Journal of Hydrology 517 (2014) 411–424

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Assessing hydrologic prediction uncertainty resulting from soft land cover classification

Lien Loosvelt [a,*], Bernard De Baets [b], Valentijn R.N. Pauwels [c], Niko E.C. Verhoest [a]

[a] Department of Forest and Water Management, Ghent University, Belgium
[b] Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Belgium
[c] Department of Civil Engineering, Monash University, Australia

ARTICLE INFO

Article history:
Received 3 October 2013
Received in revised form 4 March 2014
Accepted 20 May 2014
Available online 2 June 2014
This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Axel Bronstert, Associate Editor

Keywords:
Classification uncertainty
Monte Carlo simulation
Hydrology
Land cover map
Random Forests
Optical imagery

SUMMARY

For predictions in ungauged basins (PUB), environmental data is generally not available and needs to be inferred by indirect means. Existing technologies such as remote sensing are valuable tools for estimating the lacking data, as these technologies become more widely available and have a high areal coverage. However, indirect estimates of the environmental characteristics are prone to uncertainty. Hence, an improved understanding of the quality of the estimates and the development of methods for dealing with their associated uncertainty are essential to evolve towards accurate PUB. In this study, the impact of the uncertainty associated with the classification of land cover based on multi-temporal SPOT imagery, resulting from the use of the Random Forests classifier, on the predictions of the hydrologic model TOPLATS is investigated through a Monte Carlo simulation. The results show that the predictions of evapotranspiration, runoff and baseflow are hardly affected by the classification uncertainty when area-averaged predictions are intended, implying that uncertainty propagation is only advisable in case a spatial distribution of the predictions is relevant for decision making or is coupled to other spatially distributed models. Based on the resulting uncertainty map, guidelines for additional data collection are formulated in order to reduce the uncertainty for future model applications. Because a Monte Carlo-based uncertainty analysis is computationally very demanding, especially when complex models are involved, we developed a fast indicative uncertainty assessment method that allows for generating proxies of the Monte Carlo-based result in terms of the mean prediction and its associated uncertainty based on a single model evaluation. These proxies are shown to perform well and provide a good indication of the impact of classification uncertainty on the prediction result.

1. Introduction

Spatially distributed hydrologic models require information on the spatial distribution of land cover to simulate the water fluxes in a watershed. For each type of land cover, a set of biophysical parameters is defined in order to describe the land surface characteristics, usually within a land cover look-up table. As such, the effect of land cover on the hydrologic model response is ultimately controlled by the biophysical parameters. The latter play a crucial role within the model because they determine the energy and moisture exchange between the land surface and the atmosphere, such that changes in the vegetation parameters alter the energy budget modeled by the coupled land–atmosphere-transfer scheme and directly affect evapotranspiration, runoff and infiltration. Under conditions of data scarcity (including predictions in ungauged basins), land cover information is (highly) uncertain and affects the reliability of the model predictions when the uncertainty propagates through the hydrologic model. From a societal point of view, it is important to quantify this uncertainty because predictions of water fluxes are often used in early-warning systems for natural hazards or for assessing the effect of water resource infrastructures (Montanari et al., 2009).

When a land cover map of the catchment is not readily available, a classification of remote sensing data may provide this information (Singh and Woolhiser, 2002). However, one should be aware that the resulting map is an approximation of the complex reality and that substantial discrepancies between the real land cover and its representation may be present (Zhang and Goodchild, 2002). This awareness is spread among the user community as there is a growing demand to better document the quality of the produced map (Canters et al., 2002). Data quality research is often limited to simple overall measures such as Cohen's kappa coefficient, while information on the spatial distribution of the error is lacking.

* Corresponding author. Tel.: +32 92646140.
E-mail address: Lien.Loosvelt@UGent.be (L. Loosvelt).

http://dx.doi.org/10.1016/j.jhydrol.2014.05.049
0022-1694/

Nevertheless, it is obvious that an erroneous classification is most likely in areas with a high heterogeneity and for land cover types that exhibit high similarities in spectral properties or physical characteristics. It is therefore important (i) to obtain more detailed (spatial) information on the uncertainty of the classified map because the true value of the derived map cannot be assessed when this information is lacking and (ii) to understand how land cover uncertainties affect the hydrologic model predictions. As such, an effective support for decision making systems is provided.

Over the last 25 years, more attention has been given to uncertainty assessment and uncertainty propagation in models driven by remotely sensed data (Heuvelink and Burrough, 2002). For a review, we refer to Crosetto et al. (2000), Crosetto et al. (2001) and Foody and Atkinson (2002). Despite the growing insight into the effect of different land cover data sets (e.g. Pauwels and Wood, 2000; Wegehenkel et al., 2006), different spatial resolutions of the land cover map (e.g. Armstrong and Martz, 2008; Pauwels and Wood, 2000; Bormann, 2006; Bormann et al., 2009) and different spatial organizations of the landscape (e.g. Merz and Bardossy, 1998; Grayson and Blöschl, 2000; Bormann et al., 2009) on the environmental model response, research on the effect of classification error is limited and is often based on simple accuracy measures (e.g. Kyriakidis and Dungan, 2001; Miller et al., 2007; Livne and Svoray, 2011). In addition, also the impact of biophysical parameter uncertainty on environmental model predictions has been investigated (Breuer et al., 2006; Eckhardt et al., 2003; Liang and Guo, 2003), but was found to be less important than soil physical parameter uncertainty (Liang and Guo, 2003). Although an increasing number of studies on uncertainty propagation is being published, many questions about the effect of land cover information quality on hydrologic model predictions remain unanswered. With respect to this research topic, the following objectives are formulated:

- Evaluation of the uncertainty in water flux predictions due to land cover classification uncertainty. The aim of the uncertainty analysis (UA) is to identify the conditions under which land cover uncertainty has the highest impact and to better support the land cover classification schematization as a function of the modeling objective. In some cases, a lower quality or less detailed land cover map may be sufficient and resources can be saved.
- Evaluation of the sensitivity of water flux predictions to small changes in the biophysical parameters. Results of the sensitivity analysis (SA) allow to identify the biophysical parameters for which the hydrologic model is most sensitive. By decreasing the uncertainty in these parameters, the reliability of the water flux predictions can be improved.
- Development of a fast indicative uncertainty assessment method in order to estimate the uncertainty on the model predictions in a computationally efficient way. The aim is to generate uncertainty proxies based on a single model application.

The paper is structured as follows. In Section 2, a description of the study site and data is given, followed by a description of the hydrologic model set-up and the methods used for UA and SA. Section 3 discusses the results of the UA and SA and distinguishes between local predictions and area-averaged predictions. Finally, the main conclusions of this paper are summarized in Section 4 and are discussed with regard to PUB.

2. Methodology

2.1. Study area and data

The study is performed on the catchment of the Bellebeek (Belgium), which has a surface area of 91.24 km^2. A set of meteorological variables was registered with a temporal resolution of 10–60 min at the meteorological station situated near the outlet of the catchment (Samain et al., 2011). The meteorological records point out that the weather conditions in the catchment are representative for a temperate climate with a mean annual temperature of 11.5 °C and a total annual rainfall of 750 mm (uniformly distributed throughout the year). All meteorological forcings were collected for the period January 1, 2006 to December 31, 2007 and were aggregated to an hourly time step. Further, discharge observations with an hourly time step were continuously available at the catchment outlet.

A digital elevation model (DEM) of the study area is available (provided by the Flemish government) with a spatial resolution of 25 m and an accuracy of 0.07 m. It shows that elevation in the catchment ranges between 10 and 100 m a.m.s.l. A soil map of the study area was extracted from the FAO digital soil map and indicates a dominant presence of loam and silty loam soils. This coarse resolution soil map is chosen in order to limit the influence of soil variability when analyzing the effect of land cover confusion on the model prediction. A land cover map of the catchment is also available (provided by the Flemish government, derived from LANDSAT7 ETM+ imagery combined with field survey) and indicates the presence of 5 general classes: urban area (Ur), water area (Wa), broad-leaved forest (Bf), needle-leaved forest (Nf), pasture (Pa) and 7 specific classes: grass (Gr), potatoes (Po), beets (Be), maize (Ma), wheat (Wh), barley (Ba) and other crop species (Oc) (determined through mapping campaigns). This land cover map, further referred to as LCM-CROP, will serve as reference data in the classification analysis and indicates that the land use is dominated by cropland and pasture, intersected by urban area, forests and open water bodies. It is assumed that the crop information provided by LCM-CROP is accurate because the period of the field survey corresponds to the hydrologic simulation period.

Three cloud-free satellite images of the Bellebeek catchment were acquired on June 13 (DOY 164), June 30 (DOY 181) and July 5 (DOY 186) 2006, with a spatial resolution of 20 m. On June 13 and June 30, reflectance data in the green (0.50–0.59 μm), red (0.61–0.68 μm), near-infrared (0.79–0.89 μm) and mid-infrared (1.58–1.75 μm) wavelength regions were acquired by the SPOT 4 HRVIR. Mid-infrared radiances were not available for July 5 as the reflectance data were acquired by SPOT 2 HRV, on which a mid-infrared band is not present. The satellite images were orthorectified and geo-referenced. A radiometric correction was carried out to remove distortions due to differences in the sensitivity of the elementary detectors of the viewing instrument. Atmospheric correction was not carried out as the images were cloud-free such that the atmospheric influence could assumed to be constant over the entire image. Further, the optical images are predetermined for land cover classification such that comparison of the radiances among the different acquisition dates is not relevant.

2.2. Land cover classification

In this study, it is chosen to classify the land cover through the Random Forests (RF) algorithm. This classifier is an ensemble learning technique that builds multiple decision trees based on random bootstrapped samples of the training data (sampled with replacement) (Breiman, 2001). Consequently, each tree is constructed using a different bootstrap subset from the original training data, containing about two third of the cases. At each node in the decision tree, m variables are selected at random out of the n_{var} predictive variables and the best split among these m variables is used to split the node. By changing the set of predictive variables and the bootstrap subset over the different trees, the RF classifier introduces diversity among the classification trees. Through a majority vote of the classifier ensemble, the model output is determined. The cases left out of the construction of each tree (usually