Remote Sensing of Environment 174 (2016) 341–352

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Water observations from space: Mapping surface water from 25 years of Landsat imagery across Australia

N. Mueller [a,*], A. Lewis [a], D. Roberts [a,b], S. Ring [a], R. Melrose [a], J. Sixsmith [a], L. Lymburner [a], A. McIntyre [a], P. Tan [a], S. Curnow [a], A. Ip [a]

[a] Geoscience Australia, GPO Box 378, Canberra, ACT 2601, Australia
[b] Australian National University, Canberra, ACT 2601, Australia

ARTICLE INFO

Article history:
Received 30 July 2015
Received in revised form 26 October 2015
Accepted 10 November 2015
Available online 29 November 2015

Keywords:
Landsat
Surface water
Flood
Time series
Water resources

ABSTRACT

Following extreme flooding in eastern Australia in 2011, the Australian Government established a programme to improve access to flood information across Australia. As part of this, a project was undertaken to map the extent of surface water across Australia using the multi-decadal archive of Landsat satellite imagery. A water detection algorithm was used based on a decision tree classifier, and a comparison methodology using a logistic regression. This approach provided an understanding of the confidence in the water observations. The results were used to map the presence of surface water across the entire continent from every observation of 27 years of satellite imagery. The Water Observation from Space (WOfS) product provides insight into the behaviour of surface water across Australia through time, demonstrating where water is persistent, such as in reservoirs, and where it is ephemeral, such as on floodplains during a flood. In addition the WOfS product is useful for studies of wetland extent, aquatic species behaviour, hydrological models, land surface process modelling and groundwater recharge. This paper describes the WOfS methodology and shows how similar time-series analyses of nationally significant environmental variables might be conducted at the continental scale.

1. Introduction

Floods, droughts and water shortages pose global challenges (Vörösmarty et al. 2000). In Australia, severe flooding in late 2010 and early 2011 caused billions of dollars in damage and many deaths. As a result, the Australian Government announced the *Natural Disaster Insurance Review* which contained a provision to improve the availability of flood risk information. Reducing future flood impacts calls for improved mitigation planning and response underpinned by fundamental knowledge of the geography of floodplains and surface water (Shan et al. 2009). Severe floods are a feature of the Australian climate and landscape and are likely to continue with increasing regularity and severity. Therefore, from a government policy perspective, it is particularly important to understand where flooding may have occurred in the past to reduce its impact in the future through proper disaster planning and initiatives supporting communities to be better prepared and more disaster resilient. This means it is particularly important to understand the temporal aspect of these natural disasters at a continental scale.

In response to the National Disaster Insurance Review, Geoscience Australia was tasked with building, hosting, and populating a single authoritative source of flood related information in the National Flood Risk Information Project (NFRIP). One component of NFRIP is the production of historical flood information about the Australian continent from satellite imagery. Satellite imagery captures information about surface water in areas that are remote, inaccessible, extremely large or dangerous to approach, such as during floods (Frazier et al. 2000). Satellite images are especially valuable in areas that have sparse in-situ monitoring systems and areas with broad floodplains that are characterised by extreme flood events with large extent and temporal variability (Thomas et al. 2011). The detection of water from satellite imagery can also provide insight into the hydrological conditions of large rivers and the interactions between rivers and highly water-dependent ecosystems such as wetlands (Frazier et al. 2003; Thomas et al. 2011). Knowledge of the location, extent, persistence and recurrence of surface water is also needed for water resources assessments, for the allocation of water and regulation of its use (Khawlie et al. 2005; Morse et al. 1990), and for environmental water management (Kingsford 2000).

Flood mapping from the Landsat satellites began with Landsat-1 in the early 1970s (Robinove 1978), demonstrating that flood mapping from the Landsat satellites was feasible, potentially very accurate, and could assist with flood prediction, monitoring and relief (Smith 1997). The fixed orbital period of the Landsat satellites, and the presence of clouds, mean that the peak of the flood event is not always observed, which reduces the capacity of Landsat data to provide comprehensive flood mapping. However images acquired even several days after a flood peak are still able to capture a high proportion of the flood extent

* Corresponding author.

http://dx.doi.org/10.1016/j.rse.2015.11.003

in slowly changing systems, providing valuable information on the inundated area (Wang 2004). The methods for detecting water from optical satellite imagery are well established and are considered to be effectively operational at local and regional scales so long as the water target is not obscured by vegetation (Smith 1997). Detection methods typically exploit the absorption of longer wavelengths of light in water, especially the near and shortwave infra-red parts of the electromagnetic spectrum (Smith 1997; Frazier et al. 2000). This results in the corresponding infra-red spectral bands in the Landsat-5 and Landsat-7 satellites (Bands 4, 5 and 7) detecting low to no reflection from water as the wavelength increases, and hence being able to be used as a spectral indicator for water. Several methods take advantage of spectral indices exploiting the difference in reflectance between the visible and infra-red parts of the spectrum such as the Normalised Difference Vegetation Index (NDVI) (Tucker 1979) and the Normalised Difference Water Index (NDWI) (Gao 1996). Statistical classification methods applied to Landsat data have demonstrated highly accurate results, including supervised maximum likelihood (Frazier et al. 2000) and decision trees (Tulbure and Broich 2013).

Whilst these methods have proven accurate at local to regional scales, systematic surface water information products are required at the continental and multi-decadal scales. This type of continental-scale space-time analysis poses a number of challenges. First, sufficient observations must be collected over the temporal and geographic region of interest. Earth observation satellites have imaged Australia from 1972, building a large national archive of data with the potential to provide unique and comprehensive information on Australia's surface water (Draeger et al., 1997; Tulbure and Broich 2013). The temporally deep archives of Landsat data that have been acquired as a result of the Landsat long term acquisition plan (Arvidson et al., 2001) provide a consistent, long-term, continent-wide coverage for analysis (Thomas et al. 2011). Second, the time required to extract images from the tape-based archives, the manual tasks involved in processing and the difficulties of calibration and consistent rectification have all posed barriers. These barriers have only recently been overcome through bulk processing of satellite images (Purss et al. 2013) the use of physics-based processes to calibrate observations (Li et al. 2010), and systematic quality assurance of observations to remove artefacts (Sixsmith et al. 2013). Third, the processing required for these steps requires substantial technical infrastructure and high performance computing facilities capable of providing the necessary storage, processing and analysis platform. This was addressed by placing Geoscience Australia's archive and processing algorithms on the National Computational Infrastructure (NCI) which provides access to petabytes of high speed storage and 1.2 petaflops of processing power. High performance computing and storage combined with data standardisation and systematic quality flagging now makes it possible to apply sophisticated analyses on an entire continental Landsat archive over multiple decades.

In this paper, we present a consistent and continent-wide mapping of surface water through time using satellite imagery. We believe that we are the first to achieve such a large-scale continental analysis of surface water in both space and time for Australia. This involved analysing close to 200,000 Landsat images comprised of approximately 2×10^{13} individual observations covering the years 1987 to 2014 at 25 m resolution on the ground. The approach combines a number of water surface models and ancillary data sets to provide a level of confidence for the results. The base water detection method is similar to Tulbure and Broich (2013) in using a decision tree approach on a combination of spectral bands and derived indices. The decision tree method delivers a high accuracy across many environments while allowing fast processing suited to this very large dataset. The results of the decision tree classifier are summarised over all observations through time with ancillary analysis to obtain a confidence probability for each pixel value as a "confidence layer" that provides a measure of certainty in the results. This analysis provides a comprehensive and publicly-accessible product, called *Water Observations from Space* (WOfS), providing a continent-wide understanding of surface water persistence and recurrence, giving insight into which surface water bodies are frequently observed (such as dams or reservoirs), those which are observed infrequently (such as floods), and their temporal dynamics.

2. Data

2.1. Landsat archive

The core dataset is the Australian archive of Landsat-5 and Landsat-7 data from 1987 to 2014. Approximately 184,500 satellite images were produced from raw data using the USGS Landsat Product Generation System with a pixel size of 0.00025° (approximately 25 m resolution). The images are ortho-rectified, and corrected to measurements of surface reflectance using the method of Li et al. (2010). To facilitate processing, the images are spatially organised into 1×1 degree cells. Image acquisitions are organised as a set of data tiles corresponding to the cell area. The tiles therefore provide a complete time series of observations for every pixel and thus provide every observation for analysis (Fig. 1). We refer to this structure as the *Australian Geoscience Data Cube* (AGDC). The AGDC is located on the 'Raijin' supercomputer at the National Computational Infrastructure (NCI) and housed at the Australian National University (ANU) in Canberra, Australia. The Raijin supercomputer consists of approximately 10 petabytes of storage, Infiniband interconnect between the nodes, and 57,472 processing cores allowing large-scale parallel processing.

2.2. Shuttle radar topographic mission digital surface model (SRTM DSM)

The shuttle radar topography mission (SRTM) provides digital surface and elevation models on a near-global scale. The available resolution for government use is one arcsecond (approximately 30 m) for all of Australia (Gallant et al. 2011). The WOfS product makes use of the digital surface model (DSM) including on-ground features such as trees. The DSM was resampled to match the 25 m Landsat data and arranged into AGDC tiles, each covering one degree of latitude and one degree of longitude, named according to their south western corners. The SRTM DSM contributed to the analysis by providing indication of terrain shadow for pixel quality, steep slope masking, and through ancillary products (see Section 2.3).

2.3. Pixel quality

Observations from satellites are subject to many factors which can lead to poor observational quality, including instrument failure, instrument saturation, topographic shading, clouds, cloud shadows, and poor geo-location. Any pixel failing a series of automated quality tests was masked. The methodology of Sixsmith et al. (2013) was followed whereby for every pixel q and at observation time $t \in \{1,2,\ldots,N\}$ where N denotes the maximum number of observations, the pixel quality q_t is obtained by amalgamating the binary outcome of each quality indicator. If any of the indicators are triggered, q_t takes on a non-zero

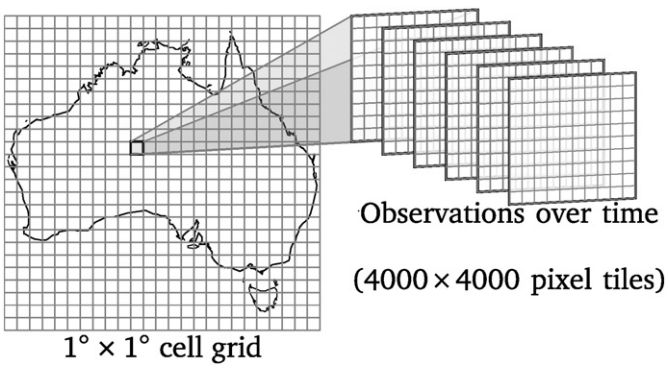

Fig. 1. The data cube. Each $1° \times 1°$ cell contains a time-series of observations as 'tiles' of data.

value and the pixel will be masked in the analysis. The quality indicators considered are:

Pixel saturation: Saturation was evaluated for each spectral band per pixel except the panchromatic band of Landsat 7. Pixel saturation can only be determined from Level 1 (L1) products produced from the Level 1 Product Generation System (LPGS). Pixels with a byte-scaled radiance value of 1 and 255 represent under and over saturated pixels respectively. For the purposes of pixel quality, both under and over saturated pixels share the same flag.

Band contiguity: Contiguity is defined as a pixel having a valid observation for every spectral band. A missing value in any one band will flag that pixel as being non-contiguous. This is considered an important factor when calculating band ratios, where a pixel may only provide a valid ratio result where the component band values are valid.

Clouded or cloud shadow: Clouded pixels were flagged using two separate methods of cloud detection, similar to that used with the *Web Enabled Landsat Data* (WELD) products (United States Geological Survey 2014). The presence of cloud and cloud shadow is computed by combining the ACCA (Irish 2000) and Fmask (Zhu and Woodcock 2012) methods to provide a combined certainty measure of a pixel being cloudy or shadowed.

Terrain shadow: Topographic shadow is similar to cloud shadow in that the underlying spectral characteristics of the surface are not truly represented. The algorithm used for identifying terrain shadowed pixels is detailed in Li et al. (2012).

2.4. Ancillary data and products

A number of datasets and products provide support to the base water classification model. These datasets and products are modelled as a set of "experts" (Cesa-Bianchi et al., 2006) that indicate the reliability of the surface water classification at any point in the landscape.

2.4.1. MrVBF. MrVBF is a multi-resolution valley bottom flatness product derived from the SRTM DSM (Gallant and Dowling 2003). The presence of a flat valley bottom is generally consistent with an observation of surface water, and MrVBF is designed to indicate deposition areas. MrVBF is an integer value with range 0 to 9. MrVBF was implemented as a 25 m resolution Australia-wide coverage to match the Landsat data in resolution and projection.

2.4.2. Slope. Slope is also derived from the SRTM DSM. The hypothesis is that, as slope increases, it becomes less likely that surface water will be present at a location. A high value of slope therefore suggests that a classification of water is less likely to be correct. Slope is a decimal value, in degrees, with range of 0–90.

2.4.3. Open water likelihood (OWL). The MODIS open water likelihood (Guerschman et al. 2011) is a classification model of the likelihood that water had been observed between 1999 and 2010 based on MODIS short-wave infra-red spectral data, the normalised difference vegetation index (NDVI), the normalised difference water index (NDWI), and the multi-resolution valley bottom flatness (MrVBF) index. The factor weights in the OWL classification model are determined using a logistic regression. The OWL product provides a comparative satellite-based surface water product across all of Australia for the period 1999 to 2010. The result is a percentage coded as a floating point variable ranging from 0 to 1.

2.4.4. Australian hydrological geospatial fabric (geofabric). The Australian hydrological geospatial fabric (Atkinson et al. 2008) is a vector GIS representation of hydrological features derived from digitisation of topographic map features and analysis of elevation models. Each polygonal Geofabric feature type was converted to raster using GDAL Transform at a resolution and projection matching the Landsat data, and coded as a binary variable with values of 0 (no feature) or 1 (hydrological feature present).

2.4.5. Australian statistical geography standard (ASGS). The Australian statistical geography standard (ASGS) by the Australian Bureau of Statistics (2012) is a geographical framework effective from July 2011. The Australian Bureau of Statistics ASGS 2011 Urban Centre and Locality dataset (for urban centres of populations of 100,000 and over) was used to derive an "urban area" indicator with values of 0 (not an urban area) or 1 (urban area).

3. Method

The requirements for the model implementation were manyfold. First, a single water classifier was required for the entire continent of Australia, that was robust against climatic changes across multiple decades, and the varied, regional climates observed across the Australian continent (tropical, subtropical, desert, grassland, and temperate). Second, as the model was applied over decades of data, an algorithm was required that was computationally efficient so that the model could be recalculated in a reasonable time frame when new data was received or the algorithm revised. Third, the model needed to accurately detect intermittent water bodies such as occur with flooding, and account for the wide variety of optical properties encountered in inland waters (Dekker et al. 1997). Finally, since the results would be released to the public through an online portal, the model output needed to be simple enough so that it could be easily interpreted and explained to a non-technical person, while at the same time provide a level of detail to suit future scientific applications and an indication of confidence of each observation. To meet these requirements the proposed model combines on a pixel-by-pixel basis: a water classification on each satellite observation, pixel quality information to determine valid from invalid observations, and ancillary information from number of third party "expert" datasets to provide a confidence assessment in the results.

3.1. Water classification

The Landsat archive provides a time series of N observations for each $1^\circ \times 1^\circ$ tile stored in the AGDC, with each tile consisting of 4000×4000 pixels. The classification of each pixel as water (coded as 1) or not-water (coded as 0) over time is denoted by $\psi_1, \psi_2, \ldots, \psi_N$. Thus $\psi_t \in \{0,1\}$ is the classification of a particular pixel at observation time $t \in \{1,2,\ldots,N\}$.

The classifications $(\psi_t)_{1 \le t \le N}$ were obtained using a regression tree classification (Breiman et al. 1984) that used individual Landsat spectral bands and normalised difference ratios (NDI) commonly used in water detection (Smith 1997; Frazier et al. 2000). The regression tree was trained on a set of water and non-water training samples that were manually selected from 20 tiles across Australia as shown in Fig. 2. Each one of these training tiles was analysed over multiple seasons to provide data from a total of 59 Landsat scenes as samples. The scenes were chosen to represent the broad variety of landscapes, vegetation, soil and water colour that occurs across Australia.

The training samples were chosen by breaking each tile into objects using eCognition (Trimble Geospatial Imaging, 2014) and then manually classifying objects within each scene into one of 26 classes, representing a variety of water and non-water targets. Objects were selected based on the criteria that they only contain pixels of the required classes and they represent a broad variety of conditions for that class (such as clear water, turbid water and water/non-water mixtures). Each classified object was then broken down into single pixels to obtain a large number of training samples with a variety of spectra to cover the range of different characteristics of each class. The data derived from each sample consisted of individual spectral bands and associated normalised difference ratios. In total, 2.8 million samples were created to produce a pool of training data. A random set was then generated from the sample pool, resulting in 180,000 samples for the regression analysis covering all 26 classes,

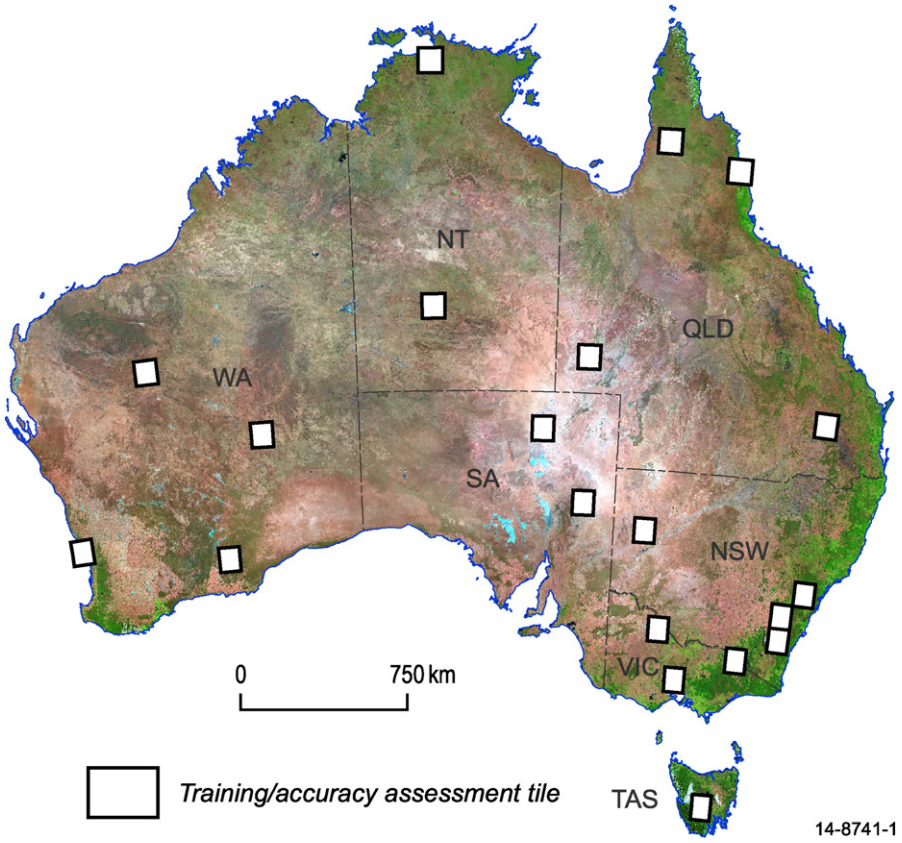

Fig. 2. Location of the twenty 1° × 1° cells used to generate training data for the model. Training data were selected from multiple Landsat scenes covering each tile.

and the regression implemented on every pixel to produce a binary classifier of water or non-water. The sample class types are shown in Table 1.

To determine the best threshold regime the regression tree outputs were tested in three scenarios: (1) where only spectral bands were used, (2) where only normalised difference ratios were used, (3) where both spectral bands and normalised difference indices were combined. The test was assessed by comparing the performance of the resulting regression trees in obtaining a predicted 95% classification accuracy in the smallest number of threshold steps. This resulted in a regression tree based on variables from both spectral bands and normalised difference ratios. The scoring of the best variables is shown in Table 2.

The resulting regression tree was optimised by analysing the relative cost of increasing the number of tree splits versus the improvement in accuracy obtained as the number of splits increased. A level of pruning was chosen beyond which each extra threshold step created more processing cost than the associated accuracy increase. The level of pruning resulted in a projected accuracy of 97% using 23 steps. The final classification tree is shown in Fig. 3.

Table 1
Sample class types used in the manual classification to create samples for the regression analysis.

Not water	Water
Bare	Cloud shadow on water
Building shadow	Estuarine
Cloud shadow on bare	Large water body
Cloud shadow on veg	River water
Cropping bare	Saline flats
Cropping dense veg	Salt lake
Dark soil	Sea
Road	Small water body
Salt	Swamp
Snow	Water and veg mix
Terrain shadow on bare	
Terrain shadow on snow	
Terrain shadow on veg	
Forest	
Grassland	
Riparian veg	

3.2. The confidence and summary layers

The model applies a logistic regression to provide a probabilistic linear classification to create a comparison between the water classification results and a set of other Australia-wide datasets related to surface water analysis. A linear regression measures the relationship between a categorical dependent variable and one or more independent variables (aka. factors, predictor variables, or features) by using probability scores as the predicted values of the dependent variable (see for example Hastie et al. (2009)). In (binary) logistic regression, the binary outcome

Table 2
Variable importance produced by the regression tree from the prospective spectral bands and normalised difference ratios for the water classifier. Higher values indicate a higher classification effectiveness of the variable. Spectral bands are designated as TM# (Thematic Mapper Band). Normalised difference ratios are designated as NDI_XY (Normalised Difference Index of Band X and Band Y).

Variable	Score
NDI_52	100.
NDI_72	98.5162
TM5	97.9127
TM7	77.2063
NDI_43	73.8958
TM1	22.0880
TM3	13.0902

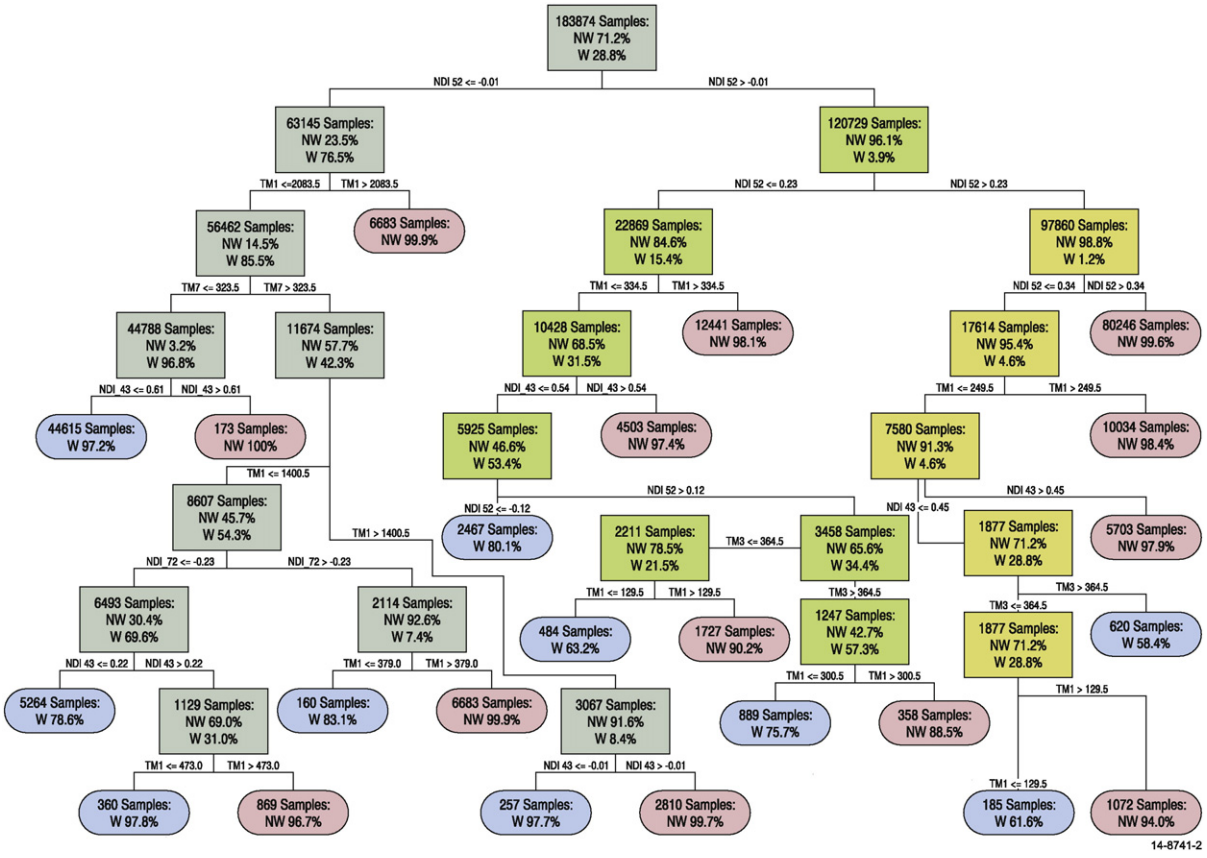

Fig. 3. Diagram of the regression tree underlying the water detection classifier. Tree branches are shown in green with endpoint for water and not-water displayed as blue and red respectively. Each branch indicates the variable used to split and the resulting balance of water and not-water samples created by the split.

of the model is usually coded as 0 or 1. The probability that a given sample Y, with factor vector X should be classified 1 is

$$\Pr(Y=1)=\sigma\left(\beta^T X\right). \tag{1}$$

where β is a vector of weights that defines a line in the factor space. The logistic function σ, defined by

$$\sigma(x):=\frac{1}{1+e^{-\beta^T x}}, \tag{2}$$

takes the output of the linear function $\beta^T x$ that ranges from $-\infty$ to ∞ and converts it to a probability that ranges from 0 to 1. The weight vector β can be found by directly optimising the log-likelihood of a training set.

For each pixel, there are M experts: $f_1, f_2, \ldots, f_M$. The experts here are third-party products and data sets that provide advice as to whether an observation of water in a location is likely. Depending on the expert, advice is provided in different ways. For example, as a binary outcome (e.g.: 1 = hydrological feature, 0 = no feature), a likelihood (of water in the range [0,1]), a real value, etc. In the implementation used here, the ancillary datasets and products were mapped to the experts as follows: MrVBF (f_1), MODIS OWL (f_2), slope (f_3), the features in the Geofabric product as f_4 to f_{12} (see Table 4), and the urban areas indicator as f_{13}. Hence the analysis used $M=13$ experts. However the methodology could potentially accommodate more ancillary products and datasets in the future.

Using p to denote a pixel instance and applying a logistic regression framework, the probability of classification of this pixel p as water (coded as 1) is.

$$\Pr(p=1)=\sigma(\alpha_1\psi_1+\cdots+\alpha_N\psi_N+\beta_1 f_1+\cdots+\beta_M f_M) \tag{3}$$

where the $N+M$ weights $\alpha_1, \ldots, \alpha_N, \beta_1, \ldots, \beta_M$ are to be determined. For the 1987 to 2014 period of the data, the number of observations N varies between 600 and 1200 depending on the location of pixel under consideration. To simplify the calculation, we used the term

$$\alpha_1\psi_1+\alpha_2\psi_2+\cdots+\alpha_N\psi_N \tag{4}$$

and made the choice of weights using

$$\alpha_t:=\begin{cases}\beta_0/C, & q_t=0,\\ 0, & q_t>0,\end{cases} \tag{5}$$

for $1\le t\le N$, where $\beta_0\in$ is a new weight to be determined, q_t is the pixel quality at observation time t, and C is the total of number of times that the pixel was not masked by pixel quality, i.e.,

$$C:=\sum_{t=1}^{N} I_{\{q_t=0\}} \tag{6}$$

where I is the indicator function ($I_A=1$ if A is true, and 0 otherwise). Next the classifications $(\psi_t)_{1\le t\le N}$ were re-indexed in terms of the C clear observations (as determined by pixel quality): $\psi_{(1)}$ is the

classification of the first clear observation, $\psi_{(2)}$ is classification of the second clear observation, and so forth up to $\psi_{(C)}$. Using the choice for the weights $\alpha_1, \ldots, \alpha_N$ gives

$$\alpha_1\psi_1 + \alpha_2\psi_2 + \cdots + \alpha_N\psi_N \tag{7}$$

$$= \frac{\beta_0}{C}\psi_{(1)} + \frac{\beta_0}{C}\psi_{(2)} + \cdots + \frac{\beta_0}{C}\psi_{(C)} \tag{8}$$

$$= \frac{\beta_0}{C}\left(\psi_{(1)} + \psi_{(2)} + \cdots + \psi_{(C)}\right) \tag{9}$$

$$= \beta_0\frac{W}{C} \tag{10}$$

where W is the number of times that the pixel is classified as water by the decision tree. This gives the simplified logistic regression

$$\Pr(p=1) = \sigma\left(\beta_0\frac{W}{C} + \beta_1 f_1 + \cdots + \beta_M f_M\right) \tag{11}$$

where the $M+1$ weights $\beta_0, \beta_1, \ldots, \beta_M$ need to be determined. At this point the model no longer depends directly on N but only on the proportion W/C which will vary between 0 and 1 and can be updated (by adding new observations) without changing the weight β_0. The term

$$S := \frac{W}{C}, \tag{12}$$

is termed the *summary* as it is the proportion of clear observations of a pixel in which water was detected (see Fig. 4). Higher values of S tend to indicate that water is consistently detected (i.e., permanent water bodies) and lower values indicate a pixel where water is detected irregularly. The summary S provides an easy to interpret (and visually striking) representation of surface water over time. This results in the model

$$\Pr(p=1) = \sigma(\beta_0 S + \beta_1 f_1 + \cdots + \beta_M f_M). \tag{13}$$

Hence, the model provides an assessment of confidence in the surface water classification, by combining a set of surface water datasets

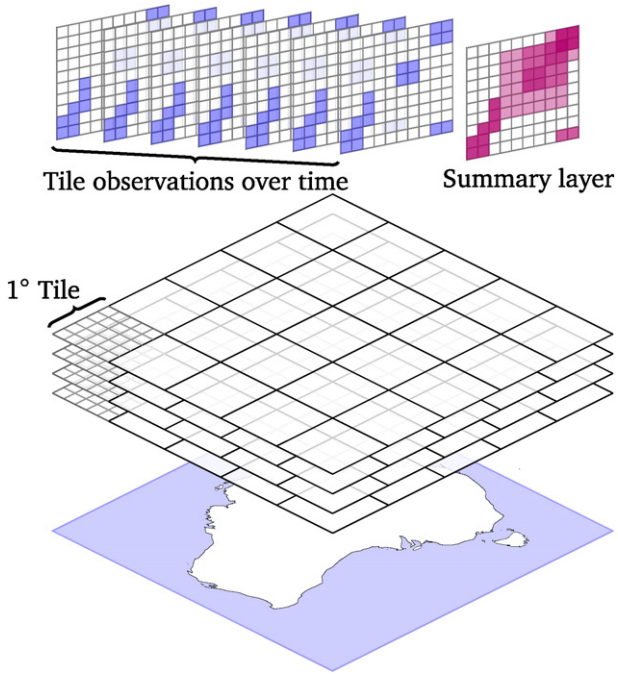

Fig. 4. Each 1° tile is composed of 4000 × 4000 pixels. The pixels in each tile are classified as water or not-water. The classifications over time are aggregated into a summary layer which gives the proportion of times that water was detected.

Table 3
Accuracy assessment of the water classifier including a breakdown into the 26 spectral subclasses. Situations where the classifier performs poorly are highlighted in red and where the classifier performs well are highlighted in green.

Spectral Subclass	Water		Not Water	
	%	#	%	#
Bare	0%	2	100%	756,974
Building shadow	6%	198	94%	2903
Cloud shadow bare	6%	4167	94%	67,852
Cloud shadow veg	2%	3033	98%	134,788
Cropping bare	0%	0	100%	60,210
Cropping dense veg	0%	0	100%	34,762
Dark soil	0%	36	100%	17,450
Road	1%	60	99%	5337
Salt	1%	1023	99%	93,141
Snow	0%	113	100%	93,695
Terrain shadow bare	11%	44,492	89%	352,121
Terrain shadow snow	1%	64	99%	7659
Terrain shadow veg	4%	3161	96%	74,022
Forest	0%	24	100%	285,875
Grassland	0%	28	100%	594,384
Riparian veg	0%	184	100%	61,587
Cloud shadow water	99%	1676	1%	11
Estuary	95%	72,585	5%	3850
Large water body	98%	124,826	2%	3057
River	80%	46,778	20%	11,651
Saline flats	92%	404	8%	33
Salt lake	99%	13,982	1%	139
Sea	98%	339,876	2%	5932
Small water body	88%	12,266	12%	1730
Swamp	63%	22,758	37%	13,519
Water veg mix	74%	34,636	26%	12,060

Producers accuracy	93%	98%
Users accuracy	92%	98%

Overall accuracy	97%

Table 4
Logistic regression model weights.

Symbol	Description	Weight	Value
S	Summary	β_0	0.1703
f_1	MrVBF	β_1	0.1671
f_2	MODIS OWL	β_2	0.0336
f_3	Slope	β_3	−0.2522
f_4	Geofabric – canal	β_4	0.0000
f_5	Geofabric – foreshore	β_5	4.2062
f_6	Geofabric – pondage	β_6	−5.4692
f_7	Geofabric – reservoir	β_7	0.6574
f_8	Geofabric – flat	β_8	0.7700
f_9	Geofabric – lake	β_9	1.9992
f_{10}	Geofabric – rapid	β_{10}	0.0000
f_{11}	Geofabric – swamp	β_{11}	1.3231
f_{12}	Geofabric – watercourse	β_{12}	1.9206
f_{13}	Urban Areas	β_{13}	−4.9358

and terrain-based factors with the Landsat observation frequency to create a confidence layer. This layer provides a point of comparison with the surface water classification to indicate whether the classification results are in line with other surface water datasets.

By itself the raw summary value S has a flaw due to classification errors that might occur in the classifiers $(\psi_{(t)})_{1\le t\le C}$, or alternatively, errors in the pixel quality mask (e.g., a cloud pixel is not flagged as having cloud cover). For example, consider the situation where $C=100$ clear observations and one of the classifiers mistakenly classifies a cloud (or shadow) as water while all the other classifiers flag it as no-water, hence $W=1$. This would give a summary score of $S=0.01$. This is a phenomenon leading to noise in the summary, often characterised by 'lone pixels' with a small summary value S. A naïve way to fix these artefacts would be to simply filter S based on some arbitrary low threshold (say, 0.05) and then set $S=0$ when $S<0.05$. Although this does remove the artefacts and noise, it also removes all traces of the low frequency surface water that characterises flooding and ephemeral water bodies. This is resolved by filtering the summary S by the probability of the pixel being an actual water pixel, by defining the *filtered summary* $\tilde{S}$ as

$$\tilde{S} := \begin{cases} S & \text{if } \Pr(p=1) > \kappa, \\ 0 & \text{otherwise}, \end{cases} \tag{14}$$

where κ is a small value to be chosen (e.g., $\kappa=0$). This approach is significantly more robust as it filters the summary using the advice from all the experts. It also removes a significant amount of noise but keeps the low frequency water observations.

4. Results and discussion

The analysis procedure generated surface water classifications for each pixel, for each observation, from 1987 to 2014. Water classifier accuracy was assessed by creating an additional set of test samples over the same locations as the original training data, but from different years, resulting in an additional 3.4 million samples, thereby ensuring that the accuracy assessment data were independent of the training data. The assessment is presented as a confusion matrix in Table 3, and indicates an overall classification accuracy assessment of 97%.

The confusion matrix highlights where the classifier performs well, and where it performs poorly. Areas identified as water are being correctly identified 93% of the time and are being misclassified as not-water 7% of the time. These errors of omission typically occur along rivers, small waterbodies and swamps where the presence of both water and vegetation within the pixel leads to a failure to identify water. This means that the WOfS product is likely to underestimate the extent of water in locations that contain mixed water and vegetation pixels. As a consequence the WOfS product may not be fit for applications that require information about the inundation characteristics of vegetated wetlands, small farm dams, and rivers with significant riparian vegetation.

Areas of water are being incorrectly identified within not-water areas in steep terrain or dense urban areas where shaded pixels are

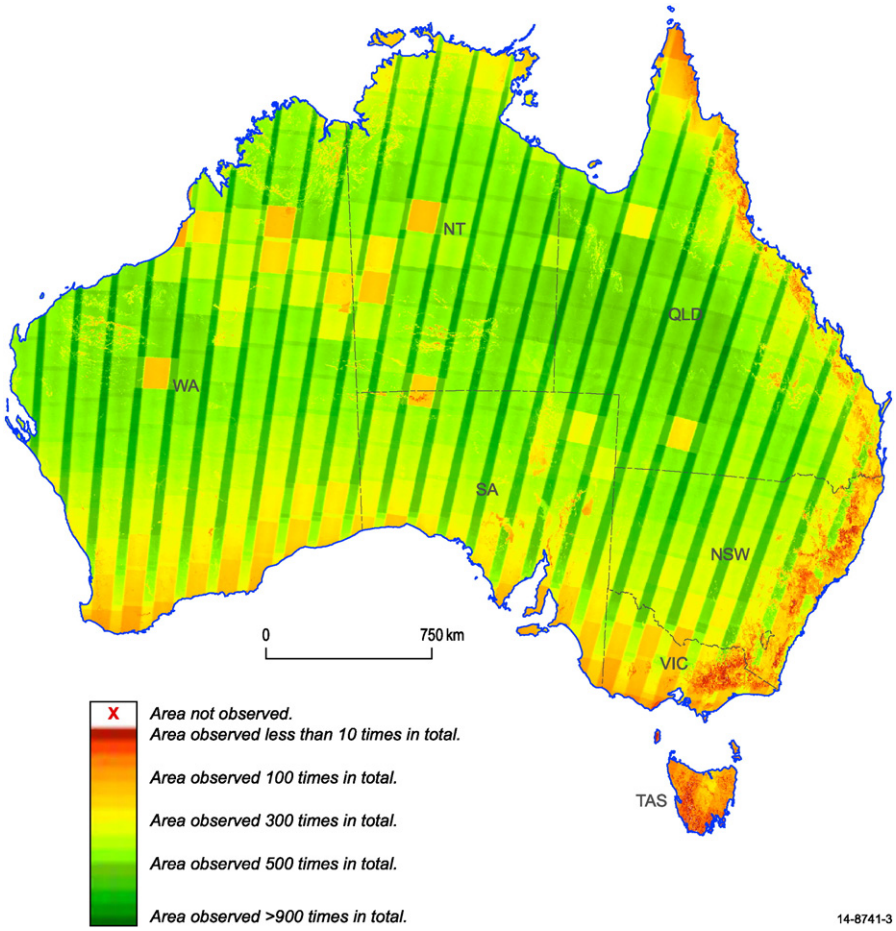

Fig. 5. Heatmap showing number of clear observations per pixel across Australia for the 1987–2014 period.

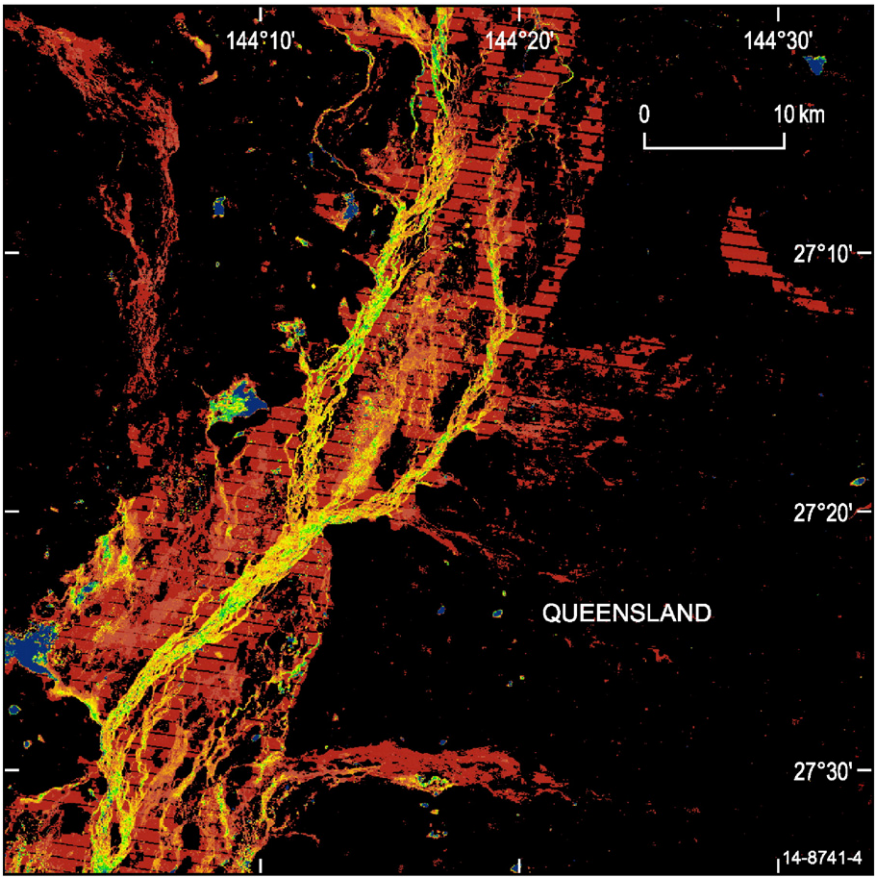

Fig. 6. An area of south-western Queensland from the WOfS product, demonstrating some instances of inundation affected by the "venetian blinds" of the Landsat-7 SLC-Off fault.

misclassified as water. These errors of commission are occurring in 8% of samples. This means that the WOfS product may overestimate the amount of water in locations that are in areas of steep terrain or in dense urban areas. The terrain and urban data used in the confidence layer help to reduce this overestimation, however some residual errors remain. As a consequence of this the WOfS product may not be fit for applications that require information about the inundation characteristics of urban areas or in mountainous regions.

A significant issue for large water bodies is signal noise for very clear water (Nichol and Vohora 2004). Data values in areas of very clear water are extremely low, often only 1 to 2 DN in the uncorrected Landsat data. This results in corresponding low values once the surface reflectance correction has been implemented, with additional issues from any error in the ancillary data used to produce the correction. As such it becomes possible for the noise to exceed the measurement by the Thematic Mapper sensor and hence the observed spectra to indicate that the target is not water. The observed values of NDI_43 and NDI_52 (see Table 2) can easily result in a water pixel in the centre of a lake being detected as not-water as the noise results in unusual values and the resulting index displays a strong positive value where it should physically be equally negative. Hence some issues arise in permanent water bodies (and ocean areas) occasionally being classified as not-water. This appears as speckle within large water bodies. A curious side effect of this behaviour is that shallow areas often display as having a higher water observation frequency than deep areas, apparently due to the improved signal to noise associated with the contribution of substrate reflectance. This is a subject for further investigation.

Another error that is not accounted for by the classification algorithm derives from problems with Scan-Line-Corrector-Off (SLC-Off) problem

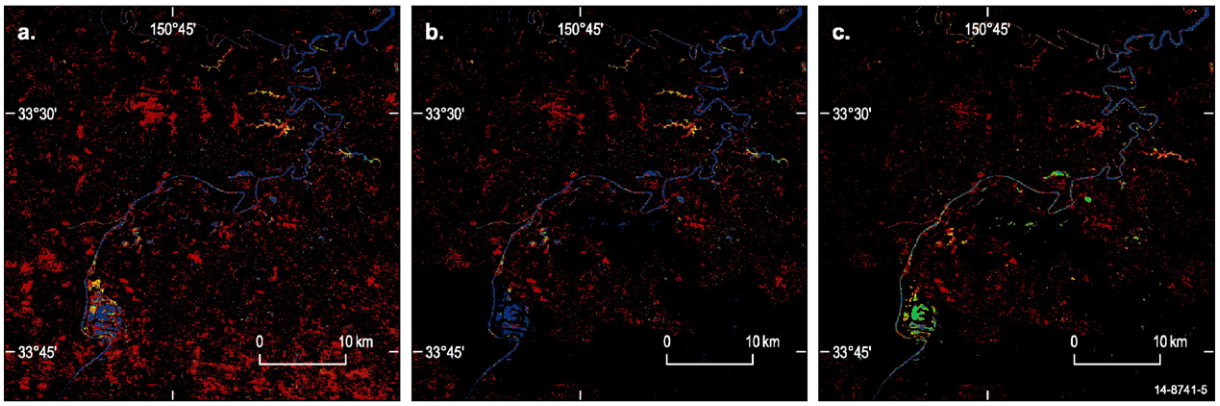

Fig. 7. Sydney area confidence filtered at $\kappa = 0\%$ (left), $\kappa = 1\%$ (centre), and $\kappa = 2\%$ (right).

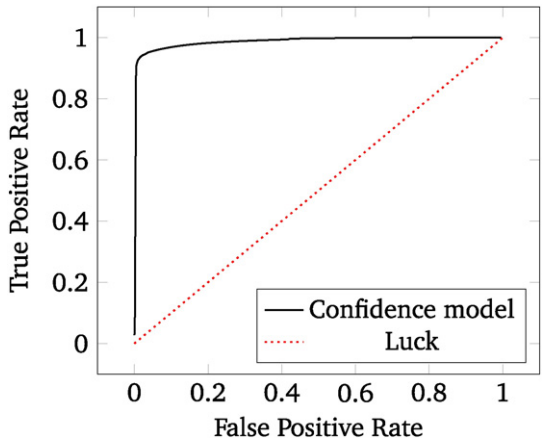

Fig. 8. Receiver operating characteristic (ROC) of the surface water model. A stratified six-fold cross-validation was applied and the average ROC is plotted. We compare the full model vs. a model where only the summary layer is used for classification (i.e., $\beta_i = 0$ for $i > 0$).

on Landsat-7 (Markham et al. 2004). The SLC-Off problem creates data gaps in the Landsat-7 imagery, worsening towards the east and west edges of the data. The classification results from SLC-Off data create corresponding gaps in the classification, and hence an inundation event will appear as though viewed through "venetian blinds" (Fig. 6). While methods exist to correct the SLC-Off gaps in Landsat 7 data, the current policy around data in the AGDC is that only "real" observations will be included, so the SLC-Off error will be an artefact in WOfS for the foreseeable future.

The confidence layer provides an assessment of the validity of the summarised water results for each pixel. The confidence was applied to the water summary layer to mask any pixels classified as water where the confidence was less that 1%. Fig. 7 demonstrates the effect that applying the confidence layer has on the water summary for confidence levels (κ) of 0, 1 and 2%.

Filtering the summarised water result at an inappropriately high confidence removes all results that indicate the presence of ephemeral water such as floods.

To assess the degree to which the logistic regression model (for the confidence layer) correctly classifies water, a stratified six-fold cross-validation was applied using the training tiles selected from across Australia. The entire training data set was randomly divided into six even subsets containing similar proportions of water and not-water samples. A different set of samples was iteratively chosen as the testing set and the other five subsets of samples were combined into a training set. An individual logistic regression model was constructed based on each training set and used for predicting the testing set. This methodology was repeated six times until all the subsets were tested. Area under

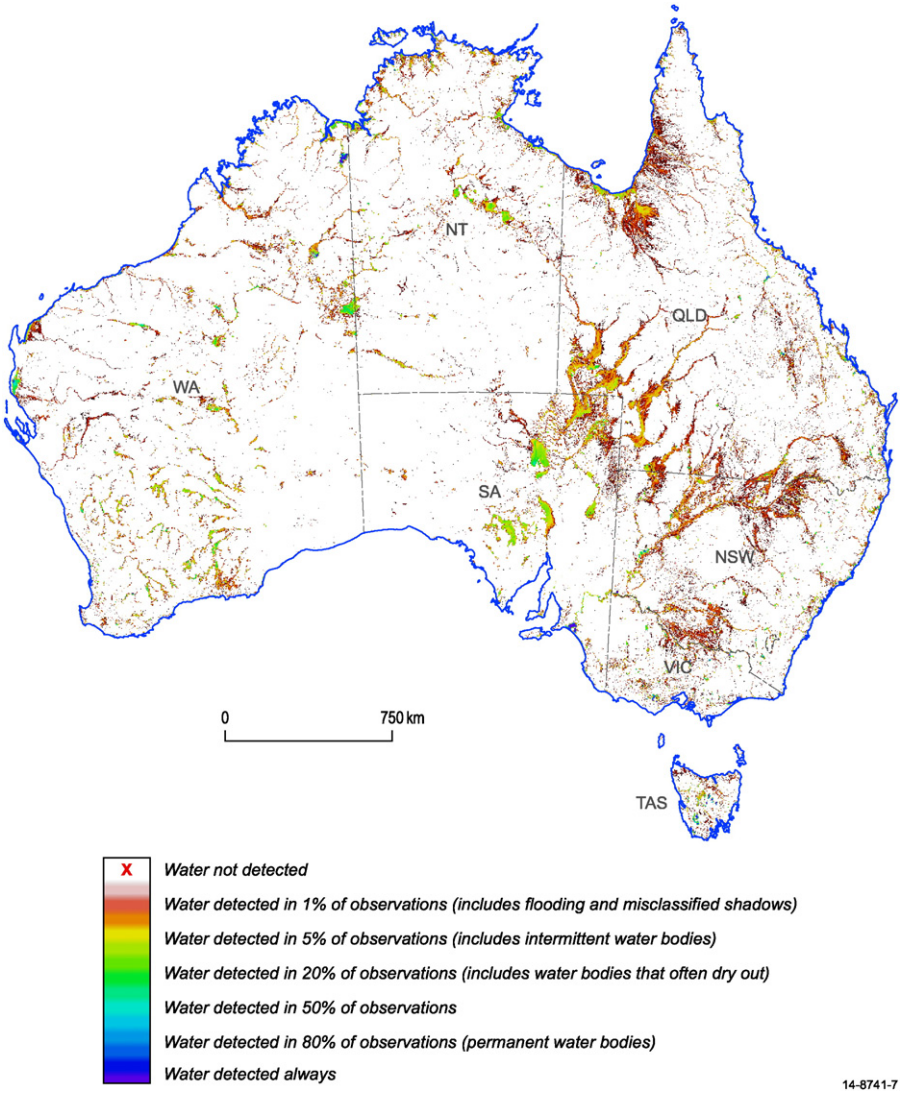

Fig. 9. WOfS filtered summary product for Australia, derived from water observations from 1987 to 2014.

the *receiver operating characteristic* (ROC) curve was used to assess the performance of the logistic regression for the confidence layer. This is a commonly used metric for evaluating statistical learning algorithms (Bradley 1997). The ROC curve in Fig. 8 was generated by varying the threshold and plotting the corresponding True Positive (TP) and False Positive (FP) rates. When interpreting such curves, models generating points closer to a TP rate of 1 for all FP rates are considered better than models with a TP rate further away from 1. At the extreme, achieving a value of (0,1) represents a perfect classification with zero FP rate and 100% TP rate. A poor classifier would yield a ROC curve exhibiting a line through the diagonal (FP rate = TP rate) which is an equivalent to the model producing nothing more than a random result (i.e., "luck") and making an equal number of correct classifications between positive and negative samples. A classifier exhibiting values below the diagonal is considered worse than a random result.

The ROC curve shown in Fig. 8 is the average ROC curve over the six-fold cross-validation with a mean area under the curve (AUC) of 0.985. Introducing "experts" into the model shows the gain in classification accuracy obtained compared to a model whereby only the summary value S is used, i.e., β_1 to β_M set to zero in (13).

The resulting weights from the logistic regression for the confidence layer demonstrate the degree to which an 'expert' agrees or disagrees with the water summary created from the water classifier. Table 4 shows the individual weightings for each factor in the logistic regression.

The 'experts' employed in the model were:

- MrVBF. The presence of a flat valley bottom is generally consistent with an observation of surface water, and MrVBF is designed to indicate deposition areas.
- Slope. A high value of slope indicates that a classification of water is less likely to be correct.
- MODIS-OWL. A high open water value indicates increased likelihood in the detection of surface water.
- Urban areas derived from the ASGS 2011 Urban Centre and Locality dataset, for urban centres of populations of 100,000 and over. In areas where there is a significant amount of urban development the water detection algorithm was confounded by the deep shadows cast by multi-storey buildings and the generally noisy spectral response created by structures.
- GeoFabric, showing known topographic surface water features. The components of Geofabric used for the analysis were: canal, foreshore, pondage, reservoir, flat, lake, rapid, swamp, and watercourse. Each component was used as a binary component in the confidence calculation.
- Quality of observations. Where observation conditions are poor, it becomes less likely that a good water observation can be made. Hence image artefacts that are masked by the pixel quality data reduce the number of times that a particular pixel was clearly observed. After applying pixel quality filters, the number of clear observations per pixel across the Australian land surface ranged from over 1200 in central Australia, to under 10 in persistently cloudy sites and mountainous areas. The number of clear observations was also greater in the overlap areas between adjacent satellite overpasses (Fig. 5). These overlap areas are larger toward southern Australia as the satellite orbits converge toward the pole.

Positive weights indicate agreement, while negative weights indicate disagreement. The expectation was to have greater confidence in the detection of water if it coincided with hydrologic features mapped in GeoFabric. Table 4 shows that the Geofabric factors have varying weights from the most positive to the most negative. The more positive features correlate with larger Geofabric polygons while the lowest correlate with the smallest or those where the resolution of Landsat is too low to reliably detect these features (such as rapids and canals). The pondage feature is strongly negative due to these being small, temporary bodies that rarely contain water. Slope shows a negative weight, which is in line with increasing slope correlating with a decreasing likelihood of the presence of significant surface water. MrVBF is positive as increasing MrVBF values correlate with larger, flatter terrain. The low, positive value of the MODIS OWL weighting corresponds to MODIS having a far lower resolution than Landsat and consequently a lower ability to detect smaller water features. Typically, in areas where there is a significant amount of urban development, the water detection algorithm is confounded by the deep shadows cast by multi-storey buildings and the generally noisy spectral response created by structures. Similar issues were found by Frazier et al. (2000) and Feng et al. (2015) for Landsat data, and Shackelford and Davis (2003) for IKONOS data. Our 'urban area' indicator derived from the ASGS dataset show a strong negative weighting, which reflects that major surface water features generally occur away from the major urban centres. Where a water feature occurs within an urban centre, it is accounted for in the Geofabric data.

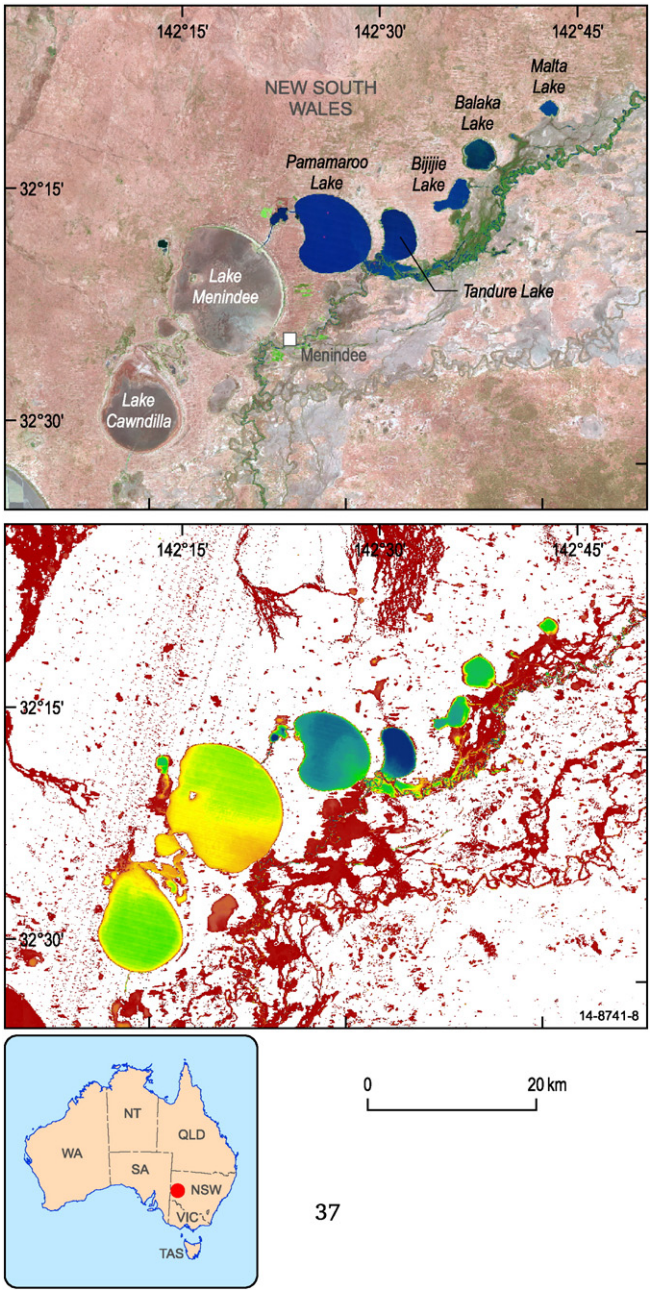

Fig. 10. The Menindee Lakes region of the Darling River in western New South Wales and the associated WOfS Filtered Summary. Dark blue areas indicate permanent water, while red areas indicate infrequent water observations including flood-related phenomena.

There are currently two global water body datasets that are derived from large collections of Landsat data: the Global Inland Water (GIW) body dataset for 2000 (Feng et al. 2015) and the GLObal WAter BOdies database (GLOWABO) by Verpoorter et al. (2014). Both demonstrate very high accuracy in mapping water bodies. GIW uses the year 2000 coverage of the Global Land Survey (GLS) of Landsat 7 ETM+ data (Gutman et al. 2008), while GLOWABO uses the GeoCover product of Landsat 7 ETM+ data from around 2000. Each produces a comprehensive coverage of water bodies of the world comprised of a mosaic of single date classifications and hence is a best snapshot of water body extent for a particular time. These products do not seek to provide an understanding of the variability in water body extent over time, whereas the primary driver for WOfS was to provide information on flooding by understanding where water was a common occurrence compared with where it was rarely observed. In addition the GIW and GLOWABO are derived from relatively cloud-free imagery, while WOfS uses every pixel of the Australian Landsat archive and masks cloud on a case-by-case basis. This generates an extensive time series of data for every pixel across the continent, and enables subsequent temporal analyses of the change in surface water over all acquisitions for a location. GIW and GLOWABO share similar issues to WOfS where water mixes with vegetation, where water areas are smaller than the Landsat pixel (although GLOWABO is enhanced using the panchromatic band to significantly enhance its resolution) and where terrain and cloud shadows remain in the data. A full comparison between these datasets is yet to be undertaken, and is a subject for a future article.

As of April 2014 the WOfS product has been delivered as a consistent continental dataset via web services, accessible at http://eos.ga.gov.au/geoserver. The services make up a suite of five layers presenting the various components of WOfS, allowing easy public access. The available layers are:

1. number of times surface water observed,
2. number of clear observations,
3. the ratio of water observations to clear observations as a percentage,
4. confidence in the water observations,
5. filtering of the water recurrence summary for confidence.

The fifth layer shown in Fig. 9, presenting the water summary filtered for confidence, represents the final state of this version of WOfS and displays the percentage of clear observations for which water was observed across Australia where the confidence value (κ) is at least one percent. Fig. 10 shows Landsat imagery for the Menindee Lakes region of western New South Wales and the associated WOfS outputs, demonstrating the presence of a wide range of permanent and ephemeral water bodies, including flood inundation. The individual water classifications from every Landsat scene are also available from the NCI at http://dap.nci.org.au.

Initial scoping of the full processing time required for the analysis indicated that one analysis of the entire Landsat archive for surface water was over four years. The analysis as conducted on the AGDC was completed in under 8 h, making it feasible to review and improve the algorithms, and repeat the analyses many times, where previously such an analysis was essentially not feasible.

5. Conclusion

The Water Observations from Space (WOfS) product provides a nationally consistent tool for understanding surface water across Australia both spatially and temporally. The maps generated using WOfS provide a new source of information on Australian floodplains, and a rich new data source for visualisation and analysis of surface water in Australia more generally. It demonstrates the power of high performance computing for remote sensing applications and the advantages of having well structured and standardised data in an accessible, high processing speed environment. The combination of large high speed storage attached to supercomputer processors, and surface reflectance data in a standard grid arrangement has enabled the development of a single analysis that can be applied systematically through decades of data.

We have demonstrated a method to develop a standard algorithm for the classification of surface water from medium resolution satellite imagery at a continental scale for decades of data. In total the process constituted some 184,500 scenes spanning over 27 years, or approximately 2×10^{13} observations, and was able to be completed in under eight hours to produce a continental map of surface water recurrence. The method shows that it is operationally feasible to apply a single algorithm over many different environmental and climatic conditions and achieve a high degree of accuracy. Some known errors remain, mainly derived from shadow produced by clouds and steep terrain. Important additional errors also derive from data anomalies (especially from Landsat-7 SLC-Off data) and urban structures. However the logistic regression method provides a mechanism of understanding the nature of the results and mitigating these errors.

WOfS is just one of many potential types of continental-scale analyses using medium resolution satellite data. Our ongoing work is focusing on the relationships between surface water and groundwater recharge, vegetation response and the dynamics of Australia's river systems. Similarly we are investigating forest and rangeland dynamics under the same workflow concepts. Our expectation is that the WOfS product itself will mature to become a regularly updated product. Public access to the water summary product is available through the Australian Flood Risk Information Portal (www.ga.gov.au/wofs).

Acknowledgements

The authors thank Dr. Brendan Brooke, Dr. Jane Sexton and Stephen Sagar, of Geoscience Australia, for their assistance in reviewing this paper.

This research was undertaken with the assistance of resources from the National Computational Infrastructure (NCI), which is supported by the Australian Government.

This paper is published with the permission of the CEO, Geoscience Australia.

References

Arvidson, T., Gasch, J., & Goward, S. N. (2001). Landsat 7's long-term acquisition plan—An innovative approach to building a global imagery archive. *Remote Sensing of Environment, 78*(1), 13–26.

Atkinson, R., Power, R., Lemon, D., O'Hagan, R., Dovey, D., & Kinny, D. (2008). *The Australian hydrological geospatial fabric—Development methodology and conceptual architecture.* CSIRO Water for a Healthy Country.

Australian Bureau of Statistics (2012). Australian statistical geography standard 2011. http://abs.gov.au/AUSSTATS/abs@.nsf/Lookup/1270.0.55.004Main+Features1July%202011?OpenDocument Accessed: 2014–12-09

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*(7), 1145–1159.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees.* CRC press.

Cesa-Bianchi, N., Lugosi, G., et al. (2006). *Prediction, learning, and games. 1*, Cambridge University Press Cambridge.

Dekker, A., Hoogenboom, H., Goddijn, L., & Malthus, T. (1997). The relation between inherent optical properties and reflectance spectra in turbid inland waters. *Remote Sensing Reviews, 15*(1–4), 59–74.

Draeger, W. C., Holm, T. M., Lauer, D. T., & Thompson, R. J. (1997). The availability of landsat data: past, present, and future. *Photogrammetric Engineering and Remote Sensing, 63*(7), 869–875.

Feng, M., Sexton, J. O., Channan, S., & Townshend, J. R. (2015). A global, high-resolution (30-m) inland water body dataset for 2000: first results of a topographic-spectral classification algorithm. *International Journal of Digital Earth.* http://dx.doi.org/10.1080/17538947.2015.1026420.

Frazier, P. S., Page, K. J., et al. (2000). Water body detection and delineation with landsat tm data. *Photogrammetric Engineering and Remote Sensing, 66*(12), 1461–1468.

Frazier, P., Page, K., Louis, J., Briggs, S., & Robertson, A. (2003). Relating wetland inundation to river flow using landsat tm data. *International Journal of Remote Sensing, 24*(19), 3755–3770.

Gallant, J. C., & Dowling, T. I. (2003). A multiresolution index of valley bottom flatness for mapping depositional areas. *Water Resources Research, 39*(12).

Gallant, J., Read, A., Dowling, T., Sims, J., Merrin, L., Ackland, R., & Herron, N. (2011). Building the national one-second digital elevation model for Australia. *Proceedings, Water Information Research and Development Alliance Science Symposium.*

Gao, B. -C. (1996). Ndwi—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment*, *58*(3), 257–266.
Guerschman, J. P., Warren, G., Byrne, G., Lymburner, L., Mueller, N., & Van Dijk, A. (2011). *Modis-based standing water detection for flood and large reservoir mapping: algorithm development and applications for the Australian continent.*
Gutman, G., Byrnes, R., Masek, I., Covington, S., Justice, C., Franks, S., & Headley, R. (2008). Towards monitoring land-cover and land-use changes at a global scale: The global land survey 2005. *Photogrammetric Engineering and Remote Sensing*, *74*(1), 6–10.
Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., & Tibshirani, R. (2009). *The elements of statistical learning. 2*, Springer.
Irish, R. R. (2000). Landsat 7 automatic cloud cover assessment. In *AeroSense* 2000, Pages 348–355. International Society for Optics and Photonics.
Khawlie, M., Shaban, A., Abdallah, C., Darwish, T., & Kawass, I. (2005). Watershed characteristics, land use and fabric: The application of remote sensing and geographical information systems. *Lakes & Reservoirs: Research & Management*, *10*(2), 85–92.
Kingsford, R. T. (2000). Ecological impacts of dams, water diversions and river management on floodplain wetlands in Australia. *Austral Ecology*, *25*(2), 109–127.
Li, F., Jupp, D. L., Reddy, S., Lymburner, L., Mueller, N., Tan, P., & Islam, A. (2010). An evaluation of the use of atmospheric and brdf correction to standardize landsat data. Selected Topics in Applied Earth Observations and Remote Sensing. *Journal of IEEE*, *3*(3), 257–270.
Li, F., Jupp, D. L., Thankappan, M., Lymburner, L., Mueller, N., Lewis, A., & Held, A. (2012). A physics-based atmospheric and brdf correction for landsat data over mountainous terrain. *Remote Sensing of Environment*, *124*, 756–770.
Markham, B. L., Storey, J. C., Williams, D. L., & Irons, J. R. (2004). Landsat sensor performance: history and current status. *IEEE Transactions on Geoscience and Remote Sensing*, *42*(12), 2691–2694.
Morse, A., Zariello, T. J., & Kramber, W. J. (1990). Using remote sensing and gis technology to help adjudicate idaho water rights. *Photogrammetric Engineering and Remote Sensing*, *56*(3), 365–370.
Nichol, J., & Vohora, V. (2004). Noise over water surfaces in landsat tm images. *International Journal of Remote Sensing*, *25*(11), 2087–2093.
Purss, M., Lewis, A., Edberg, R., Ip, A., Sixsmith, J., Frankish, G., ... Hurst, L. (2013). Exploiting data intensive applications on high performance computers to unlock Australia's landsat archive. *EGU General Assembly Conference Abstracts. 15*. (pp. 8049).
Robinove, C. J. (1978). Interpretation of a landsat image of an unusual flood phenomenon in Australia. *Remote Sensing of Environment*, *7*(3), 219–225.
Shackelford, A. K., & Davis, C. H. (2003). A hierarchical fuzzy classification approach for high-resolution multispectral data over urban areas. *IEEE Transactions on Geoscience and Remote Sensing*, *41*(9), 1920–1932.
Shan, J., Hussain, E., Kim, K., & Biehl, L. (2009). Flood mapping and damage assessment—A case study in the state of Indiana. *Geospatial Technology for Earth Observation* (pp. 473–495). Springer.
Sixsmith, J., Oliver, S., & Lymburner, L. (2013). A hybrid approach to automated landsat pixel quality. *Geoscience and Remote Sensing Symposium (IGARSS), 2013 IEEE International* (pp. 4146–4149).
Smith, L. C. (1997). Satellite remote sensing of river inundation area, stage, and discharge: A review. *Hydrological Processes*, *11*(10), 1427–1439.
Thomas, R. F., Kingsford, R. T., Lu, Y., & Hunter, S. J. (2011). Landsat mapping of annual inundation (1979–2006) of the macquarie marshes in semi-arid Australia. *International Journal of Remote Sensing*, *32*(16), 4545–4569.
Trimble Geospatial Imaging (2014). Ecognition suite. http://www.ecognition.com (Accessed: 2014–12-12)
Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, *8*(2), 127–150.
Tulbure, M. G., & Broich, M. (2013). Spatiotemporal dynamic of surface water bodies using landsat time-series data from 1999 to 2011. *ISPRS Journal of Photogrammetry and Remote Sensing*, *79*, 44–52.
United States Geological Survey (2014). Web enabled landsat data. https://landsat.usgs.gov/WELD.php (Accessed: 2014–12-04)
Verpoorter, C., Kutser, T., Seekell, D. A., & Tranvik, L. J. (2014). A global inventory of lakes based on high-resolution satellite imagery. *Geophysical Research Letters*, *41*(18), 6396–6402 2014GL060641.
Vörösmarty, C. J., Green, P., Salisbury, J., & Lammers, R. B. (2000). Global water resources: Vulnerability from climate change and population growth. *Science*, *289*(5477), 284–288.
Wang, Y. (2004). Using landsat 7 tm data acquired days after a flood event to delineate the maximum flood extent on a coastal floodplain. *International Journal of Remote Sensing*, *25*(5), 959–974.
Zhu, Z., & Woodcock, C. E. (2012). Object-based cloud and cloud shadow detection in landsat imagery. *Remote Sensing of Environment*, *118*, 83–94.