EARTH SURFACE PROCESSES AND LANDFORMS
Earth Surf. Process. Landforms **45**, 3120–3140 (2020)

Published online 18 August 2020 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/esp.4955

UAV-based training for fully fuzzy classification of Sentinel-2 fluvial scenes

P. E. Carbonneau,[1]* B. Belletti,[2,3] M. Micotti,[2] B. Lastoria,[4] M. Casaioli,[4] S. Mariani,[4] G. Marchetti[2,5] and S. Bizzi[6]
[1] Department of Geography, Durham University, Durham, UK
[2] Department of Electronics, Information and Bioengineering, Polytechnic University of Milan, Milan, Italy
[3] CNRS UMR5600-EVS, University of Lyon, Lyon, France
[4] Water Protection Department, Istituto Superiore per la Protezione e la Ricerca Ambientale (ISPRA), Rome, Italy
[5] Faculty of Science and Technology, Free University of Bozen-Bolzano, Bolzano, Italy
[6] Department of Geosciences, University of Padova, Padova, Italy

Received 17 May 2020; Revised 2 July 2020; Accepted 3 July 2020

*Correspondence to: P. E. Carbonneau, Department of Geography, Durham University, Durham, UK. E-mail: patrice.carbonneau@durham.ac.uk

ESPL
Earth Surface Processes and Landforms

ABSTRACT: An estimated 76% of global stream area is occupied by channels with widths above 30 m. Sentinel-2 imagery with resolutions of 10 m could supply information about the composition of river corridors at national and global scales. Fuzzy classification models that infer sub-pixel composition could further be used to compensate for small channel widths imaged at 10 m of spatial resolution. A major challenge to this approach is the acquisition of suitable training data useable in machine learning models that can predict land-cover type information from image radiance values. In this contribution, we present a method which combines unmanned aerial vehicles (UAVs) and Sentinel-2 imagery in order to develop a fuzzy classification approach capable of large-scale investigations. Our approach uses hyperspatial UAV imagery in order to derive high-resolution class information that can be used to train fuzzy classification models for Sentinel-2 data where all bands are super-resolved to a spatial resolution of 10 m. We use a multi-temporal UAV dataset covering an area of 5.25 km^2. Using a novel convolutional neural network (CNN) classifier, we predict sub-pixel membership for Sentinel-2 pixels in the fluvial corridor as divided into classes of water, vegetation and dry sediment. Our CNN model can predict fuzzy class memberships with median errors from −5% to +3% and mean absolute errors from 10% to 20%. We also show that our CNN fuzzy predictor can be used to predict crisp classes with accuracies from 95.5% to 99.9%. Finally, we use an example to show how a fuzzy CNN model trained with localized UAV data can be applied to longer channel reaches and detect new vegetation growth. We therefore argue that the novel use of UAVs as field validation tools for freely available satellite data can bridge the scale gap between local and regional fluvial studies. © 2020 The Authors. Earth Surface Processes and Landforms published by John Wiley & Sons Ltd

KEYWORDS: machine learning; UAV; fuzzy supervised classification; Sentinel-2; super-resolution; fluvial environments

Introduction

Fluvial remote sensing at sub-metric resolutions has been the focus of a significant body of research in recent years (e.g. Marcus and Fonstad, 2008, 2010; Carbonneau *et al.*, 2012; Piégay *et al.*, 2012, 2020; Bizzi *et al.*, 2016; Dugdale *et al.*, 2019). Many of these ideas were inspired by work arguing for an examination of rivers at very high resolutions and over very large extents (e.g. Vannote *et al.*, 1980; Fausch *et al.*, 2002). In this context, it has been argued that very high-resolution imagery, sometimes called hyperspatial imagery (Carbonneau and Piégay, 2012), could allow for a process-based analysis of image data that could be used to advance fundamental ideas in fluvial ecology and geomorphology (Carbonneau *et al.*, 2012). However, such process-focussed monitoring over regional, national or continental scales remains largely out of reach. Despite progress in manned (e.g. Carbonneau *et al.*, 2004; Dugdale *et al.*, 2013, 2015; Frechette *et al.*, 2018) or unmanned image acquisitions (e.g. Tamminga *et al.*, 2015; Woodget *et al.*, 2015; Carbonneau *et al.*, 2018), image data for entire catchments and nations is currently constrained to multi-year surveys done by national and/or regional environmental agencies and, at best, delivers imagery with a spatial resolution on the order of 0.5 m and temporal resolutions of several years or decades. Sub-metric spatial resolution imagery acquired at daily or weekly temporal frequency is available in the commercial realm but purchasing costs for entire nations or continents are beyond the most generous budgets. A potential alternative is the use of public domain multispectral satellite data such as the European Union (EU) Copernicus Sentinel-2 or the joint NASA/USGS Landsats 7 and 8. These offer a reasonable temporal repeat frequency but at much lower spatial resolutions. Sentinel-2 has four bands acquired natively at a spatial resolution of 10 m. Downing

et al. (2012) estimate that mean width of fifth-order streams is roughly 30 m and that streams with orders 5–12 (the highest order of the Amazon River) occupy 76% of global stream area. Sentinel-2 data should therefore at least be capable of detecting such streams and thus be sensitive to most rivers of the world. However, it is clear that medium to small streams (dimensions below ~100 m) will have few pixel samples that will leave a coarse digital representation of these complex landscapes. Fuzzy classification is a well researched topic that aims to infer sub-pixel scale compositions by assigning to each pixel a membership percentages for each available class (Foody and Cox, 1994; Foody, 1997; Foody *et al*., 1997; Zhang and Foody, 2001; Ling *et al*., 2019). Fuzzy classification approaches could therefore be a way to mitigate for the relatively low spatial resolution of Sentinel-2 data. Supervised classification, of either the fuzzy or crisp (i.e. assigning a single integer class number to each land-cover type representing the semantic class of a landform) variety, of large-scale river corridors does pose logistic challenges. Traditionally, ground-truth played a critical part in supervised classification of remotely sensed imagery (Curran and Williamson, 1985; Steven, 1987). However, the obvious logistic implications of large-scale ground-truth data acquisition has somewhat weakened this practice and we find an increasing usage of on-screen image interpretation as the basis for the production of training areas for supervised classification algorithms (e.g. Thanh Noi and Kappas, 2018). Low-cost drones, unmanned aerial vehicles (UAVs), might provide a cost-effective approach to the acquisition of ground-truth data for satellite data analysis. Carbonneau *et al*. (2018) have demonstrated that low-altitude imagery could be used as ground-truth data for grain size mapping algorithms. Similar to this idea, this article proposes to use UAV imagery in order to derive suitable training data for fuzzy classification algorithms applicable to Sentinel-2 imagery.

Using one of the largest UAV datasets in the published literature, we have developed a novel convolutional neural network (CNN) fuzzy classification algorithm tailored to Sentinel-2 imagery. We focus mainly on the fuzzy classification of river corridors and consider three end-member classes: water, vegetation and dry exposed sediment. In a series of experimental scenarios, we show that our CNN fuzzy classifier can predict the membership percentage of the dominant class, i.e. the class with the highest membership, for each Sentinel-2 pixel of a river corridor with median errors ranging from −5.5% to 2.4% and mean absolute errors ranging from 14.2% to 20.7%. In the case of the sub-dominant class, i.e. the class with the second highest membership, median errors range from 0.7% to 2.5% and mean absolute errors range from 10.9% to 17.9%. Also, we show that if we use our fuzzy CNN model to predict 'crisp' class (i.e. the semantic class: water, vegetation or dry exposed sediment), we can reach accuracies of 95.5% to 99.9%. We show that performance at this level is possible in a range of scenarios. CNN models trained with UAV-derived labels from a given year perform well on imagery acquired in the previous year. CNN models trained with UAV-derived labels from two rivers can be satisfactorily transferred to two new rivers, imaged on two separate Sentinel-2 tiles. Finally, CNN models trained on one part of a given river perform well when classifying other parts of the same river. Furthermore, we have tested comparator methods such as linear unmixing and dense neural network (DNN) fuzzy classifiers and have made two important findings. First, models that are trained with UAV-derived label data always perform better than models trained without the benefit of field data. Second, linear unmixing and DNN fuzzy predictors cannot match the performance of our novel CNN approach. As a final demonstration of the potential of our novel approach, we give the reader access to fuzzy classifications for 294 linear kilometres along the river corridors of our study rivers. We use a smaller portion of this data to show how fuzzy classification can be used to monitor sub-pixel vegetation growth and establish net change over a one year period. The methods developed here therefore deliver a successful integration of UAV and Satellite data and provide a pragmatic way forward for cost-effective, large-scale studies of fluvial systems.

Methods

Drone acquisitions and Sentinel-2 imagery

Ground-truth data are derived from 16 UAV surveys carried out in Italy using a DJI Phantom 4 Pro drone. The data were acquired during 2017 and 2018 and spread across eight sites located on four rivers. In northern Italy, imagery was acquired for two sites on the River Sesia, near the towns of Arborio and Caresana. The River Sesia starts in the Alpine foothills as an island-braiding channel and then, nearer its confluence with the Po, evolves into a single thread meandering channel. For the River Po, we have one sampling site. The River Po is the longest river in Italy and consists of a single-thread channel with local wandering. In central Italy, imagery was acquired for three sites on the River Paglia in Umbria. This river is a small single-thread channel with rare occurrences of localized wandering. In southern Italy, imagery was acquired for two sites on the River Bonamico in Calabria. This is a medium river with a high sediment load and active braiding. Figure 1 shows all the rivers used in this study within a national context. Each site, on each river, has one repeat survey. For each acquisition, we use 15–20 ground targets surveyed to centimetre-accuracy with a Trimble R10 RTK-GPS (real-time kinematic global positioning system) deriving its differential correction from mobile services. The images were acquired at 80% forward overlap and 50% sidelap and the flight patterns also included oblique views and multiple altitudes as recommended in Carbonneau and Dietrich (2017). Photogrammetric processing was accomplished with Agisoft Metashape. In relation to the work presented here, the primary outputs were orthomosaic images with a spatial resolution of 10 cm. If we consider repeat visits as separate acquisitions, the total area covered by the UAV surveys is 5.25 km^2. This is one of the largest UAV data acquisitions reported in the published literature. Typically, UAV data are collected for areas near or below 1 km^2 (e.g. de Haas *et al*., 2014; Tamminga *et al*., 2015; Woodget *et al*., 2015; Lindner *et al*., 2016; Rossini *et al*., 2018; Rusnák *et al*., 2018). At the upper limit, Immerzeel *et al*. (2014) report a survey area of 7.96 km^2 on a glacier surface performed with a fixed wing UAV capable of long-range flights. To our knowledge, the data used in this work represents the largest UAV survey area reported in fluvial remote sensing. We do not find other published works with repeat UAV surveys of eight sites located on four different rivers.

In addition to the UAV imagery, Sentinel-2 data was downloaded from the European Space Agency (ESA) Open-Access Copernicus hub. First, we identify suitable cloud free imagery, nearest in time within a maximum 15 days of the drone acquisition. Second, we verify the discharge from the nearest upstream gauging station and precipitation records in order to avoid rain three days prior to acquisition (dry land) and to control hydraulic conditions. Once these conditions are met, we download the full Sentinel-2 tiles at level 2A (with full atmospheric correction). In total, eight tiles were needed to match all the UAV surveys. Within these tiles, we discard Sentinel-2 bands designed to sample atmospheric quantities

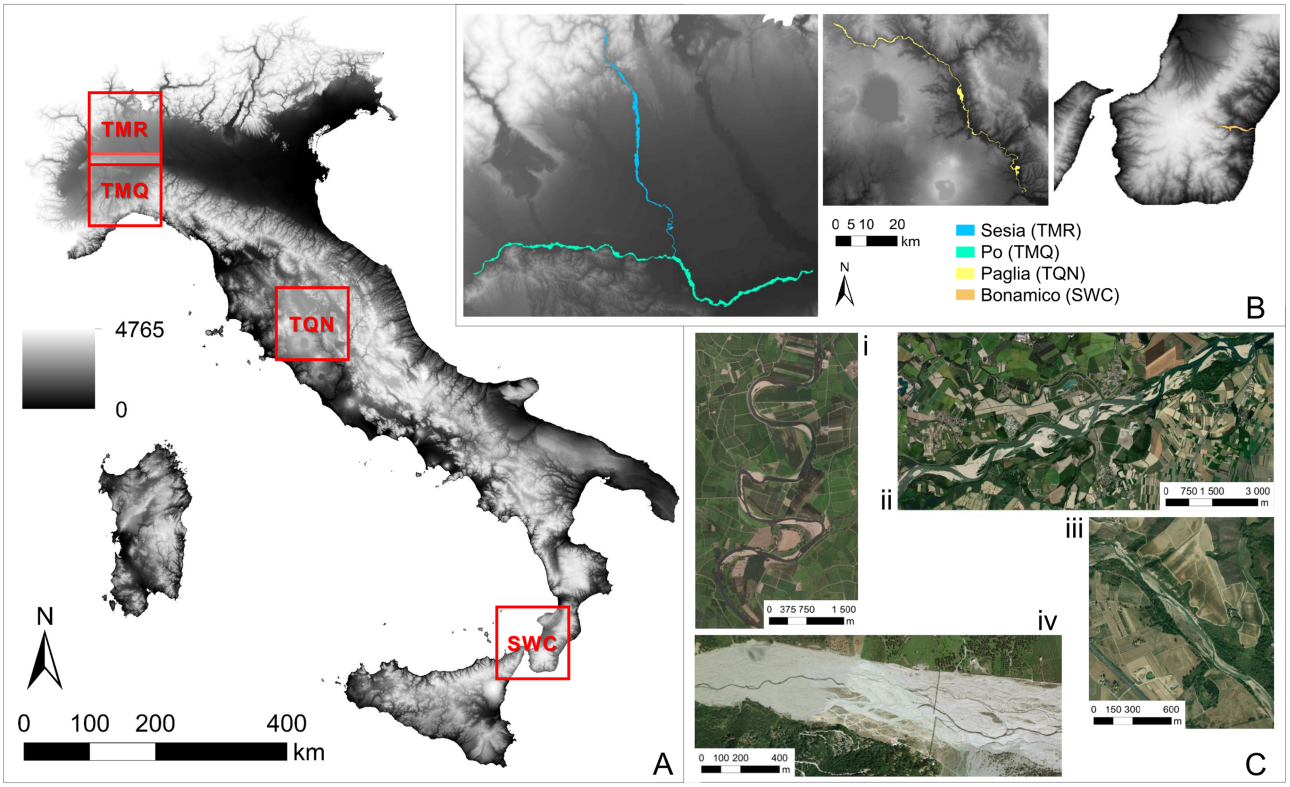

FIGURE 1. (A) Location of the Sentinel-2 tiles. Sesia River sites are located in tile TMR. Po River sites are located in tile TMQ. Paglia sites are located in tile TQN and Bonamico sites are located in tile SWC. (B) River corridors within regional topographic settings. (C) Google Earth imagery of the sites showing the diversity of river corridors. i, River Sesia; ii, River Po; iii, River Paglia; iv, River Bonamico. [Colour figure can be viewed at wileyonlinelibrary.com]

and we will only use bands 2, 3, 4, 5, 6, 7, 8, 8A, 11 and 12 (total of 10 bands). In terms of area, the UAV surveys overlap a total of 52543 Sentinel-2 pixels (@ 10m spatial resolution) which form our raw Sentinel-2 samples. Figure 1 shows the location of the study sites and associated Sentinel-2 tiles within Italy. Table 1 gives a summary of the locations and dates of the primary data acquisition.

Fuzzy label production

The core idea behind this work is to use centimetre-scale resolution UAV data in order to generate high quality training data for machine learning models. We therefore use a manual object-based image analysis (OBIA) approach to derive high quality land-cover classifications for the UAV orthoimagery that will be used to provide fuzzy training labels for our models. The manual OBIA classification applied to the 10-cm resolution UAV imagery was capable of identifying and discriminating between spatial units with similar textural and spectral characteristics that are known to constitute key geomorphic macro-units as described by Belletti *et al.* (2017): water, vegetation or dry exposed sediment. Such units represent the coarse assemblage, the external envelop, of geomorphic units of the same type (e.g. in a meandering river, a 'dry exposed sediment' macro-unit can include a 'point bar' and a dry 'chute off

Table 1. Drone and Sentinel-2 acquisition sites and dates. We use UAV data from 16 acquisitions in 2017 and 2018. The total area of the UAV surveys was 5.25km^2 sampling a total of 52543 Sentinel-2 pixels

Site (River, location)	UAV acquisition date	Sentinel-2 acquisition date	Sentinel-2 tile
Po, Nicorvo	16 September 2017	24 September 2017	TMQ
Paglia, Acquapendente	20 September 2017	21 September 2017	TQN
Paglia, Allerona	18 September 2017	21 September 2017	TQN
Paglia, Orvieto	19 September 2017	21 September 2017	TQN
Bonamico, upstream	15 November 2017	21 November 2017	SWC
Bonamico, downstream	16 November 2017	21 November 2017	SWC
Sesia, Arborio	18 April 2018	17 April 2018	TMR
Sesia, Caresana	16 April 2018	17 April 2018	TMR
Sesia, Arborio	21 September 2018	24 September 2018	TMR
Sesia, Caresana	22 September 2018	24 September 2018	TMR
Po, Nicorvo	20 September 2018	24 September 2018	TMQ
Paglia, Acquapendente	17 July 2018	20 July 2018	TQN
Paglia, Allerona	16 July 2018	20 July 2018	TQN
Paglia, Orvieto	17 July 2018	20 July 2018	TQN
Bonamico, upstream	24 October 2018	22 October 2018	SWC
Bonamico, downstream	23 October 2018	22 October 2018	SWC

channel' geomorphic units). Readers should note that from here onwards, we will refer to the 'dry exposed sediment' class merely as 'sediment'. For each orthoimage, we used the i.segment routine included with GRASS GIS 7.6 in order to segment the RGB data into spatially contiguous groups with common brightness characteristics. Trial and error was used to set the i.segment parameters in order to create relatively small but not too fragmented and consistent objects, i.e. belonging to the same class (i.e. water *versus* vegetation *versus* sediment). Parameter values for *minsize* and *threshold*, were set between 4000 and 8000 and between 0.3 and 0.85, respectively. This step allowed a first and objective delineation of riverine objects. Once segmented, the UAV orthoimages were classified by geographic information system (GIS) photointerpretation at a scale between 1:400 and 1:2000. A further manual modification of the form of some segments was occasionally performed to improve their delineation during the photointerpretation stage (e.g. to separate vegetation from water or sediment from water). Each segment was then attributed one of the three classes described earlier. All the labelled objects of a same class were then merged into semantic classes and the result exported in class raster format where each pixel holds the value of the land-cover class. Figures 2 and 3 show detailed examples of this classification process.

This manual OBIA process was extremely labour-intensive (1.5 days per orthoimage on average) but it has allowed for 10 cm resolution classifications which we will approximate as being error-free. We obviously recognize that such a process driven by human interpretation cannot be error-free. One challenge is the assignment of all features of the riverine landscape to one of only three classes. Using three classes is a design decision and we did consider using extra classes such as senescent vegetation. However, this feature is present as a minority in the data and experience with CNN training suggests that severely under-sampled classes do not produce a satisfactory response during CNN training and such classes are better merged with the most appropriate similar class. For example, on the bottom right of Figure 3, we see an area of vegetation that is a composite of dry senescent grasses and small (fresh/green) trees. Obviously the reflectance properties of senescent vegetation are not identical to those of fresh vegetation, but the vegetation class remains the most appropriate for senescent vegetation. However, we argue that such errors will have a negligible outcome on our results. First, we have classified in

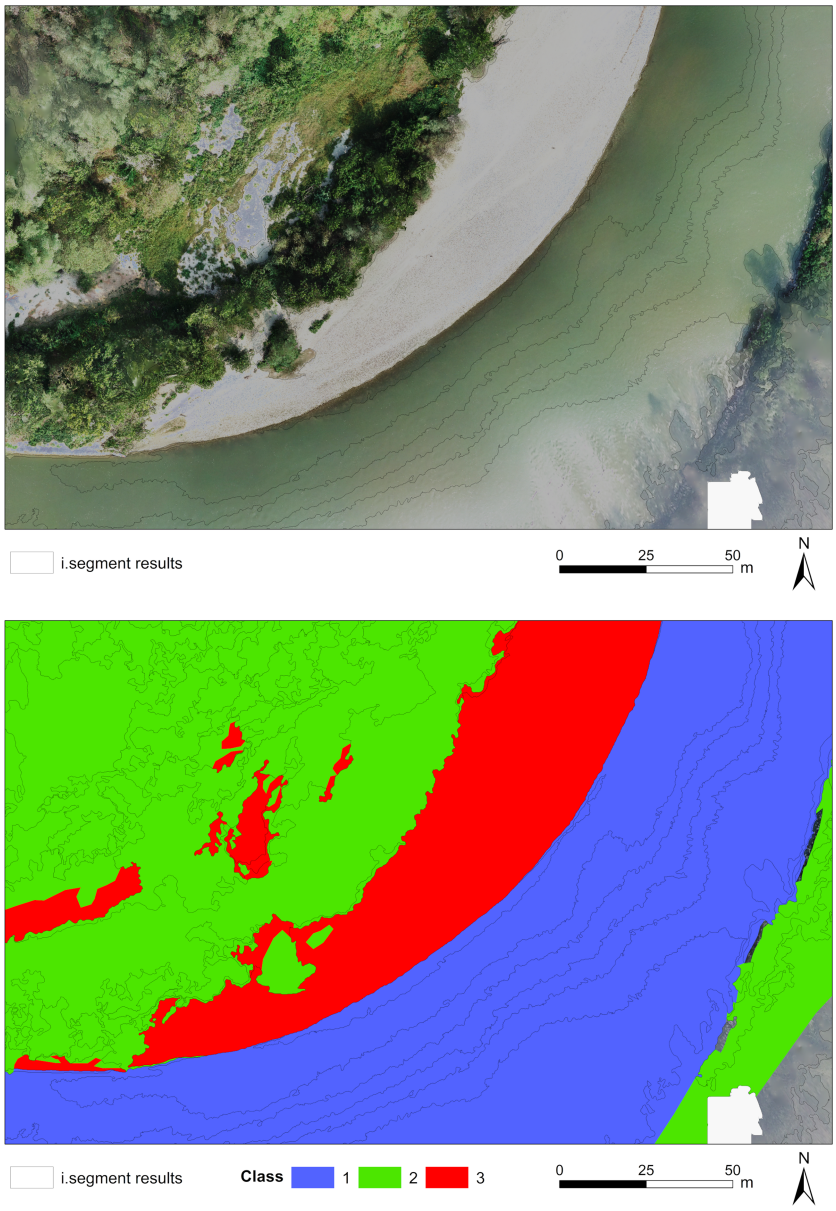

FIGURE 2. Example 1 of UAV image classification outputs for the Sesia Caresana site. [Colour figure can be viewed at wileyonlinelibrary.com]

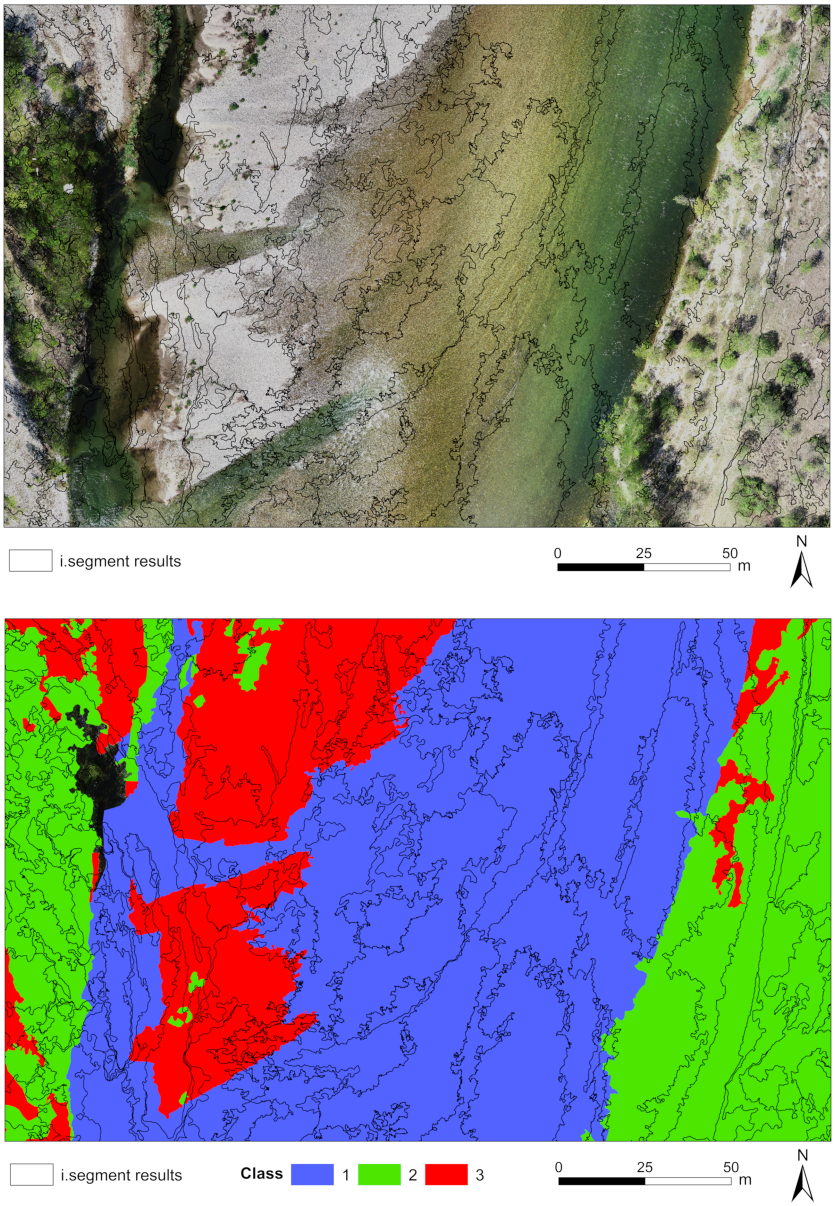

FIGURE 3. Example 2 of UAV classification outputs for the Sesia Arborio site. [Colour figure can be viewed at wileyonlinelibrary.com]

excess of 525 million UAV pixels with this method and we approximate that any classification errors leading to some confusion as to which brightness levels can be associated to a given class will average out over this large sample. Second, we intend to use this data to train deep neural networks and these have been found to be very robust to noise in the training data (Rolnick *et al.*, 2018).

Sentinel-2 image super-resolution

In order to extend the reach of this method to the smallest possible rivers, Sentinel-2 bands with 10m resolution are most relevant. Since only four Sentinel-2 bands were natively acquired at 10m, we must consider a pan-sharpening or super-resolution approach. Given that Sentinel-2 has no panchromatic band, traditional pan-sharpening methods are not appropriate (Brodu, 2017). However, a range of new methods are emerging, some of which are based on deep learning (e.g. Lanaras *et al.*, 2018; Gargiulo *et al.*, 2019). Close inspection of these methods reveals either that source code is not publicly available, or the methods are not quite integrated into a normal spaceborne remote sensing workflow that includes atmospheric correction. In our application, full compatibility with atmospheric correction is essential since we will rely on the Sen2Cor atmospheric processor used by the ESA to standardize brightness values across multiple Sentinel-2 tiles (Main-Knorn *et al.*, 2017). Brodu *et al.* (2017) describe a super-resolution method which is fully compatible with atmospherically corrected Sentinel-2 imagery. Their approach uses geometric features which are not band-dependent in order to meaningfully re-sample the 20m and 60m bands in Sentinel-2 data to 10m. This method is available as the Sen2Res plugin for the ESA SNAP open-source software designed for the processing and analysis of ESA products. The plugin delivers all 12 image bands stacked in a single tif file. However, in this work it was decided not to use bands 1 and 10 which were designed to detect atmospheric quantities and work a maximum of 10 bands. Pixels in the tif file have values normalized from 0 to 1 which is convenient for the modelling steps that will follow. At this stage, the only obvious drawback of this super-resolution approach is computational cost, it took approximately six hours to produce super-resolved bands cropped to our 16 survey areas. For the production of the larger

scale data (i.e. the full river corridor for all our sampled rivers), we needed to invest over 50 hours of computing time. The computational cost of this super-resolution method is therefore not trivial and its necessity is evaluated in this work. However, the outputs of the process do appear to be very satisfactory, if computationally expensive (Figure 4).

Data preparation

We begin this step by assessing the co-registration between the satellite and drone data. This task is not straightforward. Normally, positional accuracy of satellite is assessed with relatively large stable, often man-made, features. No such features are available in the drone imagery. We therefore limit ourselves to a qualitative check that the landform contours in both drone and Satellite imagery overlap in the same Sentinel-2 pixel. However we note that even this determination is not simple as partial occupancy pixels can raise questions as to the exact location of a boundary in the Sentinel-2 data. However, a validation of the UAV-derived orthoimage with additional data from the RTK-GPS gives an average positional accuracy for the drone orthoimagery of 0.09 m. Co-registration was therefore deemed adequate.

In order to compare the classification performance of models trained with UAV-derived models to those trained without field data, we first generate a set of so-called 'desk-based' polygons. In this case we use photo-interpretation of the Sentinel-2 images themselves in order to delineate areas falling within our three classes. This is simply done in QGIS 3.4 by displaying the Sentinel-2 image in false-colour with an infrared component and using a polygon shapefile to manually digitize class samples. These vector class samples are then rasterized to outputs where each pixel holds the class value. Areas that were not digitized are coded as 0 and will eventually be ignored. We will refer to this data as the *desk-based* label raster. After this step, we organize the data by cropping the UAV-derived labels, the Sentinel-2 imagery and the desk-based class rasters to small areas around the study sites. Therefore for each of the 16 UAV acquisitions, we have three small georeferenced rasters: a Sentinel-2 sub-image (10 bands), a UAV-derived label raster (1 band) and a desk-based label raster (1 band). Fuzzy membership class values for each Sentinel-2 pixel can now be determined by extracting the map coordinates of the Upper Left (UL) and Lower Right (LR) corners of each Sentinel-2 pixel. Within this bounding box, we extract the associated 10000 classified pixels from the 10 cm ×10 cm UAV-derived labels and then calculate membership percentage for each class in each Sentinel-2 pixel.

Finally, the data is prepared for machine learning. For readers less familiar with machine learning, we recommend Burkov (2019) as a concise text leading to more advanced texts (e.g. Goodfellow *et al.*, 2016; Chollet, 2017). The models used here will require two types of label data: crisp semantic classes and fuzzy classes. We aim to use CNNs as our main classification algorithm. CNNs have a huge range of applications across all fields of science (e.g. language processing, time series, video processing, biochemistry, etc.). In this work we intend to use CNN models specific to 'multiband' imagery (i.e. multiple two-dimensional (2D) rasters repeated for several spectral bands), we will therefore format our GIS data as four-dimensional (4D) tensors immediately ready for CNN processing. The tensors used for CNN processing are 4D digital objects that store multiple images, usually having several bands, in a single digital object. Video files are an example of a tensor as they store multiple static RGB images together and display them according to a fourth dimension, time. In the case of CNN data, the fourth dimension is merely an indexed list of

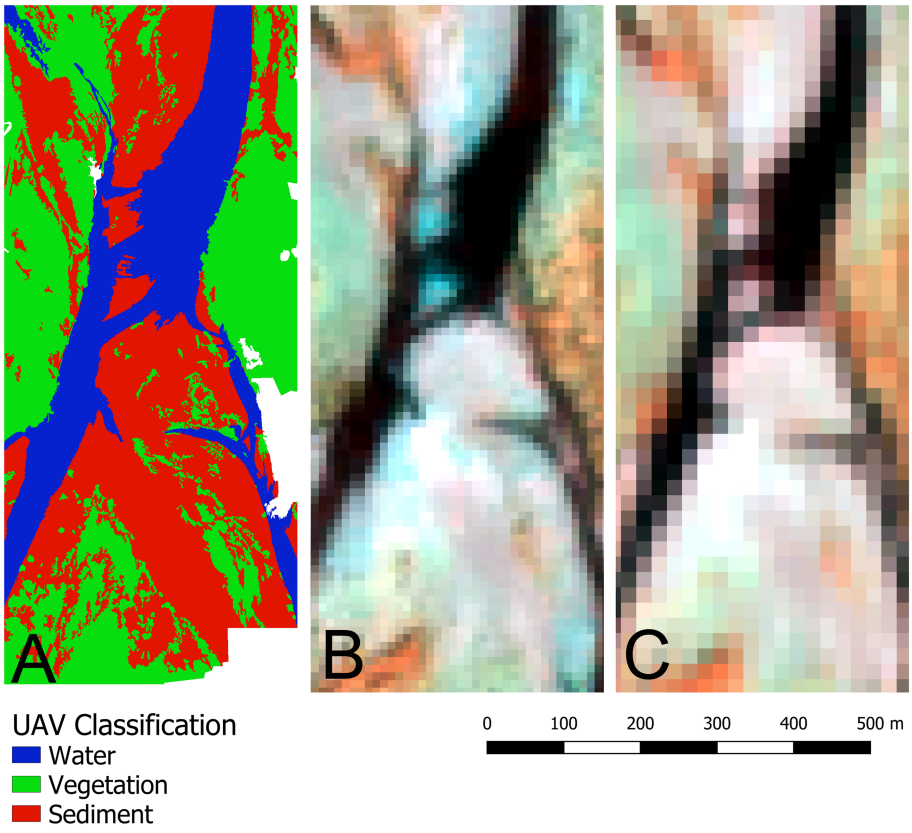

FIGURE 4. Example of (A) UAV-derived label data and corresponding Sentinel-2 imagery at (B) 10 m of spatial resolution and (C) 20 m of spatial resolution. [Colour figure can be viewed at wileyonlinelibrary.com]

images and there is no implication of time. The tensor format is therefore a convenient way of stacking multiple images in a single digital file. In the case of large images, CNN can be used to model local information such as semantic class (e.g. Buscombe and Ritchie, 2018). However, such image-specific CNN models usually require the use of a small tile which gives a local sub-sample of the image. In our case of sub-pixel fuzzy classification, we will initially assume that the exact size of this local region is small and will determine the optimal size later. We therefore use Python to create and format our data as 4D tensors of small image tiles. We scan each Sentinel-2 image and for each pixel extract an (S, S, 10) sub-image tile where S is the size of the tile, in pixels, and 10 is the number of bands. We only use odd numbers for S in order to have a unique central pixel for each extracted image tile. These tiles are compiled as tensors of dimension (Ns, S, S, 10) where Ns is the number of samples, i.e. the number of extracted tiles. In parallel to the tile extraction, we extract the class information associated to the central pixel of each tile. In the case of crisp classes, we have two options: we can use the desk-based labels or we can convert the UAV-derived fuzzy labels. Desk-based labels are already in crisp format. For the UAV-fuzzy labels used as crisp classes, we keep only those pixels with a pure class which we define as those pixels where the highest membership percentage is >95%. This class becomes the crisp class. Therefore for the crisp classes, each tile in the tensor has an associated integer value of 1 to 3 giving the class of the central pixel. This information is stored in a separate file. In the case of fuzzy classes, each tensor tile is associated to three floating point values giving the membership fraction (0–1) for each of three classes (water, vegetation, sediment) as sampled at the central pixel of the tile. This information is also stored in a separate file. In addition, we use a data augmentation procedure to increase our sample sizes. Data augmentation is common in deep learning and consists in creating new samples based on small modifications of existing samples (Chollet, 2017). In our case, for each sampled tile, we perform three rotations of 90°, 180° and 270°. After each rotation, we add a small amount of noise generated separately for each image pixel. For example, in the case of a 5 × 5 × 10 image tile, we generate 250 samples of noise. The noise values ranged 1E-4 to 1E-3 and are added to pixel values normalized from 0 to 1 by the super-resolution process. Each combination of rotation and noise addition is considered as a new sample. In the preliminary stages of this work, we found that the addition of noise was crucial to the numeric stability of model training process. The augmentation therefore allows us to generate four effective samples from each tile initially extracted from the Sentinel-2 imagery. Each of the four samples will have the same class labels, but the image tiles will be rotated and have very slightly different values. The results are a total of 210172 tile samples extracted from our UAV-derived labels and 79650 samples extracted from the desk-based labels. These augmented samples will be used to enhance the training of our models. However, when we validate these models, the augmentation will always be removed and we will only use raw data.

CNN model selection

We now select a CNN model for use in this work from a range of possible designs. We use the Python coding language and the Tensorflow library (Abadi *et al.*, 2015). The tile size mentioned earlier must be selected. In addition, there is a very long list of tunable parameters to be adjusted and architecture choices to be made. We tested a total of 102 candidate models that spanned a range of values for the tile size, the depth of the neural networks, the number of convolution filters (discussed later) and the optimal input bands to use from the possible 10 Sentinel-2 bands intended to sample surface characteristics. For brevity we only present summary findings here. The full analysis is available in the Supporting Information document accompanying this article. Our model selection procedure arrives at three conclusions. First, the use of 10 bands of super-resolved image data delivers performance improvements that are statistically significant when compared to the limited use of four bands natively acquired at 10 m of spatial resolution. We must therefore conclude that the computationally expensive super-resolution process is justified. Second, deeper neural network architectures with a larger number of trainable parameters did not deliver statistically significant performance improvements when compared to smaller networks. Third, a model using image tiles of 5×5 pixels across Sentinel-2 bands 2, 3, 4, 5, 6, 7, 8, 9, 11 and 12 with 32 convolution filters delivered best performance and was significantly better than the other candidate models (in the statistical sense). In our model selection analysis we also examine model performance as a function of the number of training epochs and find that our optimal model can be trained for 200 epochs with a learning rate of 5E-4 without overfitting (Burkov, 2019). Figure 5 (right) shows the final CNN model architecture. In total the model has 12931 trainable parameters. This number of trainable parameters is relatively low compared to other CNN models. The CNN models that are the basis of media headlines and set benchmarks for tasks such as facial recognition often have in excess of one million parameters. For example, Buscombe and Ritchie (2018) use the MobileNetV2 CNN which has in excess of 3.7 million trainable parameters. Given that our model is much smaller, we will refer to our model as a compact convolutional neural network (cCNN) as in Samarth *et al.* (2019). In parallel to our cCNN, we wish to test a non-convolutional DNN since such networks have been used in the past for fuzzy classification (Foody, 1997). We therefore use a similar architecture to our cCNN where the only difference is the removal of the 2D convolution layer (Figure 5, left). Our DNN model has 3362 trainable parameters. The DNN model will not use tensors as input. By construction, this model can only accept single pixel brightness values in the training process. We refer to this as a 'pixel-based' operation. This data can easily be extracted from the tensor format as the central pixel (in the *XY* plane) of the 4D tensor object. Interestingly, both these architectures can be used for crisp and fuzzy classification with minimal modifications. Neural networks are in fact inherently fuzzy classifiers. Any neural network designed for classification, of any type, will terminate in a layer that has as many nodes as classes in the label data. In a traditional crisp classification problem, a trained neural network will take a new sample and output a likelihood of class membership in each node. For example, in a three class problem, the network terminates with three nodes. When a sample returns the highest membership in node 2, the sample is predicted as class 2. In order to move between fuzzy and crisp classification we only need to change the activation function of the final layer to train the network with either integer (crisp) or float (fuzzy) numeric values and when final classes are attributed, fuzzy classification does not require the determination of the class from the highest membership and all three class membership predictions are saved. Both the DNN and cCNN are trained for 200 epochs with a low learning rate of 5E-4. We used an NVIDIA GTX-1070 GPU to accelerate training. Given the relatively low number of parameters in our models, training is completed in less than two minutes. We also tested the training duration on a smaller laptop with no graphics processing unit (GPU) acceleration and an i5 processor. This unit completed the 200

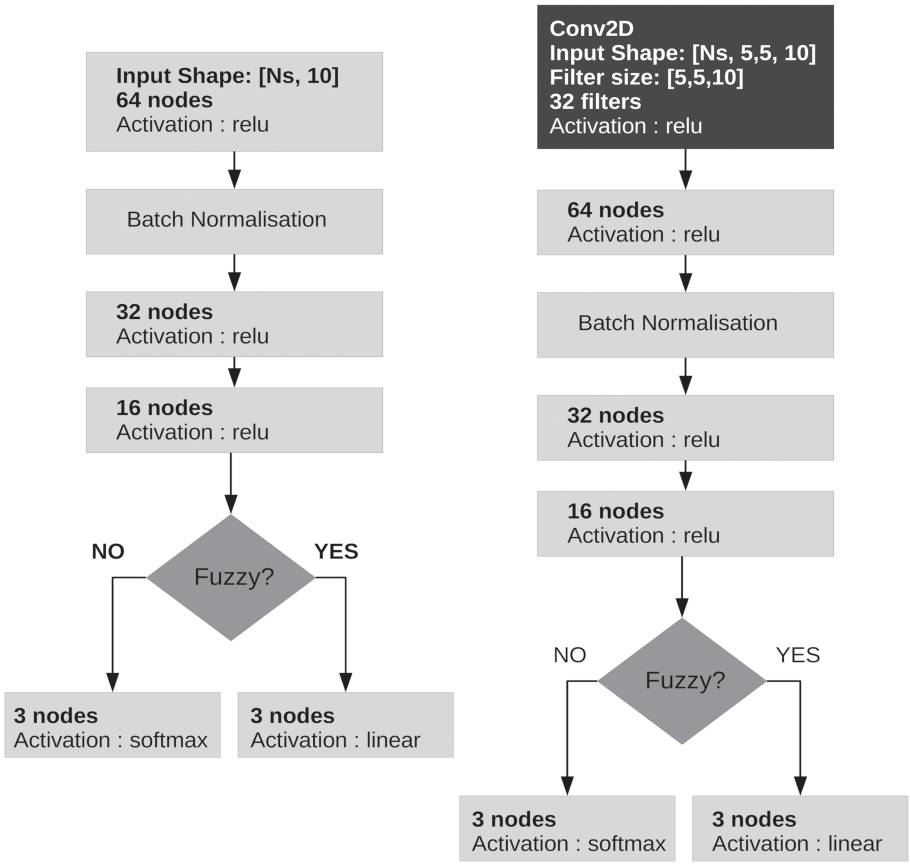

FIGURE 5. Neural network architectures. (Left) Dense neural network (DNN) architecture with three main node layers of 64, 32 and 16 nodes with a batch normalization layer. This network has 3362 trainable parameters. (Right) Optimal compact convolutional neural network (cCNN) chosen during the sensitivity analysis show in the Supporting Information. This cCNN uses a single convolution layer and has 12931 trainable parameters. Both networks can be made to deliver fuzzy or crisp classification with a change of the activation function in the three node output layer.

epochs of training in four minutes for the cCNN. GPU support is therefore desirable but not essential for this work.

A crucial aspect of our model selection procedure was an examination of the cCNN filter patterns. A CNN filter is a convolution filter, sometimes called a template or a kernel. During convolution, the kernel scans the image raster and for each kernel position within the image, a dot-product is performed and the result is mapped to a new raster. In cases where the convolution kernel has the same size as the image, the convolution output is a single scalar value. Readers familiar with image processing may recall that by organizing the values in a convolution kernel, we can engineer filter operations that detect lines along preferential orientations or basic shapes. Deep learning with CNN obviates this engineering step by allowing the values in the convolution to be learned during network training. It therefore becomes possible to examine the resulting filter values and assess the CNN performance. Readers should note from Figure 5 and from the Supporting Information document that our optimal model uses a filter size which is the same as the size of our image tiles. Therefore, since we use 32 filters, each filter will produce a single scalar value passed to the subsequent densely connected layers. In effect, the CNN operation will produce one new predictor per filter. In our case, the local image area of the small 5×5 tile (through 10 bands) will be transformed into 32 new predictors via the convolution operation. We can examine the convolution filter values in order to get a better understanding on exactly how the CNN is converting the 5×5 spatial neighbourhood into new predictors. Figure 6 summarize Figure S4 from the Supporting Information document. Here we see the two dominant patterns resulting from the training of our final model. On the left, we see a pattern where the central pixel has the highest value and will therefore control the output of the convolution dot-product. With this pattern, this CNN filter is using the brightness of the central pixel to contribute to the final prediction. Readers are reminded that during tensor construction, the label information for each tensor is in fact the label data for the central pixel in the tensor. This use of the central pixel therefore amounts to a pixel-based classification. Our cCNN has learned to behave like a DNN. Additionally, Figure 6 (right) shows an opposite pattern where the central pixel weight is minimized and where the outer pixels of the filter make the largest contribution. The cCNN is therefore also using the outer pixels in the template to contribute to the final prediction. This is clear evidence of exactly how our cCNN can use both the central pixel and an image tile neighbourhood to contribute to predictions.

Experimental structure

We structure this work around three experiments. These use our available data in different partitions of training and validation samples that will allow us to test and demonstrate key aspects of model performance. The main condition in all these experiments is that data used to train models will never be used to validate these models. There will always be a clear separation of time and/or space between training and validation datasets. Readers are also reminded that our initial 52543 samples have undergone a data augmentation procedure and that we now have a total of 210172 samples at our disposal for training UAV-based models and 79650 samples to train desk-based models. However, for both UAV- and desk-based models in validation, we remove the augmented samples and only use raw data. Our first experiment aims to test temporal resilience

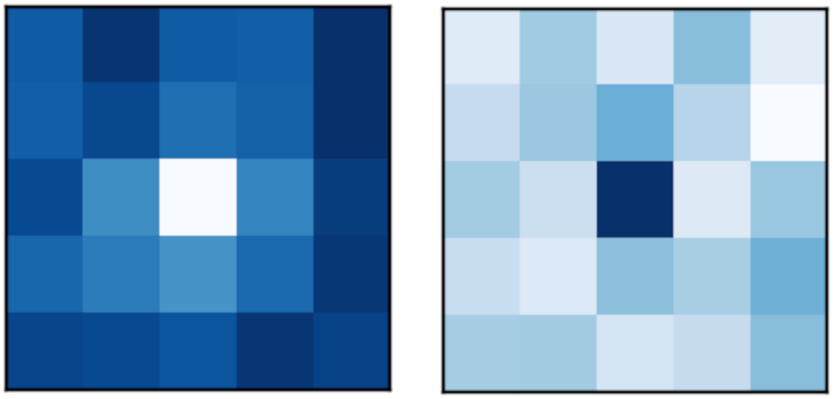

FIGURE 6. Example of CNN filter patterns showing two key responses. Patches in light blue/white have the strongest response and those in dark blue have the weakest response. (Left) Central pixel dominates the response, this filter uses the central pixel to contribute to the fuzzy membership prediction. (Right) Outer pixels dominate the response. This filter uses the spatial neighbours to contribute to the prediction of fuzzy membership. This is a key function that a DNN cannot perform. [Colour figure can be viewed at wileyonlinelibrary.com]

of trained models. We use label data acquired in 2018 to train models that will be validated against data acquired in 2017. Our second experiment aims to test the multi-river spatial transferability of models. We use label data acquired for the Rivers Sesia (2018) and Bonamico (2017 and 2018) to train models that will be validated with data from the Rivers Po (2017 and 2018) and Paglia (2017 and 2018) and imaged on separate Sentinel-2 tiles. The third experiment tests local spatial transferability of models. We wish to assess if local label data acquired from one or two sites could be used to classify whole, but single, rivers. This experiment is therefore separated in three parts (Rivers Sesia, Paglia and Bonamico) and we use label data from one or two sites to train models validated against an additional separate site. Table 2 summarizes the three experiments and gives the number of available samples for each.

Benchmark methods

We compare our CNN model to other more established methods. This aspect of our work has two linked objectives: we wish to assess if the field-effort and cost required to collect the UAV imagery is justified and; if our novel CNN actually delivers better performance than established methods. In the case of crisp classification performance, we will use the DNN as presented in Figure 5 (left) as a comparator method. Training and validation site selection follows Experiments 1–3 (Table 2).

But here we use different combinations of desk-based and UAV-derived labels as follows: (1) we use the desk-based labels to train a model validated with desk-based labels (from other sites as per Table 2); (2) we use desk-based labels to train a model validated with UAV-derived labels; (3) we use UAV-derived labels to train models validated against UAV-derived labels. For the fuzzy classification tests, we use two comparator methods: linear unmixing and the DNN from Figure 6. We implement linear unmixing with the Orfeo Toolbox used as a plugin for QGIS 3.4 (Grizonnet *et al.*, 2017). This workflow starts by determining endmembers with the automated method of Nascimento and Dias (2005). Here we input three classes. Then, the abundance fraction of each component (the fuzzy classification) is estimated with an unmixing procedure using the Minimum Dispersion Constrained Nonnegative Matrix Factorization algorithm (Huck *et al.*, 2010). This is a fully unsupervised procedure and it runs on individual images. It was therefore executed on each Sentinel-2 cropped image of our original dataset and we produced a total of 16 outputs, one output for each line of Table 1. As this is a fully unsupervised process, the three classes are not necessarily in the same order as per our classification scheme. We therefore used visual interpretation of each output to map the linear unmixing class scheme to our own scheme. Furthermore, because this output is unsupervised, training data is not relevant. When used in the experiments described earlier, we will collate the results of linear unmixing results for the validation sites only, the

Table 2. Summary of the three main experiments. Sample numbers include data augmentation which transforms the initial 52543 samples to an augmented possible maximum of 210172 samples. Each experiment uses of a portion of the data to train models and reserves the rest for validation. In the case of validation samples, we do not use augmentation and the data corresponds to the actual Sentinel-2 pixel samples

		UAV-based		Desk-based	
	Description	Training [#]	Validation [#]	Training [#]	Validation [#]
Experiment 1	All data from 2018 in training, all data from 2017 in validation.	150996	14794	53630	6505
Experiment 2	Rivers Sesia and Bonamico in training, Rivers Po and Paglia in validation.	112676	24375	38813	10231
Experiment 3a	Bonamico downstream in training, Bonamico upstream in validation.	47271	3380	11932	1985
Experiment 3b	Paglia Acquapendente + Orvieto in training, Paglia Allerona in validation.	25069	4472	8835	1624
Experiment 3c	Sesia Arborio in training, Sesia Caresana in validation.	35808	4021	8427	2633

training sites are not relevant to this output. The DNN and CNN are trained with two different types of data and we once again use our desk-based labels to assess the value of UAV fieldwork. We will use the desk-based labels to train fuzzy DNN and CNN models. The desk-based labels do not have any fuzzy values (all are assumed pure classes) but we can convert this data to 100% membership values (i.e. fraction of 1.0) and let the models infer fuzzy composition from the end-member values. Also, we will use the UAV-derived label data to train and validate models according to the experiments in Table 2. In summary, for each experiment described earlier, we have three sets of results. First, we present crisp classification performances using a mix of desk-based and UAV-based data. Second, we have a set of fuzzy classification results from linear unmixing, DNN and CNN models that have not used any field data as training inputs but which will still be validated against the UAV-based data. These are uniformly described as 'desk-based' results/approaches. Finally, we have a set of fuzzy classifications outputs from the DNN and CNN that were both trained and validated with UAV-based data. These will be described as UAV-based results/approaches.

Error metrics

We quantify the earlier results with the following error metrics. For crisp classification we use a simple accuracy metric defined as the percentage of pixels having the correctly predicted class. In reporting fuzzy classification errors, other works such as Foody *et al*. (1997) use an overall root-mean-square (RMS) value as a main reported error and then subdivide this into class errors. We argue that this is overly optimistic, the overall RMS value includes many small error predictions from the minority class that will draw the RMS error down and therefore cannot be used as a single metric to characterize fuzzy prediction errors across a whole image. We therefore propose two new error categories: dominant and sub-dominant class errors. We define the dominant class as the class that has the highest membership fraction for each pixel and the sub-dominant class as the class which has the second highest membership fraction. The dominant class error is calculated by finding the dominant class of a pixel in the ground-truth data (UAV-derived labels) and differencing it from the predicted membership fraction in the same class. A dominant class error of −1 means that the predicted membership fraction is 0 while the actual observed membership fraction is 1. A similar calculation can be made for the sub-dominant class errors. In terms of actual error metrics and quantification, we will report error distributions, the mean absolute error, the median error (useful for statistical tests) and the error variance. We find that this new dominant/sub-dominant conceptualization of errors has several advantages. It is a uniform error metric that can be applied to all the pixels in an image irrespective of land-cover type. In the Supporting Information, we compared the values for dominant/sub-dominant errors to error values explicitly calculated for each class (see Table S5 and Figure S6). We find similar but statistically different error distribution. Crucially, the dominant error is higher than each individual class error and therefore provides a more conservative error estimate. This can be understood since the dominant class selection process concentrates the largest classification errors into a single category. Unlike the overall RMS values reported by works such as Foody *et al*. (1997), our dominant class median and mean absolute errors can be applied to any pixel in a fuzzy prediction raster and we have confidence that our dominant class error does not under-estimate the actual error. This of course assumes that over-estimating errors is preferred to under-estimating errors. We also find that the dominant class error is more suited to change detection research by allowing us to directly establish thresholds for meaningful detections of a sub-dominant class overtaking a dominant class such as cases where vegetation colonizes a stable bar of river sediment.

Results

Table 3 presents the results for crisp classification performance. We note some key overall patterns. When we compare the accuracies of desk-based models validated with desk-based data against the accuracies of desk-based models validated with UAV data, we see that desk-based validation always over-estimates the quality of a classification, sometimes by as much as 16%. The use of a fuzzy classifier to predict pure crisp classes gives the best performance. Our cCNN outperforms the DNN with accuracies ranging from 95.5% to 99.9% and performs above 90% in each experiment. In terms of performance across the different experiments, Experiment 2 has slightly lower results. This shows that the task of classifying new rivers

Table 3. Crisp classification results. The term 'desk-based' refers to the use of on-screen image interpretation and digitization of regions-of-interests with specified pure class areas defined by a human user. These regions of interest can be used as training data or as validation data. The terms 'UAV training' or 'UAV valid.' refer to the use of our UAV-derived labels in either training or validation. Final column presents results of the UAV-based method. Fuzzy training means that we train the classifier with fuzzy membership and then crisp the results to obtain classes that are predicted to be 95% pure

		Desk-based training: Desk-based valid.	Desk-based training: UAV valid.	UAV training: UAV valid.	Fuzzy training: UAV valid.
DNN	Experiment 1	85.6	82.4	93.2	99.2
	Experiment 2	87.5	82.9	91.5	97.8
	Experiment 3a	90.4	82.3	86.4	89.1
	Experiment 3b	97.6	83.6	96.5	98.1
	Experiment 3c	80.6	76.8	89.5	97.7
CNN	Experiment 1	83.4	82.5	94.0	99.3
	Experiment 2	79.9	72.1	88.9	95.5
	Experiment 3a	86.0	81.0	89.4	98.8
	Experiment 3b	97.0	81.0	94.7	99.9
	Experiment 3c	82.6	68.7	90.7	96.9

Table 4. Validation results for desk-based approaches. Left: results for the Linear Unmixing method (LuM) which is purely unsupervised and has no training data. The classes of the outputs were interpreted and correctly associated to our class labelling for water, vegetation and sediment. Middle: results of a fuzzy DNN trained with the desk-based image interpretation. Right: results of a fuzzy CNN trained with desk-based polygons. When training a fuzzy classifier from desk-based data, we treat the digitized classes as pure class and attribute them a 100% membership for their given class observation. The network is left to infer partial memberships from these endmembers. We give the mean absolute error (MAE), the median error (MDE) and the error variance (EVAR) expressed in percentages

	LuM			DNN			CNN		
	MAE	MDE	EVAR	MAE	MDE	EVAR	MAE	MDE	EVAR
Dominant class errors									
Experiment 1	51.6	−52.9	4.2	20.9	−0.1	10.5	24.4	−2.8	13.1
Experiment 2	71.2	−67.6	4.8	26.2	−5.5	11.3	25.5	−5.0	12.4
Experiment 3a	60.1	−64.4	5.1	24.2	−3.9	10.5	57.1	−56.9	3.0
Experiment 3b	75.0	−81.3	4.5	22.8	−2.1	10.1	57.0	−50.3	2.7
Experiment 3c	65.1	−68.2	2.9	52.8	−56.9	16.2	71.6	−73.9	3.3
Sub-dominant class errors									
Experiment 1	24.2	−22.3	5.4	11.1	0.4	5.4	11.2	0.6	5.6
Experiment 2	13.7	−20.2	4.4	10.6	0.0	4.7	10.5	0.0	5.6
Experiment 3a	20.2	−18.5	3.1	16.7	0.4	7.1	10.7	2.3	2.4
Experiment 3b	17.6	−8.4	3.2	10.2	0.5	9.5	9.0	2.2	1.9
Experiment 3c	12.3	15.7	2.5	13.0	0.1	4.0	8.3	1.2	1.9

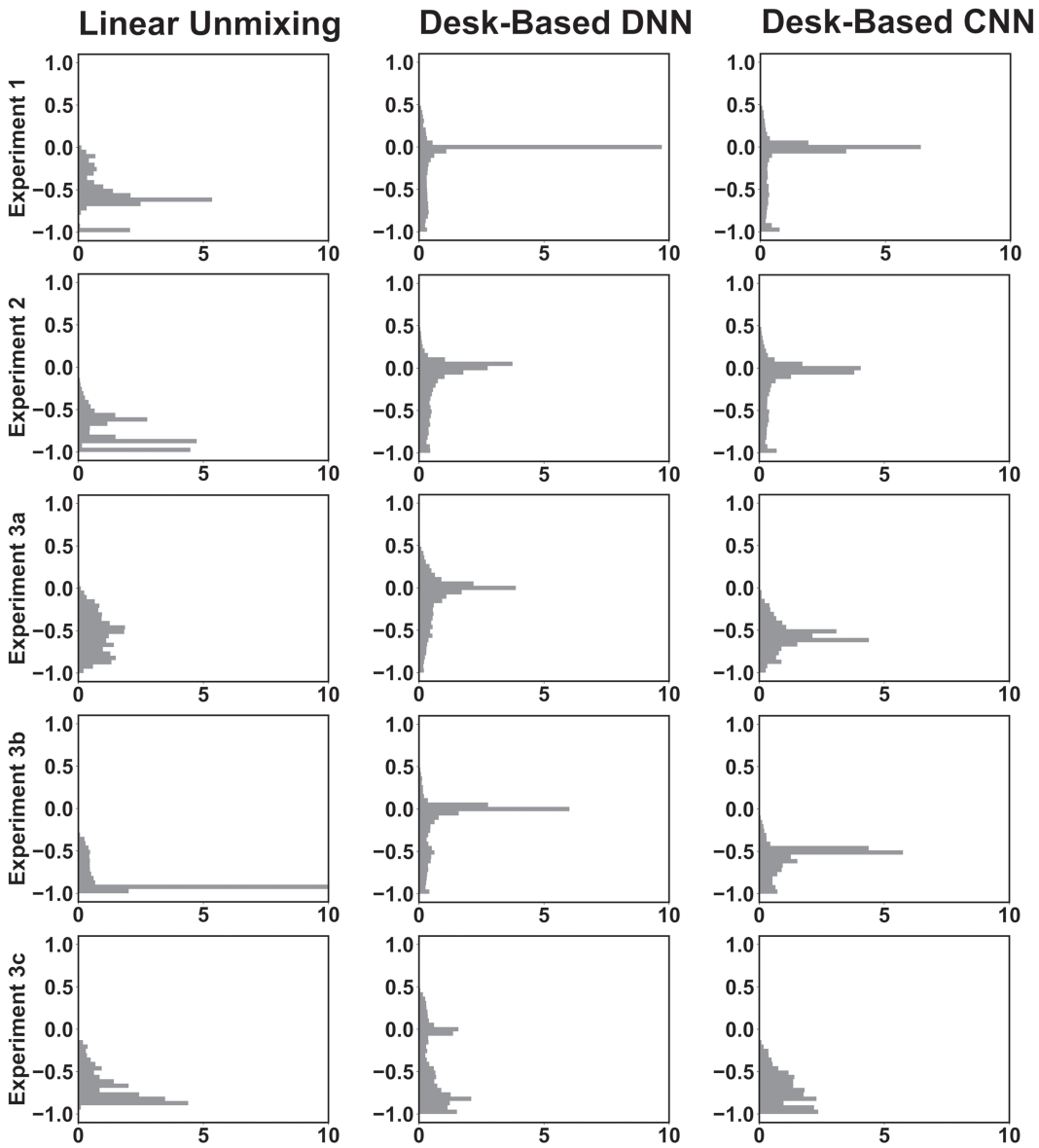

FIGURE 7. Dominant class error histograms from desk-based methods. Horizontal axes give probability density and the vertical axes give the error expressed as a fraction from −1 to 1. An error of −1 signifies that the predicted dominant class membership value for a pixel is 0 with associated ground-truth observation of 1. Desk-based models are fuzzy models trained with the manually digitized polygons.

not seen in the training data is the most difficult. However, our reported accuracy of 95.5% can still be considered as a state-of-the-art performance.

We now consider fuzzy classification results. Table 4 presents the error statistics for the desk-based approaches where UAV-derived data are not used to train models. Figure 7 presents the error distributions for the associated dominant class errors. Figure 8 presents the error distributions for the associated sub-dominant class errors. Table 4 shows that the linear unmixing approaches performed very poorly with dominant class prediction errors as high as 75%. Visual checks of the data confirm that whilst the overall pattern seems correct, actual membership predictions are small and clearly underestimated. This is confirmed by the fact that the median errors for linear unmixing are strongly negative. If we look at Figure 7 we see that the dominant class prediction error distributions have modal peaks well below 0. In the case of sub-dominant errors for linear unmixing, the median errors remain large. Dominant class errors for the desk-based fuzzy DNN are somewhat better with the exception of Experiment 3b which again has a very high median error. However if we look at the error distributions for the desk-based DNN in the middle column of Figure 7, we see that they are often characterized by a strong modal peak for errors of 0 and then a nearly uniform distribution of errors between −1 and 0. The peak at errors of 0 is encouraging, and strongly impacts the statistics in Table 4. However, the quasi-uniform portion of this distribution is problematic. It means that where the error is not 0, the actual magnitude of the error has a nearly equal probability of having any value from 1 to 0. Ideally, we would prefer to have error distributions closer to a normal curve where low magnitude errors are more probable than high magnitude errors. The subdominant class errors reported in Figure 8 show similar behaviour. The performance of the desk-based CNN is in fact worse than that of the desk-based DNN. Dominant and sub-dominant class errors for Experiment 3 are all very high. Overall, we again find that Experiment 2 seems to be the hardest case.

We now consider the performance of fuzzy models trained with the benefit of UAV-derived field labels. Table 5 gives error statistics and Figures 9 and 10 give error distributions for dominant and sub-dominant classes, respectively. In Table 5, we

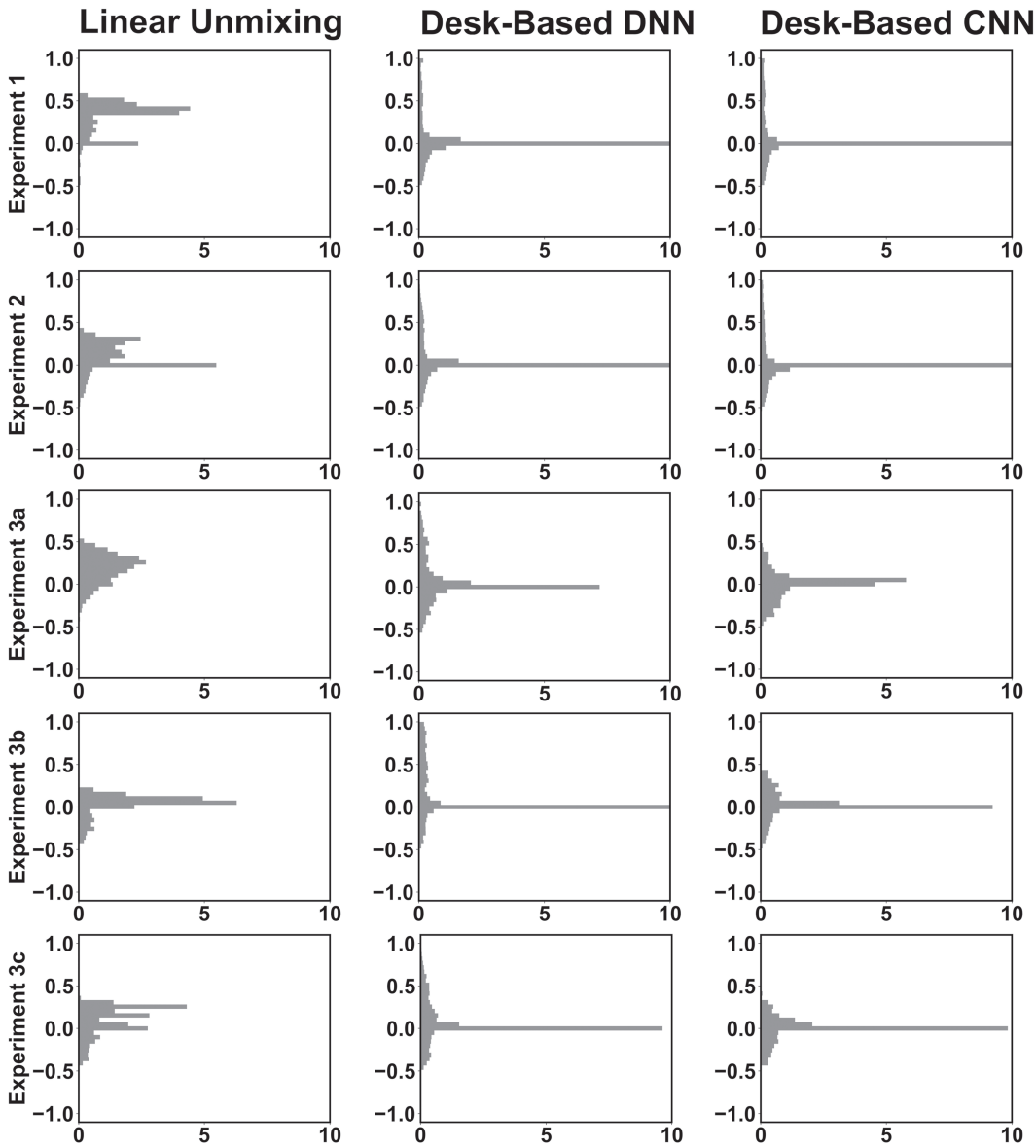

FIGURE 8. Sub-dominant class error histograms from desk-based methods. Horizontal axes give probability density and the vertical axes give the error expressed as a fraction from −1 to 1. An error of −1 signifies that the predicted dominant class membership value for a pixel is 0 with associated ground-truth observation of 1. Desk-based models are fuzzy models trained with the manually digitized polygons.

Table 5. Validation results for UAV-based approaches. Results for fuzzy DNN and CNN models trained with the UAV-derived fuzzy labels. Mean absolute error (MAE), the median error (MDE) and the error variance (EVAR) are expressed in percentage

	DNN			CNN		
	MAE	MDE	EVAR	MAE	MDE	EVAR
Dominant class errors						
Experiment 1	18.5	−11.3	4.0	14.2	−4.2	4.6
Experiment 2	17.2	−3.8	6.1	18.0	−5.5	5.9
Experiment 3a	23.0	−10.9	8.1	20.7	−4.6	7.6
Experiment 3b	21.0	−10.4	5.0	14.8	−2.4	4.1
Experiment 3c	19.0	−10.0	4.8	17.0	−5.0	6.2
Sub-dominant class errors						
Experiment 1	13.8	4.7	3.5	11.4	2.2	3.6
Experiment 2	13.3	1.8	4.3	13.8	2.5	3.1
Experiment 3a	20.0	6.8	7.3	17.9	1.8	6.0
Experiment 3b	11.7	4.2	2.1	10.9	0.7	3.1
Experiment 3c	14.4	10.2	2.6	11.2	1.6	3.1

note the absence of catastrophic errors. The highest mean absolute error in Table 5 is 23% for the DNN model. The highest magnitude median error is 11.3% again for the DNN model. The error distributions of Figures 9 and 10 show a much more desirable pattern when compared to the desk-based DNN outputs (reproduced in Figures 9 and 10 from Figures 7 and 8, respectively) with low magnitude errors being more probable than higher magnitude errors. In Table 5, the CNN generally performs better with the exception of Experiment 2 where the DNN has slightly out-performed the CNN.

In order to add context to the statistical results presented earlier, we will now examine a sample of fuzzy class rasters in map format (Figures 11-13). Figure 11 shows a sample from Experiment 1. We show the original cropped Sentinel-2 image, the UAV-derived labels, the fuzzy class obtained by the desk-based DNN and the fuzzy class obtained by the UAV-based DNN. Here we clearly see that despite seemingly encouraging error statistics (see Table 4), the performance of the desk-based model is not acceptable in the case of this single image (one of six validation sites for Experiment 1). In terms of areas, the desk-based model has successfully predicted the large area of sediment and some of the vegetated area. But the prediction has a high percentage of vegetation in the wetted perimeter which is clearly wrong. The error statistics are dominated by the large area of correctly predicted pixels, but the outcome of the desk-based model cannot be reliably used in any subsequent analyses. Figure 12 shows an example from

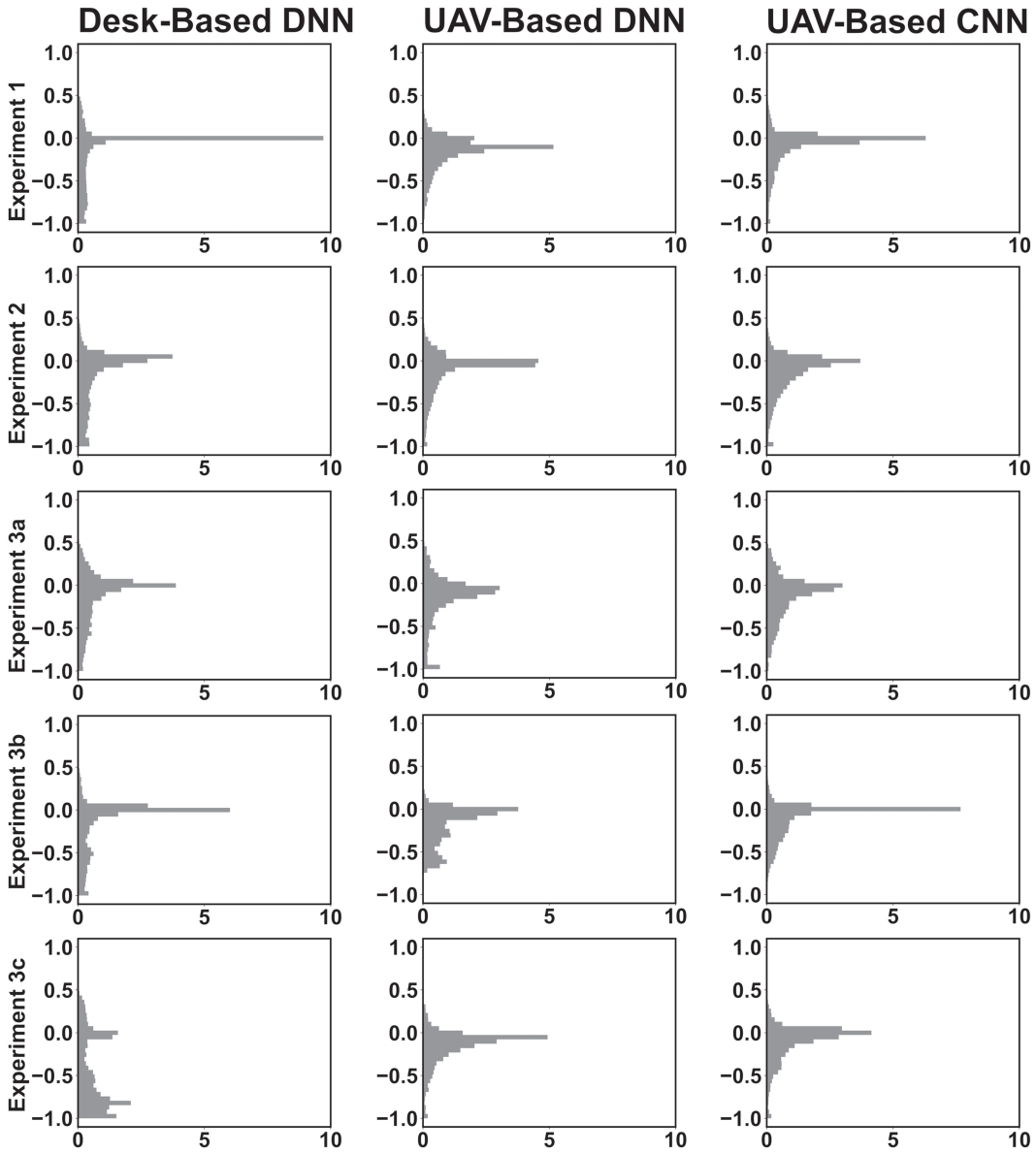

FIGURE 9. Dominant class error histograms from UAV-based methods. Desk-based DNN method results from Figure 7 are reproduced for comparison. Horizontal axes give probability density and the vertical axes give the error expressed as a fraction from −1 to 1.

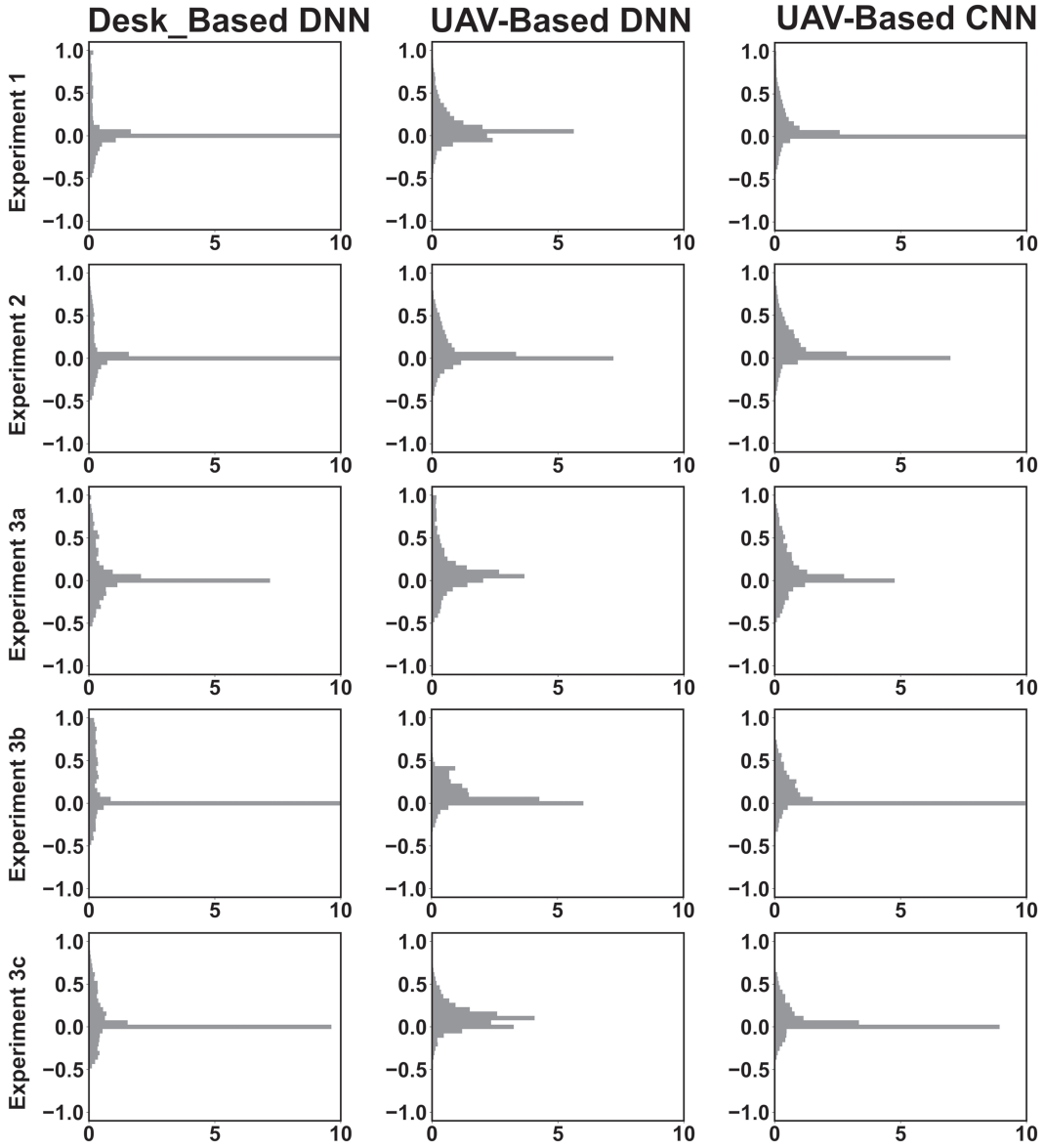

FIGURE 10. Sub-dominant class error histograms from UAV-based methods. Desk-based DNN method results from Figure 8 are reproduced for comparison. Horizontal axes give probability density and the vertical axes give the error expressed as a fraction from −1 to 1.

Experiment 2. In this case, the error statistics shown in Table 5 show that the DNN model had a slightly better performance. However, Figure 12 shows both the UAV-based DNN and CNN models. Close examination will show a systematic narrowing of the sediment banks and an under-estimation of the wetted perimeter. In Figure 12, the CNN model predictions are much closer to the UAV-derived ground-truth labels. Similarly, Figure 13 shows an example from Experiment 3. Here we can see that the CNN fuzzy predictor has produced a closer semblance of the actual pattern of channels. Both Figures 12 and 13 show examples of small errors that do not have a large weight in the calculation of statistics but which are disproportionately important for fluvial geomorphology studies.

Finally, we show how fuzzy class rasters produced with our cCNN model can be deployed at larger scales. We trained a master model using all 210172 training samples. We then applied it to produce a total of 294 linear kilometres of river corridor classifications for the Po, Sesia, Paglia and Bonamico Rivers as imaged in 2018 by Sentinel-2. These cannot be clearly displayed in static format, but they have been made available to the reader in the online data and can be opened by any GIS package. In order to demonstrate a large-scale application, we use a 10km stretch of the River Paglia and show how our approach can detect meaningful net change over a oneyear period. Figure 14 (along the diagonal) shows this 10km reach of the Paglia for years 2017 and 2018. Figure 14 also shows insets for three sediment bars, A, B and C, where vegetation growth can be seen. Figure 15 shows membership distributions for the dominant and sub-dominant classes of vegetation and sediment for each bar and for 2017 and 2018. Bar A is included in the Paglia Allerona UAV acquisition site. Bars B and C were not included in any UAV survey. From the UAV data for bar A, we count the number of pixels (@ 10cm) as a percentage of the total for each class. Bar A had 36% sediment pixels in 2017 and 21% in 2018. Vegetation pixels occupy 57% in 2017 and 73% in 2018. For the fuzzy model predictions for bar A, the median sediment membership for 2017 is 34% (−2% error) and 16% (−5% error) for 2018. Vegetation membership was 62% (+5% error) in 2017 and 79% (+6% error) in 2018. In the case of the fuzzy predictions for bar A, a Mann–Whitney U-test confirms that the medians of these distributions are significantly different with a p-value of 5E-28 for the 2017/2018 sediment membership pairing and 1E-18 for the 2017/2018 vegetation membership pairing. We now consider fuzzy predictions for bars B and C and use U-tests to establish significance of the observed changes. For

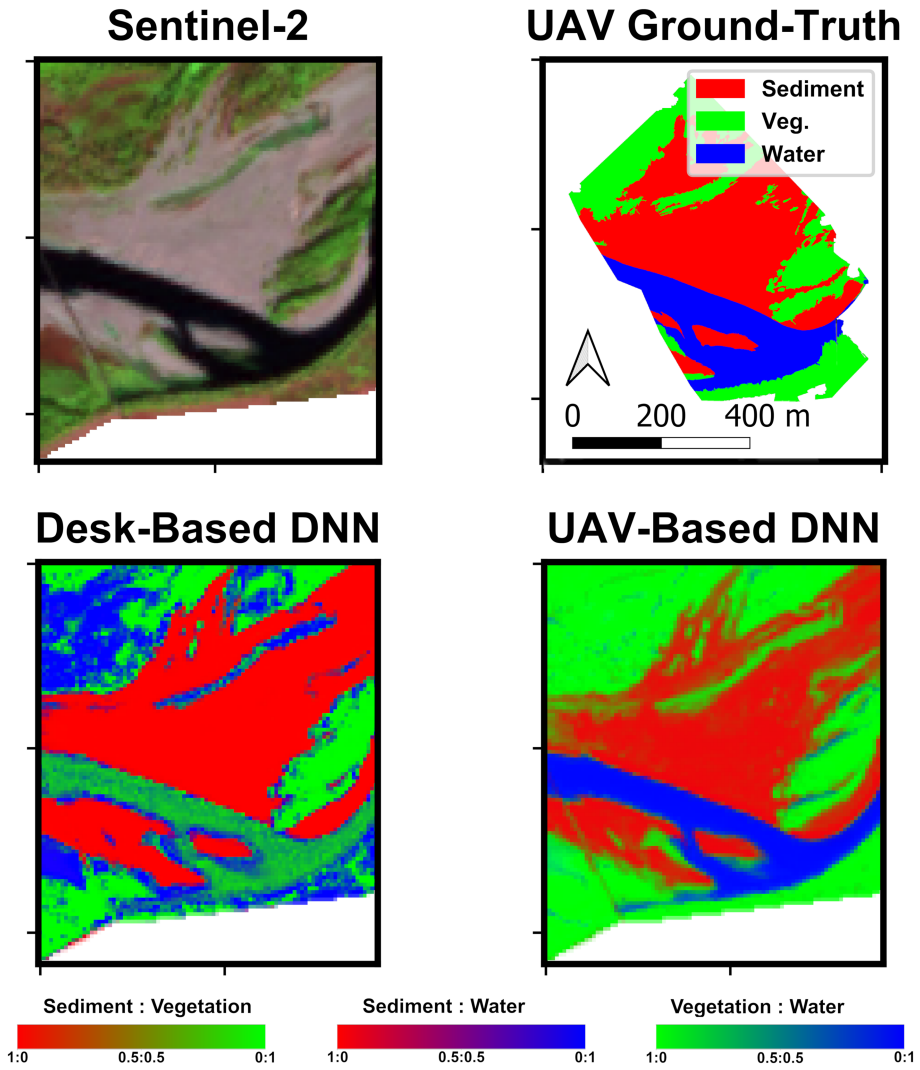

FIGURE 11. Example of results from Experiment 1. The four sub-plots have the same scale, orientation and extent. Colour bars give the legend for two-class mixture ratios with fuzzy memberships expressed as 0–1 fractions. Despite seemingly encouraging statistics, the desk-based DNN approach has performed poorly when predicting areas dominated by water and vegetation. [Colour figure can be viewed at wileyonlinelibrary.com]

bar B, median sediment membership in 2017 was 46% and 27% in 2018 (p-value = 6E-40). Vegetation median membership was 50% in 2017 and 66% in 2018 (p-value = 1E-11). For bar C, median sediment membership was 75% in 2017 and 59% in 2017 (p-value = 2E-03). Vegetation median membership was 24% in 2017 and 40% in 2018 (p-value = 1E-03).

Discussion

We have demonstrated that classification models, both crisp and fuzzy, trained with the benefit of ground-truth data derived from low-altitude UAV flights systematically deliver better performance. Our cCNN, trained with UAV-derived label data, delivers optimal performance for both crisp and fuzzy classification of Sentinel-2 data and it can identify pure class pixels with an accuracy up to 99.9%. Similar models trained with desk-based data, not having the benefit of field observations, identify pure class pixels with a reduced accuracy of 81%. In terms of fuzzy classifications, we find that our cCNN delivers the optimal error statistics (Table 4) confirmed by well structured error distributions (Figures 9 and 10) and visually accurate fuzzy classification outputs (Figures 11–14). Conversely, any modelling approach that did not benefit from the UAV-derived labels had markedly degraded performances (Table 3), poorly structured error distributions (Figures 7 and 8) and unsatisfactory fuzzy classification outputs (Figure 11). Similarly, Feng *et al.* (2018) note that linear mixing approaches do not always give consistent and easy to interpret results when predicting class memberships.

Our findings therefore support the continued importance of fieldwork as a primary data source (Lane, 2020). In this first published example of UAV imagery applied to the problem of satellite image classification, we show that drone surveys can be used as a cost-effective tool to extend the value of local fieldwork to regional scales. We have evidenced three possible application scenarios where we attempt to predict fuzzy membership for the main land-cover elements of the river corridor: water, vegetation and sediment. In our first experiment we show that UAV surveys acquired at multiple sites and for multiple rivers can satisfactorily predict fuzzy memberships for the same sites and rivers, but acquired in the previous year (median error of −4.2% and mean absolute error of 14.2%). This demonstrates that fuzzy models can transfer to different Sentinel-2 tiles of the same location, but for different times. This is an interesting finding for regional studies aiming to monitor/characterize several rivers over a period of several years. As a consequence, it is not necessary to fund annual field survey, but it is sufficient to guarantee a biennial UAV acquisition to achieve a good model performance.

In our second experiment we show that UAV surveys acquired at for two rivers can produce models that can satisfactorily predict fuzzy memberships for two new rivers as imaged

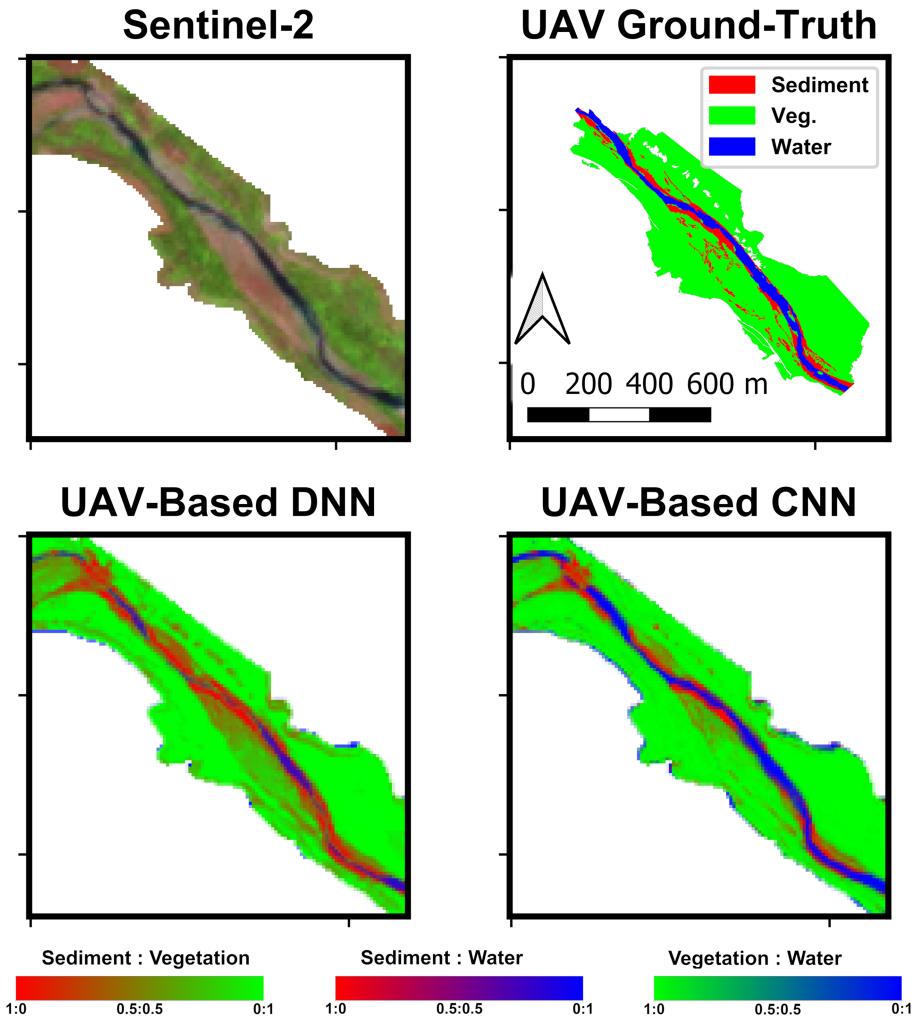

FIGURE 12. Sample results from Experiment 2. The four sub-plots have the same scale, orientation and extent. Despite the DNN have slightly better outcome statistics, we see that in this case, the wetted area of the channel seems under-estimated. This is a small area which mitigates the impact of the errors on the statistics, but for fluvial studies, this is an important error. [Colour figure can be viewed at wileyonlinelibrary.com]

in two separate Sentinel-2 tiles (median error of 5.5% and mean absolute error of 18.0%). The obvious caveat here is that all our field sites are in Italy. Whilst Italy does have a very wide range of river types with strong north–south gradients of temperature, hydrology and geology, our data obviously does not include tropical rivers, arctic rivers, bedrock rivers, etc. Therefore, from a global perspective, our data spans a relatively similar range of conditions. However, whilst the Rivers Po and Sesia are in the same catchment and are arguably similar, the Rivers Paglia and Bonamico are markedly different with the Paglia being a small single thread channel with exposed paleo-marine clays in the central Appenines and the Bonamico being an actively braiding channel with high sediment supply and transport rates in the Southern Appenines but also with a very episodic hydrology characterized by periods of flash floods contrasted with extreme, almost ephemeral, low flows (Figure 1). Nevertheless both the statistics and the visual appraisal of outcomes were satisfactory. This finding therefore shows that classification models trained from localized UAV survey data can be extended on a regional scale. Our first two experiments used in excess of 100k augmented samples in training supplied by 8 to 10 UAV surveys.

In our third experiment we explored the lower limit of training requirements by testing the classification of single rivers based on only one or two UAV acquisitions. The median errors range from −5% to 2.4% and mean absolute errors range from 14.8% to 20.7%. Interestingly, within Experiment 3, the ranking of errors follows the same order as that of the number of training samples. However this does not hold for the overall results. The errors for Experiment 3 are similar to those reported for Experiments 1 and 2 and we conclude that successful fuzzy classification models can be produced with a one or, preferably two, UAV surveys. In terms of surface area, Experiment 3 uses surveys covering from 0.6 km^2 to 1.1 km^2. We therefore propose a rule of thumb that for a given river at a given time, UAV surveys totalling ~1 km^2 offer a reliable training sample for classification models applicable to Sentinel-2 imagery. With the current generation of low-cost drones and associated flight planning software apps, this target is readily achievable in a single day. Given that Sentinel-2 tiles are 100 km wide, this local acquisition has the potential to train a classification model applicable to river corridor lengths in excess of 100 km. However, if the river corridor undergoes significant morphological transitions within the Sentinel-2 tile, we recommend caution and the deployment of additional drone surveys. If several square kilometres of UAV surveys are available, the methods presented here are capable of large-scale surveys and they can deliver nuanced change detection analysis. Even at smaller scales, we see repeated evidence that the fuzzy classification approach does partially mitigate for the resolution of the Sentinel-2 data with smaller, pixel-scale, features readily visible in fuzzy class rasters (Figures 12 and 13). Similarly, the results shown in Figure 15 could not be derived from a traditional crisp, semantic, classification workflow. Here we see the growth of vegetation and cases where vegetation overtakes sediment as the dominant class of the bar (bars A and B), or alternatively, where the

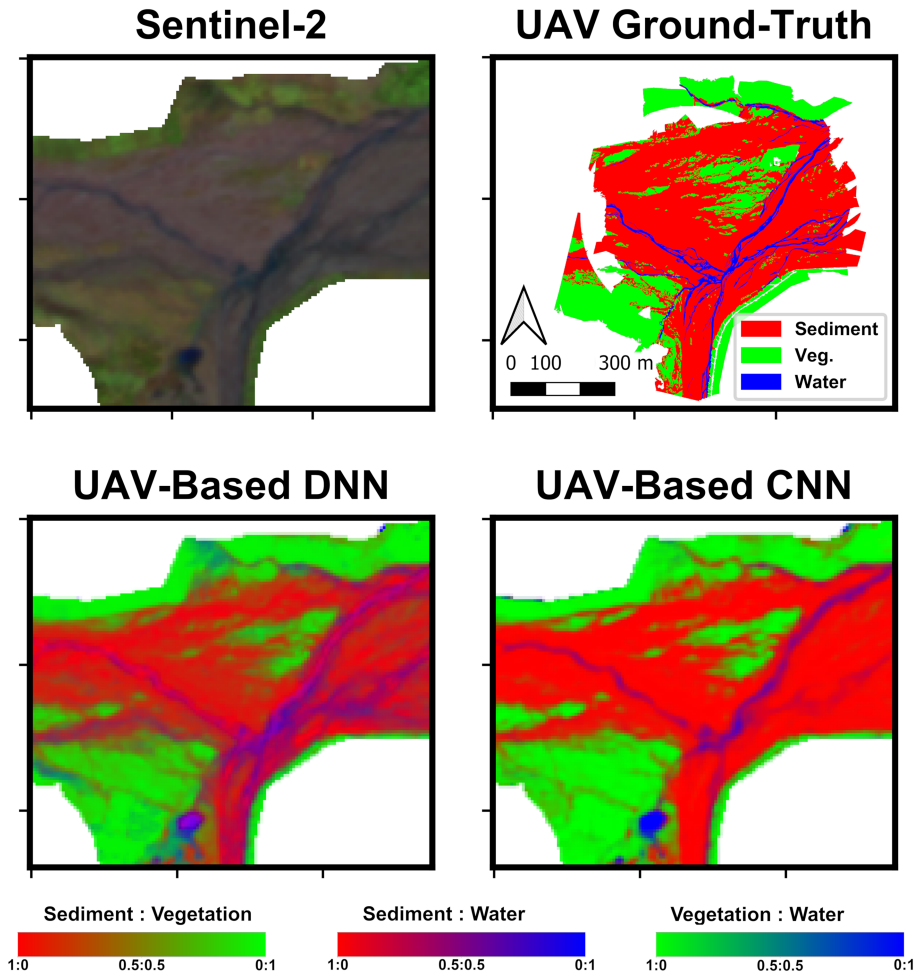

FIGURE 13. Sample result from Experiment 3a. The four sub-plots have the same scale, orientation and extent. In this case the DNN mean absolute error was 3% lower than for the CNN. Nevertheless, we see errors that are small in scale but potentially very important in the context of fluvial study. For example, the small water body on the bottom has been interpreted as an equal mixture of water and sediment. [Colour figure can be viewed at wileyonlinelibrary.com]

vegetation class increases in membership but marginally remains as the sub-dominant class of the bar (bar C).

A first limitation to note in this work is the lack of significant seasonal variability in our data. Surveys in central and northern Italy were carried out in a window from late spring to late summer and the climate in southern Italy displays less variation across the year. The models and results we present here must only be considered as valid during summer and thus useable for the monitoring of net annual change. In the case of applications where seasonal changes are required, especially if vegetation and/or snow classification is important, we recommend additional drone survey deployment in order to sample seasonal changes and the addition of extra classes in the models as appropriate. Another limitation to this work has been the requirement for computationally expensive super-resolved Sentinel-2 imagery. In the Supporting Information, we find that the use of 10 Sentinel-2 bands delivered the best performance. However, we did find that some models using only Sentinel-2 bands 2, 3, 4 and 8 (natively acquired at 10 m of spatial resolution) could deliver median errors below 10% and mean absolute errors below 20% (Table S2, Supporting Information). If the computational overhead of super-resolution is not an option, the methods presented here can still be used with the 2, 3, 4 and 8 band combination with an acceptable sacrifice of data quality. When evaluating the performance of this limited band-set, we recommend using multiple approaches. During this work it was observed that scalar error statistics are not entirely reliable if taken alone. In addition to the mean absolute error, median error and error variance we also experimented with the RMS error and the standard deviation of error. We considered these errors for each individual class and also for our newly developed dominant/sub-dominant class approach. All of these metrics have strong and weak points. Similarly to other authors, we found that the RMS was too heavily influenced by a small percentage of classification errors which lead to incorrect rankings of overall model performances (Willmott and Matsuura, 2005, 2006). But we found that using median error and error variance were effective because median values can readily be tested for significant change with a standard Mann–Whitney U-test and variance with a Brown–Forsyth test. We found that using class-specific error statistics made it difficult to rank overall model performance and were not practicable in the determination of the optimal CNN architecture and parametrization. In fact, we found that the production of a single error statistic based on a concatenation of all the class-based error predictions gave mean absolute and median errors that were artificially small and which under-estimate the error a user can expect from a fuzzy class prediction (Table S6 of Supporting Information). Therefore, in the end, we decided to use dominant and sub-dominant class median errors, mean absolute errors and error variances but to accompany the scalar statistics with a display of error distributions and an explicit examination of some actual fuzzy classification outputs. Overall, the errors reported here appear slightly higher than reported by Foody *et al.* (1997). These authors use Landsat 7 data at 20 m in a similar three-class problem in order to develop a fully fuzzy model for AVHRR data with a resolution of 1.1 km. We note that for the testing data, these authors find

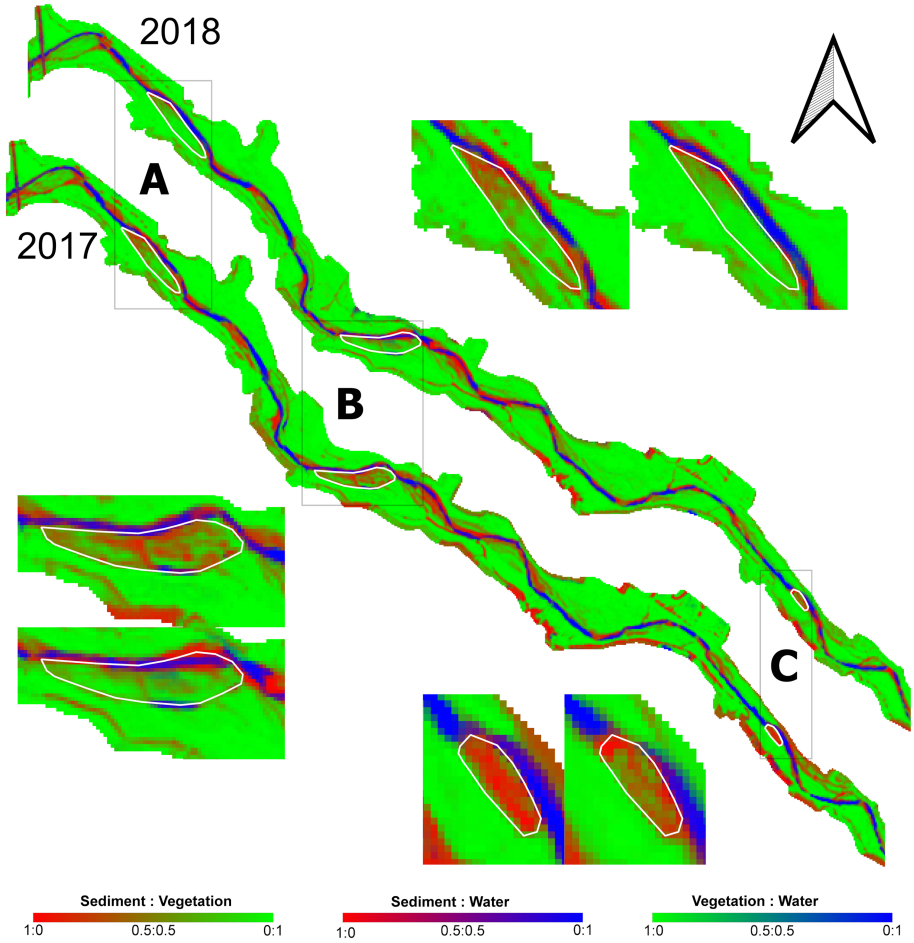

FIGURE 14. Fuzzy classifications for a 10km reach of the River Paglia for years 2017 and 2018. We show three sediment bars, A, B and C, where new vegetation growth is visible. Bar A is included in the Paglia Allerona UAV acquisition site. Bars B and C were not included in any UAV survey. [Colour figure can be viewed at wileyonlinelibrary.com]

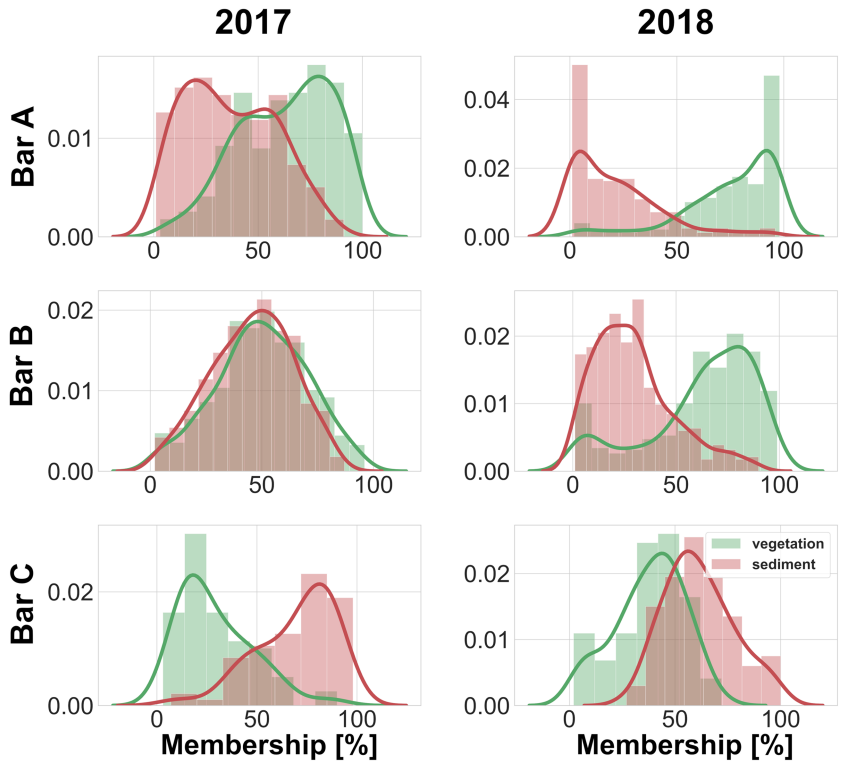

FIGURE 15. Membership histograms for sediment and vegetation classes of bars A, B and C from Figure 14. [Colour figure can be viewed at wileyonlinelibrary.com]

a best RMS error of roughly 8% to 9%. We argue that this difference in quality can be explained by the resolution of the label data. When using Landsat data at 30m spatial resolution, the scale of the Landsat pixel remains much larger than the small features such as individual trees and channels. The Landsat acquisition at 30m will have the usual smoothing effect on the edges of natural features smaller than 30m. In our case, the use of hyperspatial UAV-derived label data at 10cm implies that our training data approaches the full spatial variability found in our study landscape. This will increase the scatter in the fuzzy models because small isolated patches as small as 10cm ×10cm can make a contribution to the fuzzy membership calculation even if they do not impact the reflected intensity of radiation.

Finally, we note that alternate sources of training data may be acceptable. In the Supporting Information document, task 8, we present a brief analysis of the effect of resolution. Using a local median filter, we have downsampled our UAV-derived labels to spatial resolutions of 50cm, 1m and 3m. Whilst the effect of downsampling was to systematically degrade the quality of the fuzzy class rasters (see Table S7 and Figure S8 in Supporting Information), we find mean absolute errors in the area of ~20 to 25% which might be deemed acceptable for certain applications. This opens the way for other sources of training data such as lower resolution airborne surveys or inexpensive satellite data such as Ikonos or Planet Scope.

Conclusion

We have demonstrated a workflow where low-cost drone data is integrated to satellite imagery. By leveraging new super-resolution algorithms and the sub-pixel information made available through fully fuzzy classification models trained with UAV-derived label data, our method delivers continuous metric-scale information from freely available satellite imagery which is suited to fluvial geomorphology investigations. For example, a loss of pure-class sediment pixels accompanied with an increase of the vegetation class membership for a gravel bar indicates re-colonization of this bar by young plants and hence the creation of potential new habitats (Figures 14 and 15). Fuzzy classification allows us to detect growing vegetation well before it becomes dominant and shifts the crisp class of the pixel. This in turn has implications for the stability and age of the bar surface. In future applications, we expect the labour intensive manual OBIA element of the workflow to be replaced with emerging deep learning classifiers that are now achieving pixel-level classification performances as high as 99% (Buscombe and Ritchie, 2018; Carbonneau *et al.*, 2020). Furthermore, the cost of the method could be lowered by developing advanced co-registration methods which could correct for the expected metric-scale offsets that occur when using direct georeferencing drone surveys as presented in Carbonneau and Dietrich (2017). We also perceive a need for progress in the area of deep learning based super-resolution algorithms (e.g. Lanaras *et al.*, 2018), specifically trained to atmospherically correct Sentinel-2 data for fluvial corridors, this would facilitate the mass-production of Sentinel-2 data with 10m of spatial resolution across all bands. All these elements of progress are incremental and should not represent a major challenge. This work therefore opens a pathway to operationalized fluvial remote sensing suitable for both research and management applications at regional and national scales.

Acknowledgements—This work was performed in the framework of the 'IRIS – Italian Research and development Initiative for Spaceborne river monitoring' project and funded by the Italian National Institute for Environmental Protection and Research (ISPRA) in the form of a technical-scientific collaboration agreement between ISPRA and the Polytechnic University of Milan, Department of Electronics, Information and Bioengineering, within the context of the ASI-ISPRA initiative 'Habitat Mapping'. Impact activity on this project has been supported by Department of Geography Impact Fund, Durham University. The authors are grateful to the EU Copernicus Programme for freely providing Sentinel-2 data that was exploited in this work. Our thanks go also to the regional Institutions of Piedmont, Lombardy and Calabria Regions that made available discharge and precipitation records used to control hydraulic and hydrologic conditions of the monitored sites. We thank Dr Martina Bussettini (ISPRA, Italy) and Ms Francesca Piva (ISPRA, Italy) for their useful support during the UAV acquisitions. We also thank Prof. Andrea Castelletti for crucial support during the initiation of the project. Finally, we thank Dr Daniel Buscombe and another anonymous reviewer for helpful and constructive comments.

Data Availability Statement

Python code is available at: https://github.com/Pcdurham/UAV2Sen and can be cited as Carbonneau (2020). Data is available at: https://collections.durham.ac.uk/files/r1v692t6239 (Carbonneau *et al.*, 2020).

Conflicts of Interest

The authors have declared no conflicts of interest.

References

Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, Mané D, Monga R, Moore S, Murray D, Olah C, Schuster M, Shlens J, Steiner B, Sutskever I, Talwar K, Tucker P, Vanhoucke V, Vasudevan V, Viégas F, Vinyals O, Warden P, Wattenberg M, Wicke M, Yu Y, Zheng X. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems.

Belletti B, Rinaldi M, Bussettini M, Comiti F, Gurnell AM, Mao L, Nardi L, Vezza P. 2017. Characterising physical habitats and fluvial hydromorphology: A new system for the survey and classification of river geomorphic units. *Geomorphology* **283**: 143–157. https://doi.org/10.1016/j.geomorph.2017.01.032.

Bizzi S, Demarchi L, Grabowski RC, Weissteiner CJ, de Bund WV. 2016. The use of remote sensing to characterise hydromorphological properties of European rivers. *Aquatic Sciences* **78**: 57–70. https://doi.org/10.1007/s00027-015-0430-7.

Brodu N. 2017. Super-resolving multiresolution images with band-independent geometry of multispectral pixels. *IEEE Transactions on Geoscience and Remote Sensing* **55**: 4610–4617. https://doi.org/10.1109/TGRS.2017.2694881.

Burkov A. 2019. *The Hundred-Page Machine Learning Book by Andriy Burkov*. Self-Published.

Buscombe D, Ritchie AC. 2018. Landscape classification with deep neural networks. *Geosciences* **8**: 244 https://doi.org/10.3390/geosciences8070244.

Carbonneau PE. 2020. UAV2Sen: Fuzzy Classification of Sentinel 2 Imagery with UAV-based label data. Zenodo. https://doi.org/10.5281/zenodo.3911028

Carbonneau PE, Dietrich JT. 2017. Cost-effective non-metric photogrammetry from consumer-grade sUAS: implications for direct georeferencing of structure from motion photogrammetry. *Earth Surface Processes and Landforms* **42**: 473–486. https://doi.org/10.1002/esp.4012.

Carbonneau PE, Piégay H. 2012. Introduction: The growing use of imagery in fundamental and applied river sciences. In *Fluvial Remote Sensing for Science and Management*, Carbonneau PE, Piégay H (eds). John Wiley & Sons: Chichester; 1–18. https://doi.org/10.1002/9781119940791.ch1.

Carbonneau PE, Lane SN, Bergeron NE. 2004. Catchment-scale mapping of surface grain size in gravel bed rivers using airborne digital imagery. *Water Resources Research* **40**: W07202. https://doi.org/10.1029/2003WR002759.

Carbonneau P, Fonstad MA, Marcus WA, Dugdale SJ. 2012. Making riverscapes real. *Geomorphology* **137**: 74–86.

Carbonneau PE, Bizzi S, Marchetti G. 2018. Robotic photosieving from low-cost multirotor sUAS: A proof-of-concept. *Earth Surface Processes and Landforms* **43**: 1160–1166. https://doi.org/10.1002/esp.4298.

Carbonneau PE, Dugdale SJ, Breckon TP, Dietrich JD, Fonstad MA, Miyamoto H, Woodget AS. 2020.Adopting deep learning for RGB fluvial scene classification. EarthArXiv.https://osf.io/74kdg.

Chollet F. 2017. *Deep Learning with Python*. Shelter Island, New York: Manning, 1–384.

Curran PJ, Williamson HD. 1985. The accuracy of ground data used in remote-sensing investigations. *International Journal of Remote Sensing* **6**: 1637–1651. https://doi.org/10.1080/01431168508948311.

Downing JA, Cole JJ, Duarte CM, Middelburg JJ, Melack JM, Prairie YT, Kortelainen P, Striegl RG, McDowell WH, Tranvik LJ. 2012. Global abundance and size distribution of streams and rivers. *Inland Waters* **2**: 229–236. https://doi.org/10.5268/IW-2.4.502.

Dugdale SJ, Bergeron NE, St-Hilaire A. 2013. Temporal variability of thermal refuges and water temperature patterns in an Atlantic salmon river. *Remote Sensing of Environment* **136**: 358–373. https://doi.org/10.1016/j.rse.2013.05.018.

Dugdale SJ, Bergeron NE, St-Hilaire A. 2015. Spatial distribution of thermal refuges analysed in relation to riverscape hydromorphology using airborne thermal infrared imagery. *Remote Sensing of Environment* **160**: 43–55. https://doi.org/10.1016/j.rse.2014.12.021.

Dugdale SJ, Malcolm IA, Hannah DM. 2019. Drone-based Structure-from-Motion provides accurate forest canopy data to assess shading effects in river temperature models. *Science of the Total Environment* **678**: 326–340. https://doi.org/10.1016/j.scitotenv.2019.04.229.

Fausch KD, Torgersen CE, Baxter CV, Li HW. 2002. Landscapes to Riverscapes: Bridging the gap between research and conservation of stream fishes. *BioScience* **52**: 483–498. https://doi.org/10.1641/0006-3568(2002)052[0483:LTRBTG]2.0.CO;2.

Feng R, Wang L, Zhong Y. 2018. Least angle regression-based constrained sparse unmixing of hyperspectral remote sensing imagery. *Remote Sensing* **10**: 1546. https://doi.org/10.3390/rs10101546.

Foody GM. 1997. Fully fuzzy supervised classification of land cover from remotely sensed imagery with an artificial neural network. *Neural Computing & Applications* **5**: 238–247. https://doi.org/10.1007/BF01424229.

Foody GM, Cox DP. 1994. Sub-pixel land cover composition estimation using a linear mixture model and fuzzy membership functions. *International Journal of Remote Sensing* **15**: 619–631. https://doi.org/10.1080/01431169408954100.

Foody GM, Lucas RM, Curran PJ, Honzak M. 1997. Non-linear mixture modelling without end-members using an artificial neural network. *International Journal of Remote Sensing* **18**: 937–953. https://doi.org/10.1080/014311697218845.

Frechette DM, Dugdale SJ, Dodson JJ, Bergeron NE. 2018. Understanding summertime thermal refuge use by adult Atlantic salmon using remote sensing, river temperature monitoring, and acoustic telemetry. *Canadian Journal of Fisheries and Aquatic Sciences* **75**: 1999–2010. https://doi.org/10.1139/cjfas-2017-0422.

Gargiulo M, Mazza A, Gaetano R, Ruello G, Scarpa G. 2019. Fast super-resolution of 20m Sentinel-2 bands using convolutional neural networks. *Remote Sensing* **11**: 2635 https://doi.org/10.3390/rs11222635.

Goodfellow I, Bengio Y, Courville A. 2016. *Deep Learning*. MIT Press.

Grizonnet M, Michel J, Poughon V, Inglada J, Savinaud M, Cresson R. 2017. Orfeo ToolBox: Open source processing of remote sensing images. *Open Geospatial Data, Software and Standards* **2**: 15 https://doi.org/10.1186/s40965-017-0031-6.

de Haas T, Ventra D, Carbonneau PE, Kleinhans MG. 2014. Debris-flow dominance of alluvial fans masked by runoff reworking and weathering. *Geomorphology* **217**: 165–181. https://doi.org/10.1016/j.geomorph.2014.04.028.

Huck A, Guillaume M, Blanc-Talon J. 2010. Minimum dispersion constrained nonnegative matrix factorization to unmix hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing* **48**: 2590–2602. https://doi.org/10.1109/TGRS.2009.2038483.

Immerzeel WW, Kraaijenbrink PDA, Shea JM, Shrestha AB, Pellicciotti F, Bierkens MFP, de Jong SM. 2014. High-resolution monitoring of Himalayan glacier dynamics using unmanned aerial vehicles. *Remote Sensing of Environment* **150**: 93–103. https://doi.org/10.1016/j.rse.2014.04.025.

Lanaras C, Bioucas-Dias J, Galliani S, Baltsavias E, Schindler K. 2018. Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network. *ISPRS Journal of Photogrammetry and Remote Sensing* **146**: 305–319. https://doi.org/10.1016/j.isprsjprs.2018.09.018.

Lane SN. 2020. Editorial 2020 Part II: Data from nowhere? *Earth Surface Processes and Landforms* **45**: 5–10. https://doi.org/10.1002/esp.4775.

Lindner G, Schraml K, Mansberger R, Hübl J. 2016. UAV monitoring and documentation of a large landslide. *Applied Geomatics* **8**: 1–11. https://doi.org/10.1007/s12518-015-0165-0.

Ling F, Boyd D, Ge Y, Foody GM, Li X, Wang L, Zhang Y, Shi L, Shang C, Li X, Du Y. 2019. Measuring river wetted width from remotely sensed imagery at the sub-pixel scale with a deep convolutional neural network. *Water Resources Research* **55**: 5631–5649. https://doi.org/10.1029/2018WR024136.

Main-Knorn M, Pflug B, Louis J, Debaecker V, Müller-Wilm U, Gascon F. 2017. Sen2Cor for Sentinel-2, in: Image and Signal Processing for Remote Sensing XXIII. Presented at the Image and Signal Processing for Remote Sensing XXIII, International Society for Optics and Photonics, p. 1042704. https://doi.org/10.1117/12.2278218

Marcus WA, Fonstad MA. 2008. Optical remote mapping of rivers at sub-meter resolutions and watershed extents. *Earth Surface Processes and Landforms* **33**: 4–24. https://doi.org/10.1002/esp.1637.

Marcus WA, Fonstad MA. 2010. Remote sensing of rivers: the emergence of a subdiscipline in the river sciences. *Earth Surface Processes and Landforms* **35**: 1867–1872. https://doi.org/10.1002/esp.2094.

Nascimento JMP, Dias JMB. 2005. Vertex component analysis: a fast algorithm to unmix hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing* **43**: 898–910. https://doi.org/10.1109/TGRS.2005.844293.

Piégay H, Alber A, Lauer JW, Rollet A-J, Wiederkehr E. 2012. *Biophysical characterisation of fluvial corridors at reach to network scales, in: Fluvial Remote Sensing for Science and Management*. John Wiley & Sons: Chichester; 241–269. https://doi.org/10.1002/9781119940791.ch11.

Piégay H, Arnaud F, Belletti B, Bertrand M, Bizzi S, Carbonneau P, Dufour S, Liébault F, Ruiz-Villanueva V, Slater L. 2020. Remotely sensed rivers in the Anthropocene: State of the art and prospects. *Earth Surface Processes and Landforms* **45**: 157–188. https://doi.org/10.1002/esp.4787.

Rolnick D, Veit A, Belongie S, Shavit N. 2018. Deep learning is robust to massive label noise. arXiv:1705.10694 [cs].

Rossini M, Di Mauro B, Garzonio R, Baccolo G, Cavallini G, Mattavelli M, De Amicis M, Colombo R. 2018. Rapid melting dynamics of an alpine glacier with repeated UAV photogrammetry. *Geomorphology* **304**: 159–172. https://doi.org/10.1016/j.geomorph.2017.12.039.

Rusnák M, Sládek J, Kidová A, Lehotský M. 2018. Template for high-resolution river landscape mapping using UAV technology. *Measurement* **115**: 139–151. https://doi.org/10.1016/j.measurement.2017.10.023.

Samarth G, Neelanjan B, Breckon TP. 2019. Experimental exploration of compact convolutional neural network architectures for non-temporal real-time fire detection. arXiv:1911.09010 [cs, eess].

Steven MD. 1987. Ground truth an underview. *International Journal of Remote Sensing* **8**: 1033–1038. https://doi.org/10.1080/01431168708954745.

Tamminga A, Hugenholtz C, Eaton B, Lapointe M. 2015. Hyperspatial remote sensing of channel reach morphology and hydraulic fish habitat using an unmanned aerial vehicle (UAV): A first assessment in the context of river research and management. *River Research and Applications* **31**: 379–391. https://doi.org/10.1002/rra.2743.

Thanh Noi P, Kappas M. 2018. Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine classifiers for land cover classification using Sentinel-2 imagery. *Sensors* **18**: 18 https://doi.org/10.3390/s18010018.

Vannote RL, Minshall GW, Cummins KW, Sedell JR, Cushing CE. 1980. The river continuum concept. *Canadian Journal of Fisheries and Aquatic Sciences* **37**: 130–137. https://doi.org/10.1139/f80-017.

Willmott CJ, Matsuura K. 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research* **30**: 79–82. https://doi.org/10.3354/cr030079.

Willmott CJ, Matsuura K. 2006. On the use of dimensioned measures of error to evaluate the performance of spatial interpolators. *International Journal of Geographical Information Science* **20**: 89–102. https://doi.org/10.1080/13658810500286976.

Woodget AS, Carbonneau PE, Visser F, Maddock IP. 2015. Quantifying submerged fluvial topography using hyperspatial resolution UAS imagery and structure from motion photogrammetry. *Earth Surface Processes and Landforms* **40**: 47–64. https://doi.org/10.1002/esp.3613.

Zhang J, Foody GM. 2001. Fully-fuzzy supervised classification of sub-urban land cover from remotely sensed imagery: Statistical and artificial neural network approaches. *International Journal of Remote Sensing* **22**: 615–628. https://doi.org/10.1080/01431160050505883.

Carbonneau P, Belletti B, Micotti M, Bizzi S. 2020. Fuzzy Classification of Sentinel-2 data with UAV-derived label data [dataset] [WWW Document].

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Data S1 Document.

Figure S1. Convolutional neural network (CNN) architectures. (Left) The basic CNN architecture used in this work. Ns is the number of samples, S is the X-Y size of the image tile and Nb is the number of image bands used. Note that this network has a single convolutional layer with the filter size equal to the tile size. This returns 1 scalar value per filter and therefore creates 1 new predictor per filter. (Right) Deeper network architectures. These networks use an increasing number of convolution layers as tile size S increases and always with $3 \times 3 \times$ Nb size filters. For example, for tile sizes of 9, we use a total of 4 convolution layers. The final result is again a single scalar predictor for each convolution filter. Note that we do not use image padding after the convolution filters because for such small tiles, the padded area (e.g. 2 pixels out of a width of 9 pixels) would represent a significant portion of the image.

Figure S2. Training performance of compact architectures. Top: model using 5×5 pixel tiles and 128 filters with 10 bands. Bottom: model using 7×7 pixel tiles and 128 filters with 10 bands. Both models have a single convolution layer.

Figure S3. Training performance of deep architectures. Top: model using 5×5 pixel tiles and 128 filters with 10 bands. This model has two convolution layers. Bottom: model using 7×7 pixel tiles and 128 filters with 10 bands. This model has three convolution layers.

Table S1. Long-list model selection results. We use five-fold cross-validation scored with the mean absolute error. The table gives the mean of the five-folds with standard deviation of the five-folds in brackets. Entries highlighted in yellow have a mean absolute error of < 0.10 and an interfold standard deviation < 0.01. These models are kept for the next step.

Table S2. Dominant class error and number of parameters for long-list models. Model names are coded following their parameters: D or C for deep or compact, 5, 7 or 9 for the size of the image tiles, 8, 32, 128 or 512 for the number of filters and 4B or 10B for the 4 band or 10 band cases. For example, the compact architecture model using 7×7 image tiles, 8 filters and 4 bands will be notated: S_7_8_4B. We give mean absolute error (MAE), median error (MDE) and error variance (EVAR).

Table S3. Mann–Whitney p-values for short-listed models A–AB, p-values of less than 0.01 are notated as 0.00. Results in red highlight model pairs which are not significantly different in median rank at the 99% level. Model pairs highlighted in blue fail at the 99% level, but pass at the 95% level. Two top performing models are highlighted in green.

Table S4. Brown–Forsythe p-values for short-listed models A–O, p-values of less than 0.01 are notated as 0. Results in red highlight model pairs which are not significantly different in variance at the 99% level. Model pairs highlighted in blue fail at the 99% level, but pass at the 95% level. Two top performing models are highlighted in green.

Figure S4. Filter responses for the selected compact model using 5×5 image tiles, 32 filters and 10 bands.

Figure S5. Filter responses for the selected deep model using 9×9 image tiles, 128 filters and 10 bands.

Table S5. Comparison of error metrics for the dominant and sub-dominant classes to error metrics for each individual class in the case of Experiment 1.

Table S6. Overall class-based median errors (MDE) and mean absolute errors (MAE) for all experiments.

Figure S6. Class errors. This figure compares the dominant and sub-dominant class error distributions to those associated to each specific class.

Figure S7. Results of naive experiment with all sites of the Po basin (i.e. Rivers Sesia and Po) acquired in 2017 used as both training and validation. Data from sites on the Sesia River presents abnormalities and were discarded from further analysis.

Table S7. Mean absolute errors for downsampling scenarios. Fuzzy classifications were produced with the optimal compact convolutional neural network (CNN) model using a tile size of 5×5 pixels and 32 filters. In the yellow highlighted area, we reproduce values from Table 5 in the main article produced from the original unmanned aerial vehicle (UAV) class rasters at 10 cm spatial resolution.

Figure S8. Effect of resolution. Distributions for the mean absolute dominant class error for optimal compact convolutional neural network (CNN) fuzzy predictions based on unmanned aerial vehicle (UAV) data downsampled from 10 cm to 50 cm, 100 cm and 300 cm. Distributions for Experiments 3b and 3c are poor, but in other cases, the pattern of errors is reasonable indicating that this data may be an acceptable compromise.

