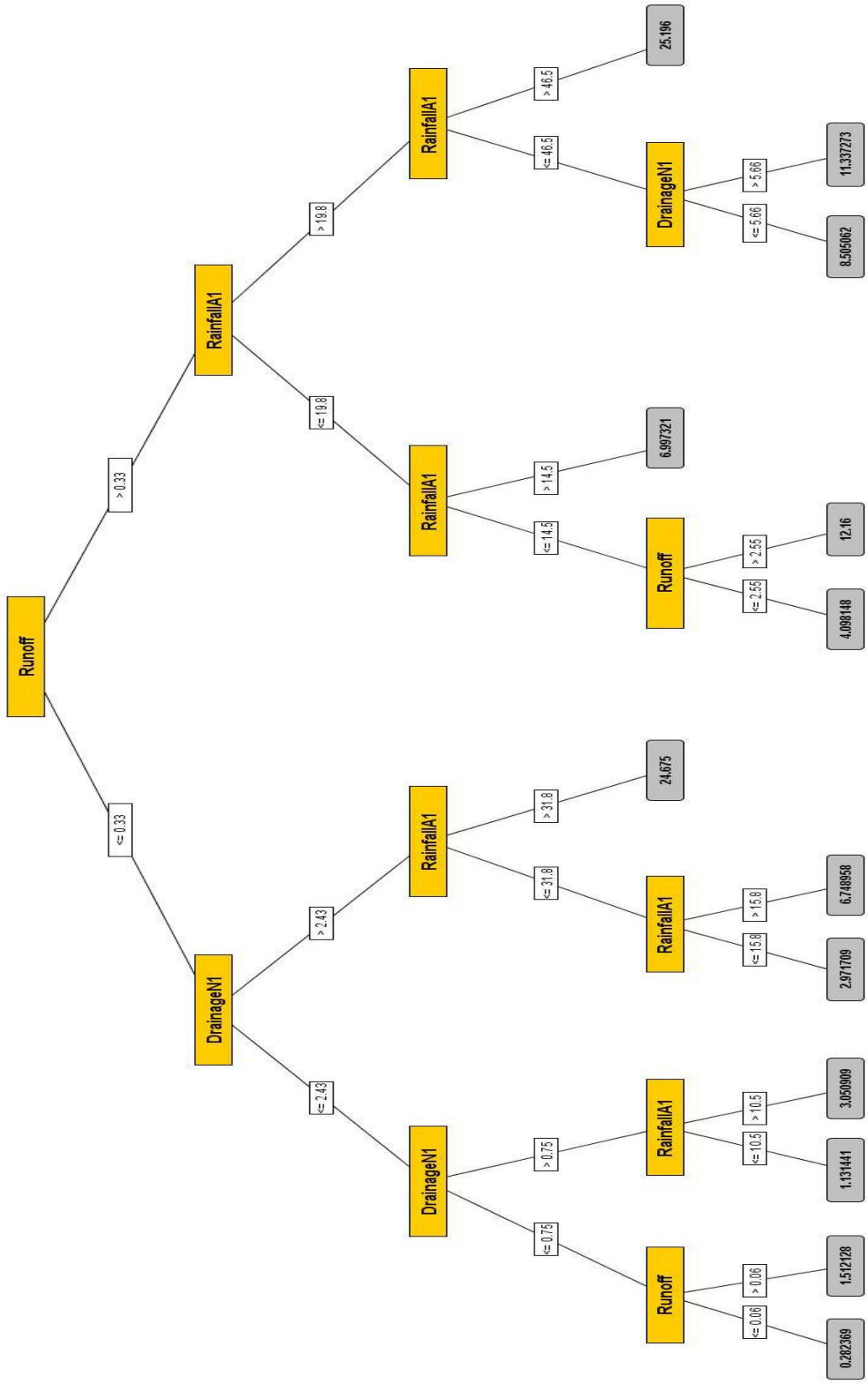
(c)

Figure B.1 *(continued)*: *Campaign based predictive models – Regression trees.*



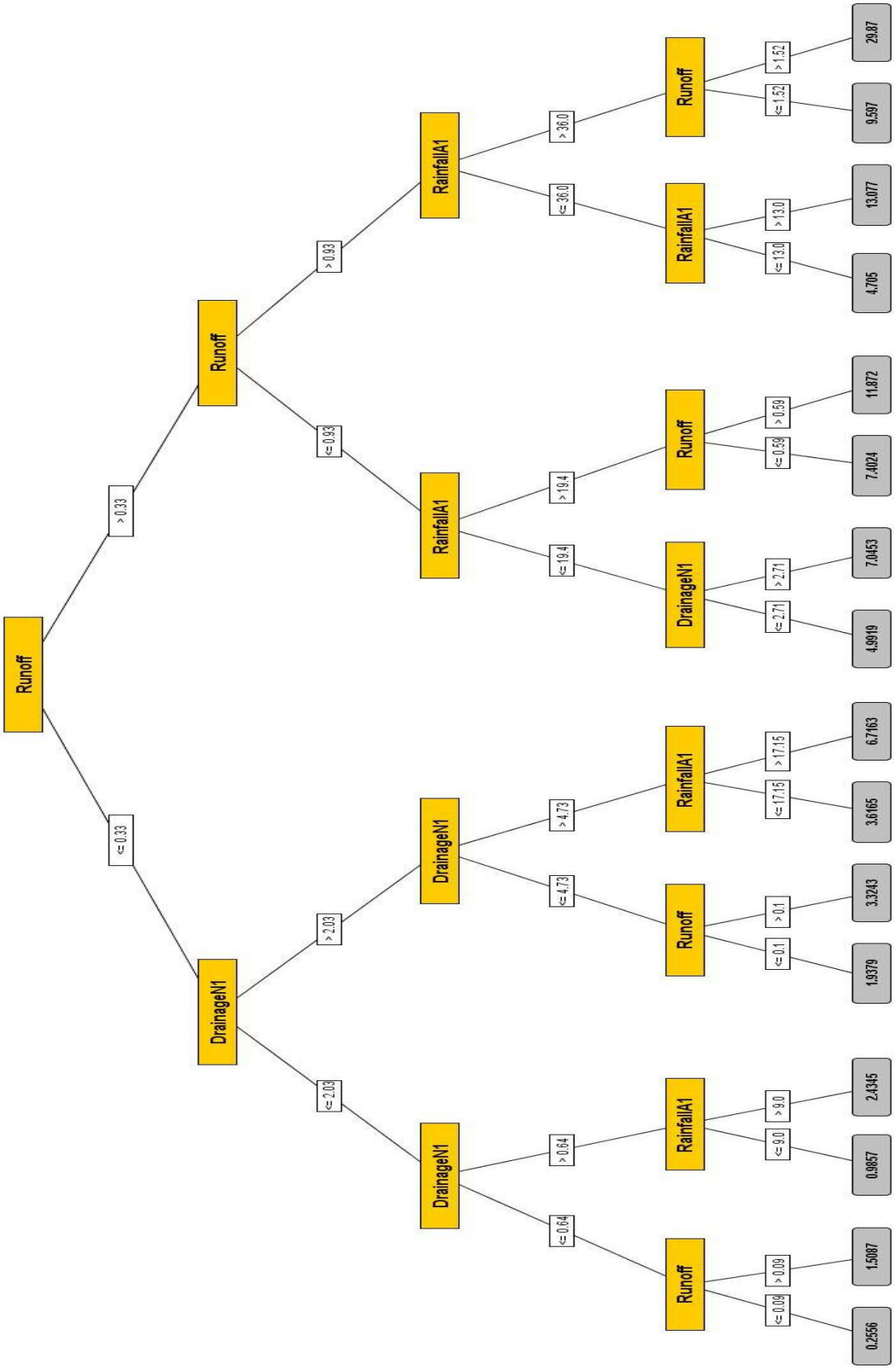
(a)

Figure B.2: *Drainage season based predictive models – Regression trees*. The models have been learned on data from all fields (a), only field T3 (b), and only field T6 (c), respectively.



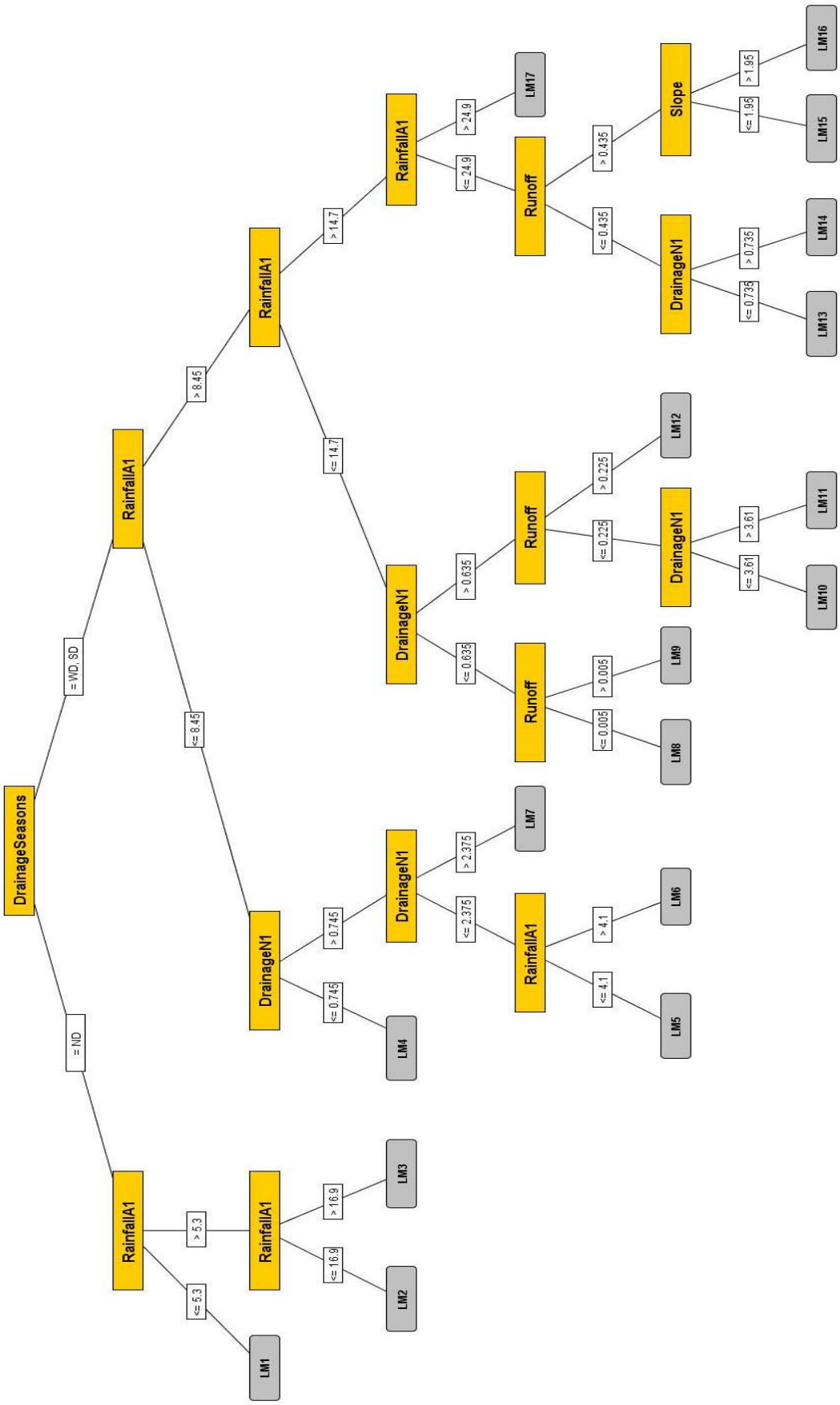
(b)

Figure B.2 *(continued)*: *Drainage season based predictive models – Regression trees*.



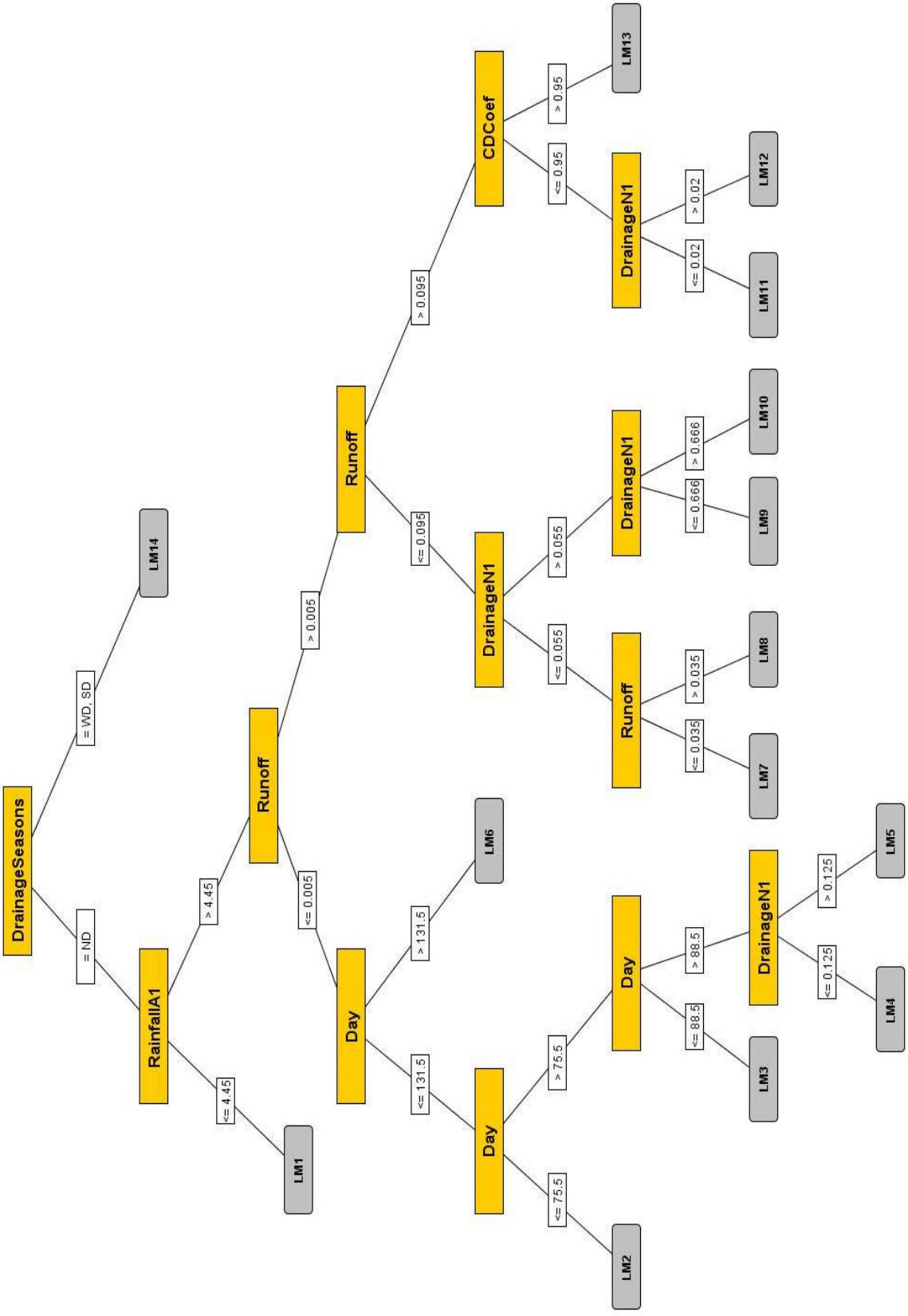
(c)

Figure B.2 *(continued)*: *Drainage season based predictive models – Regression trees*.



(a)

Figure B.3: *Campaign based predictive models – Model trees*. The models have been learned on data from all fields (a), only field T3 (b), respectively.



(b)

Figure B.3 *(continued)*: *Campaign based predictive models – Model trees*.

The following are linear models that have been built as part of model tree learned from data of all fields (Figure B.3 (a)):

LM1

Drainage =

0.0002 * *DrainageSeason*=WD
+ 0.0011 * *Runoff*
+ 0.0003 * *DrainageN1*
+ 0.0082

LM2

Drainage =

– 0.0001 * *Day*
– 0.0271 * *CDCoef*
– 0.0029 * *Temp*
+ 0.0038 * *RainfallA1*
+ 0 * *DrainageSeason*=SD,WD
+ 0.0002 * *DrainageSeason*=WD
+ 0.8776 * *Runoff*
+ 0.4598 * *DrainageN1*
+ 0.0581

LM3

Drainage =

0.1607 * *CDCoef*
– 0.0243 * *Temp*
+ 0.0143 * *RainfallA1*
+ 0 * *DrainageSeason*=SD,WD
+ 0.0002 * *DrainageSeason*=WD
+ 1.7458 * *Runoff*
+ 0.6303 * *DrainageN1*
– 0.0248

LM4

Drainage =

– 0.0001 * *Day*
+ 0.0002 * *Crop*=Wheat,Winter_horse_bean,Winter_peas,Rgi
+ 0.0193 * *RainfallA1*
+ 0.0001 * *Slope*
+ 0.0001 * *DrainageSeason*=SD,WD
+ 0.0011 * *DrainageSeason*=WD
+ 0.7271 * *Runoff*
+ 0.8432 * *DrainageN1*
– 0.0215

LM5

Drainage =

+ 0.0008 * *Crop*=Wheat,Winter_horse_bean,Winter_peas,Rgi
+ 0.0481 * *RainfallA1*
+ 0.0003 * *Slope*
+ 0.0001 * *DrainageSeason*=SD,WD
+ 0.0011 * *DrainageSeason*=WD
+ 0.012 * *Runoff*
+ 0.4885 * *DrainageN1*
+ 0.2036

LM6

Drainage =

+ 0.0008 * *Crop*=Wheat,Winter_horse_bean,Winter_peas,Rgi
+ 0.1759 * *RainfallA1*
+ 0.0003 * *Slope*
+ 0.0001 * *DrainageSeason*=SD,WD
+ 0.0011 * *DrainageSeason*=WD
+ 0.0195 * *Runoff*
+ 0.6544 * *DrainageN1*
– 0.5662

LM7

Drainage =

+ 0.2241 * *Crop*=Wheat,Winter_horse_bean,Winter_peas,Rgi
+ 0.0129 * *Temp*
+ 0.174 * *RainfallA1*
+ 0.0967 * *Slope*
+ 0.0001 * *DrainageSeason*=SD,WD
+ 0.0011 * *DrainageSeason*=WD
+ 0.0119 * *Runoff*
+ 0.1263 * *DrainageN1*
+ 0.6724

LM8

Drainage =

+ 0.0003 * *Crop*=Wheat,Winter_horse_bean,Winter_peas,Rgi
+ 0.0099 * *RainfallA1*
– 0.001 * *Slope*
+ 0.0001 * *DrainageSeason*=SD,WD
+ 0.0136 * *DrainageSeason*=WD
+ 0.1666 * *Runoff*
+ 0.0069 * *DrainageN1*
+ 0.2937

LM9

Drainage =

+ 0.0003 * *Crop*=Wheat,Winter_horse_bean,Winter_peas,Rgi
+ 0.1624 * *RainfallA1*
– 0.001 * *Slope*
+ 0.0001 * *DrainageSeason*=SD,WD
+ 0.0136 * *DrainageSeason*=WD
+ 2.5947 * *Runoff*
+ 0.0069 * *DrainageN1*
– 0.856

LM10

Drainage =

+ 0.0003 * *Crop*=Wheat,Winter_horse_bean,Winter_peas,Rgi
+ 0.2093 * *RainfallA1*
– 0.001 * *Slope*
+ 0.0001 * *DrainageSeason*=SD,WD
+ 0.0136 * *DrainageSeason*=WD
+ 0.0499 * *Runoff*
+ 0.438 * *DrainageN1*
– 0.6719

LM11

Drainage =

+ 0.0003 * *Crop*=Wheat,Winter_horse_bean,Winter_peas,Rgi
+ 0.0115 * *RainfallA1*
– 0.001 * *Slope*
+ 0.0001 * *DrainageSeason*=SD,WD
+ 0.0136 * *DrainageSeason*=WD
+ 0.0499 * *Runoff*
+ 0.0145 * *DrainageN1*
+ 3.5582

LM12

Drainage =

+ 0.0003 * *Crop*=Wheat,Winter_horse_bean,Winter_peas,Rgi
+ 0.0101 * *RainfallA1*
– 0.001 * *Slope*
+ 0.0001 * *DrainageSeason*=SD,WD
+ 0.0136 * *DrainageSeason*=WD
+ 0.0785 * *Runoff*
+ 0.2112 * *DrainageN1*
+ 3.9982

LM13

Drainage =

+ 0.0003 * *Crop*=Wheat,Winter_horse_bean,Winter_peas,Rgi
+ 0.0102 * *RainfallA1*
– 0.0066 * *Slope*
+ 0.0001 * *DrainageSeason*=SD,WD
+ 1.7618 * *DrainageSeason*=WD
+ 0.0381 * *Runoff*
+ 0.0157 * *DrainageN1*
+ 0.8439

LM14

Drainage =

+ 0.0003 * *Crop*=Wheat,Winter_horse_bean,Winter_peas,Rgi
+ 0.1839 * *RainfallA1*
– 0.0066 * *Slope*
+ 0.0001 * *DrainageSeason*=SD,WD
+ 0.1152 * *DrainageSeason*=WD
+ 0.0381 * *Runoff*
+ 0.1847 * *DrainageN1*
+ 0.9917

LM15

Drainage =

+ 0.0003 * *Crop*=Wheat,Winter_horse_bean,Winter_peas,Rgi
+ 0.0088 * *RainfallA1*
– 0.0743 * *Slope*
+ 0.0001 * *DrainageSeason*=SD,WD
+ 0.1275 * *DrainageSeason*=WD
+ 0.0628 * *Runoff*
+ 0.0128 * *DrainageN1*
+ 9.7515

LM16

Drainage =

+ 0.0003 * *Crop*=Wheat,Winter_horse_bean,Winter_peas,Rgi
+ 0.0088 * *RainfallA1*
– 0.034 * *Slope*
+ 0.0001 * *DrainageSeason*=SD,WD
+ 0.1275 * *DrainageSeason*=WD
+ 0.0628 * *Runoff*
+ 0.0128 * *DrainageN1*
+ 8.1495

LM17

Drainage =

+ 0.0003 * *Crop*=Wheat,Winter_horse_bean,Winter_peas,Rgi
+ 0.3514 * *RainfallA1*
– 0.006 * *Slope*
+ 0.0001 * *DrainageSeason*=SD,WD
+ 0.0723 * *DrainageSeason*=WD
+ 1.2498 * *Runoff*
+ 0.3046 * *DrainageN1*
– 3.6868

The following are linear models that have been built as part of model tree learned from data of field T3 (Figure B.3 (b)):

LM1

Drainage =

0.0006 * *Season*=Spring,AW
+ 0.0014 * *Season*=AW
– 0.0002 * *Crop*=CIPAN,Rgi
+ 0.0011 * *Crop*=Rgi
+ 0.0001 * *Temp*
+ 0.0002 * *RainfallA1*
+ 0.0011 * *DrainageSeason*=SD,WD
+ 0.0008 * *DrainageSeason*=WD
+ 0.0091 * *Runoff*
+ 0.0029 * *DrainageN1*
+ 0.007

LM2

Drainage =

0.0021 * *Season*=Spring,AW
+ 0.0054 * *Season*=AW
+ 0.0001 * *Day*
– 0.0002 * *Crop*=CIPAN,Rgi
+ 0.0011 * *Crop*=Rgi
+ 0.0001 * *Temp*
+ 0.0004 * *RainfallA1*
+ 0.0011 * *DrainageSeason*=SD,WD
+ 0.0008 * *DrainageSeason*=WD
+ 0.0352 * *Runoff*
+ 0.042 * *DrainageN1*
- 0.0025

LM3

Drainage =
0.0021 * *Season*=Spring,AW
+ 0.0054 * *Season*=AW
+ 0.001 * *Day*
– 0.0002 * *Crop*=CIPAN,Rgi
+ 0.0011 * *Crop*=Rgi
+ 0.0001 * *Temp*
+ 0.0006 * *RainfallA1*
+ 0.0011 * *DrainageSeason*=SD,WD
+ 0.0008 * *DrainageSeason*=WD
+ 0.0352 * *Runoff*
+ 0.2359 * *DrainageN1*
– 0.0855

LM4

Drainage =
0.0021 * *Season*=Spring,AW
+ 0.0054 * *Season*=AW
+ 0.0009 * *Day*
– 0.0002 * *Crop*=CIPAN,Rgi
+ 0.0011 * *Crop*=Rgi
+ 0.0001 * *Temp*
+ 0.0006 * *RainfallA1*
+ 0.0011 * *DrainageSeason*=SD,WD
+ 0.0008 * *DrainageSeason*=WD
+ 0.0352 * *Runoff*
+ 0.3487 * *DrainageN1*
– 0.0575

LM5

Drainage =
0.0021 * *Season*=Spring,AW
+ 0.0054 * *Season*=AW
+ 0.0009 * *Day*
– 0.0002 * *Crop*=CIPAN,Rgi
+ 0.0011 * *Crop*=Rgi
+ 0.0001 * *Temp*
+ 0.0006 * *RainfallA1*
+ 0.0011 * *DrainageSeason*=SD,WD
+ 0.0008 * *DrainageSeason*=WD
+ 0.0352 * *Runoff*
+ 0.49 * *DrainageN1*
+ 0.1518

LM6

Drainage =

0.0021 * *Season*=Spring,AW
+ 0.0054 * *Season*=AW
– 0.0002 * *Crop*=CIPAN,Rgi
+ 0.0011 * *Crop*=Rgi
+ 0.0001 * *Temp*
+ 0.0003 * *RainfallA1*
+ 0.0011 * *DrainageSeason*=SD,WD
+ 0.0008 * *DrainageSeason*=WD
+ 0.0352 * *Runoff*
+ 0.0251 * *DrainageN1*
– 0.0004

LM7

Drainage =

0.0234 * *Season*=Spring,AW
+ 0.0656 * *Season*=AW
+ 0.0002 * *Day*
– 0.0002 * *Crop*=CIPAN,Rgi
+ 0.0011 * *Crop*=Rgi
– 0.0004 * *Temp*
+ 0.0012 * *RainfallA1*
+ 0.0011 * *DrainageSeason*=SD,WD
+ 0.0008 * *DrainageSeason*=WD
+ 0.4674 * *Runoff*
+ 0.2514 * *DrainageN1*
– 0.0911

LM8

Drainage =

0.0234 * *Season*=Spring,AW
+ 0.0656 * *Season*=AW
+ 0.0002 * *Day*
– 0.0002 * *Crop*=CIPAN,Rgi
+ 0.0011 * *Crop*=Rgi
– 0.0008 * *Temp*
+ 0.0012 * *RainfallA1*
+ 0.0011 * *DrainageSeason*=SD,WD
+ 0.0008 * *DrainageSeason*=WD
+ 0.5748 * *Runoff*
+ 0.3382 * *DrainageN1*
– 0.0366

LM9

Drainage =

0.0234 * *Season*=Spring,AW
+ 0.0656 * *Season*=AW
– 0.0005 * *Day*
– 0.0002 * *Crop*=CIPAN,Rgi
+ 0.0011 * *Crop*=Rgi
+ 0.0001 * *Temp*
+ 0.0012 * *RainfallA1*
+ 0.0011 * *DrainageSeason*=SD,WD
+ 0.0008 * *DrainageSeason*=WD
+ 2.6787 * *Runoff*
+ 0.6299 * *DrainageN1*
+ 0.093

LM10

Drainage =

0.0234 * *Season*=Spring,AW
+ 0.0656 * *Season*=AW
– 0.0005 * *Day*
– 0.0002 * *Crop*=CIPAN,Rgi
+ 0.0011 * *Crop*=Rgi
+ 0.0001 * *Temp*
+ 0.0289 * *RainfallA1*
+ 0.0011 * *DrainageSeason*=SD,WD
+ 0.0008 * *DrainageSeason*=WD
+ 4.38 * *Runoff*
+ 0.4328 * *DrainageN1*
+ 0.1469

LM11

Drainage =

0.0624 * *Season*=Spring,AW
+ 0.9494 * *Season*=AW
+ 0.0024 * *Day*
– 0.0002 * *Crop*=CIPAN,Rgi
+ 0.0011 * *Crop*=Rgi
+ 0.0001 * *Temp*
+ 0.0179 * *RainfallA1*
+ 0.0011 * *DrainageSeason*=SD,WD
+ 0.0008 * *DrainageSeason*=WD
+ 0.7167 * *Runoff*
+ 0.9865 * *DrainageN1*
– 1.1612

LM12

Drainage =

0.0624 * *Season*=Spring,AW
+ 0.7086 * *Season*=AW
+ 0.005 * *Day*
– 0.0002 * *Crop*=CIPAN,Rgi
+ 0.0011 * *Crop*=Rgi
+ 0.0001 * *Temp*
+ 0.0071 * *RainfallA1*
+ 0.0011 * *DrainageSeason*=SD,WD
+ 0.0008 * *DrainageSeason*=WD
+ 0.6125 * *Runoff*
+ 1.2544 * *DrainageN1*
– 0.3346

LM13

Drainage =

0.0624 * *Season*=Spring,AW
+ 1.1715 * *Season*=AW
+ 0.0039 * *Day*
– 0.0002 * *Crop*=CIPAN,Rgi
+ 0.0011 * *Crop*=Rgi
+ 0.0001 * *Temp*
+ 0.0108 * *RainfallA1*
+ 0.0011 * *DrainageSeason*=SD,WD
+ 0.0008 * *DrainageSeason*=WD
+ 0.5701 * *Runoff*
+ 0.9718 * *DrainageN1*
– 1.4305

LM14

Drainage =

0.001 * *Season*=Spring,AW
+ 0.4886 * *Season*=AW
– 0.1648 * *Crop*=Spring_peas, Winter_peas, RGA, Winter_horse_bean, Rapeseed, Barley(spring), Wheat, CIPAN, Rgi
– 0.0004 * *Crop*=CIPAN,Rgi
+ 0.8263 * *Crop*=Rgi
+ 0.0167 * *Temp*
+ 0.1794 * *RainfallA1*
+ 0.0021 * *DrainageSeason*=SD,WD
+ 0.3181 * *DrainageSeason*=WD
+ 1.2638 * *Runoff*
+ 0.2849 * *DrainageN1*
–0.5682

Appendix C. List of publications

Kuzmanovski, V.; Džeroski, S.; Debeljak, M. *Integration of structured expert knowledge*. In: Proceeding of 4th Jožef Stefan International Postgraduate School – Student Conference (IPSSC) - Ljubljana, Slovenia. 137–143 (2012).

Kuzmanovski, V.; Džeroski, S.; Schulte, R.; Debeljak, M. Synergetic leaching model based on pathway and pressure factors. In: Proceedings of 17th International Nitrogen Workshop – Wexford, Ireland. 230–232 (2012)

Appendix D. Biography of the candidate

Vladimir Kuzmanovski was born in Kumanovo, Republic of Macedonia. He finished secondary school in his birthplace. In 2005 he started his studies at the Institute of Informatics, Faculty of Mathematics and Natural Sciences, University Ss Cyril and Methodius in Skopje, Republic of Macedonia. He was enrolled in 8-semester BSc program in the area of Computer and Software Engineering. He defended his BSc thesis titled “FastBit optimization of bitmap indices” under the supervision of Professor Goran Velinov.

In 2011 he started his master studies at the Jožef Stefan International Postgraduate School - Ljubljana, Republic of Slovenia.

His research is in the field of data mining and decision support systems and includes the study, developments and application of different data mining algorithms. His current research is concerned with developing models for accurate predictions of water flows from a filed.