A rapid non-iterative proper orthogonal decomposition based outlier detection and correction for PIV data

2016 Meas. Sci. Technol. 27 125303

(http://iopscience.iop.org/0957-0233/27/12/125303)

OPEN ACCESS
IOP Publishing
Measurement Science and Technology
Meas. Sci. Technol. **27** (2016) 125303 (16pp)
doi:10.1088/0957-0233/27/12/125303

A rapid non-iterative proper orthogonal decomposition based outlier detection and correction for PIV data

J E Higham, W Brevis and C J Keylock

Department of Civil and Structural Engineering, Sheffield Fluid Mechanics Group, Sheffield, South Yorkshire, UK

E-mail: jhigham1@sheffield.ac.uk

Received 15 July 2016, revised 16 September 2016
Accepted for publication 5 October 2016
Published 26 October 2016

Abstract

The present work proposes a novel method of detection and estimation of outliers in particle image velocimetry measurements by the modification of the temporal coefficients associated with a proper orthogonal decomposition of an experimental time series. Using synthetic outliers applied to two sequences of vector fields, the method is benchmarked against state-of-the-art approaches recently proposed to remove the influence of outliers. Compared with these methods, the proposed approach offers an increase in accuracy and robustness for the detection of outliers and comparable accuracy for their estimation.

Keywords: outlier detection, proper orthogonal decomposition, particle image velocimetry, image processing, experimental fluid mechanics

(Some figures may appear in colour only in the online journal)

1. Introduction

Particle image velocimetry (PIV) is a powerful experimental tool used in fluid mechanics to obtain a sequence of two- or three-dimensional vector fields. For this, a turbulent flow is seeded with particles, which respond to, but do not affect, the turbulent flow structures (Adrian and Westerweel 2011). After image pairs are acquired, a vector field can be estimated using cross-correlation, typically in the Fourier domain. Any minor error, such as flaws in the image acquisition, or inhomogeneities of the flow seeding, can lead to poor correlations between image pairs resulting in errors within the vector fields: these errors are often referred to as outliers. Ideally one should try to mitigate against all of these problems, but they are often unavoidable. As PIV sequences can contain thousands of vector fields, numerous contributions have suggested approaches to automatically reduce the influence of outliers.

Typically, these methods fall into three categories: methods which use local spatial statistics of the vector fields to separately detect and estimate outliers; methods which spread or smooth the influence of outliers within the data fields; and methods which use spatio-temporal features obtained from statistical approaches, such as proper orthogonal decompositions (POD), to detect and or estimate outliers.

The most common of these methods are based on spatial statistics. Westerweel (1994) suggested three methods for outlier detection by comparison of local statistics: 'local-mean'; 'local-median'; and 'global-mean'. The 'local-median' method was found to be most accurate, but not practical, as ad hoc thresholds are required for different flow regimes. By normalising the residuals of the local medians with respect to a robust estimate of the local variation of the velocity, the 'local median' method was improved, resulting in the 'universal outlier detection' (UOD) approach (Westerweel and Scarano 2005). This method is popular, but struggles to detect groups of outliers due to their influences on the local statistics. As a consequence, the 'adaptive weighted angle and magnitude threshold method' (AWAMT) (Masullo and Theunissen

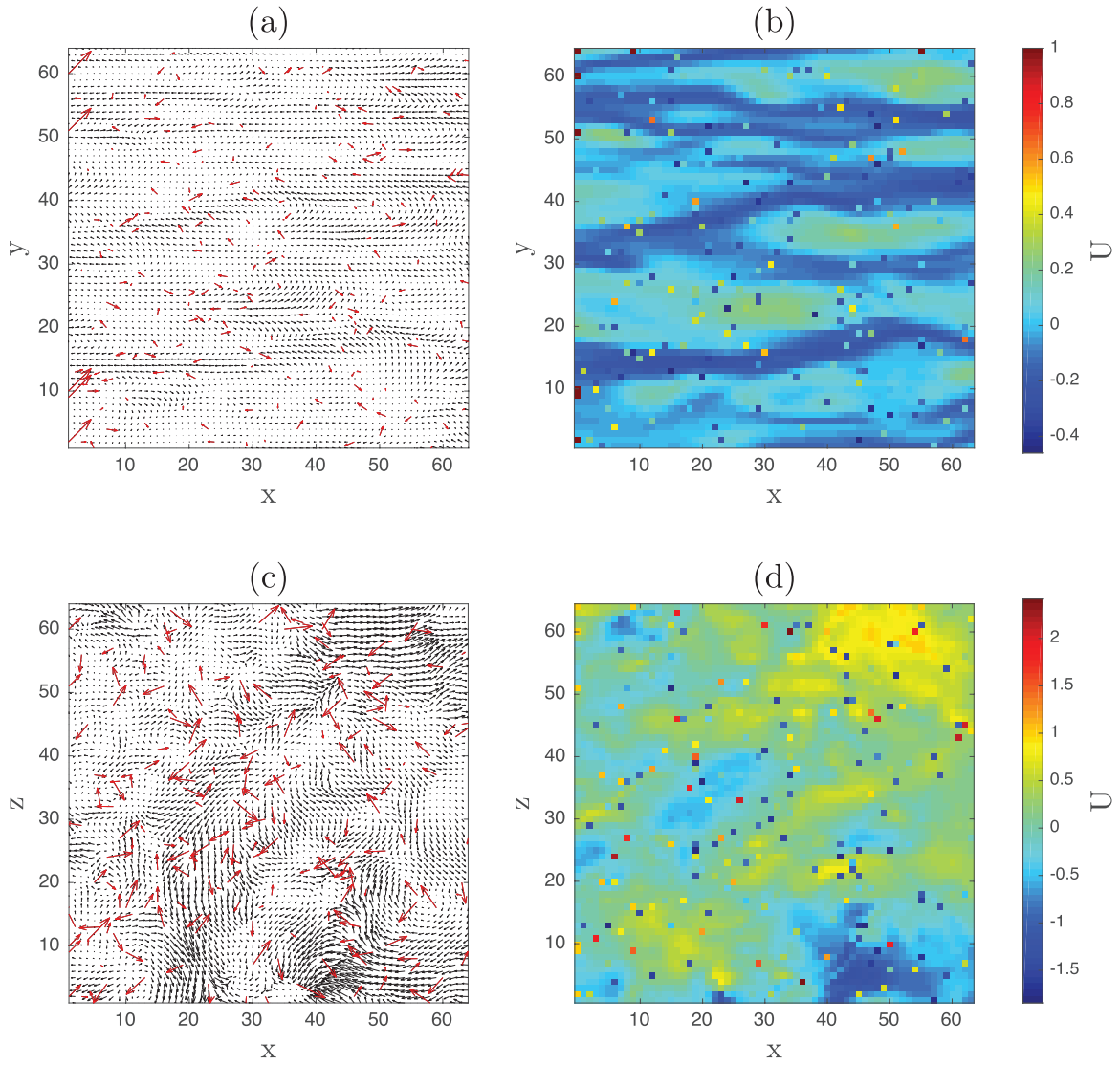

Figure 1. Synthetic vector fields from JHTDB numerical time series. Synthetic outliers are highlighted in red. $Q = 5\%$ and $N_c = 1$ have been used for these examples. (a) and (b) Two-dimensional vector fields and longitudinal velocity magnitude for channel flow, respectively. (c) and (d) Two-dimensional vector fields and longitudinal velocity magnitude for isotropic turbulence, where U is the streamwise component. The 500th vector field in the sequence is shown.

2016) was developed to improve the UOD approach. As in the UOD method, AWAMT detects outliers by comparing local statistics in the local neighbourhood. However, the AWAMT dynamically adapts the size of the neighbourhood to account for larger clusters. Furthermore, AWAMT normalises residuals with respect to a vector's magnitude and angle, adopting a modified Gaussian weighted distance-based averaging median. Masullo and Theunissen (2016) found AWAMT to improve on the UOD method for the detection of clusters of outliers and in the overall accuracy of detection. A more complex method of outlier detection uses cellular neural networks (CNN) to create a detection scheme by obtaining stable states of neurons. However, the robustness and accuracy of the method was found to be comparable only to the 'local-median' method (Liang *et al* 2003).

These methods described above only detect outliers which means that an interpolation scheme is still required. As these methods calculate outliers locally i.e. in single vector fields, it is intuitive to use simple local statistical methods such as linear, bi-linear, spline or more complex mathematical models, such as Kriging (Gunes *et al* 2006). Consequently, these local methods are dependent on the characteristic length scales of the flow and on the resolution of the acquired images. Alternatively, if outliers in several vector fields of the sequence are detected, an iterative POD based method such as 'Gappy POD' (Everson and Sirovich 1995) can be used. Gappy-POD and Kriging are comparable in effectiveness and Gappy-POD has been further developed with the adaptive Gappy-POD formulation (Raben *et al* 2012). However, these methods are computationally expensive and impractical for the long vector field sequences found in some PIV measurements (Gunes *et al* 2006). An alternative method recently proposed in fluid mechanics is the 'all-in-one' method (Garcia 2010, 2011), based on the combination of penalised least squares techniques, discrete cosine

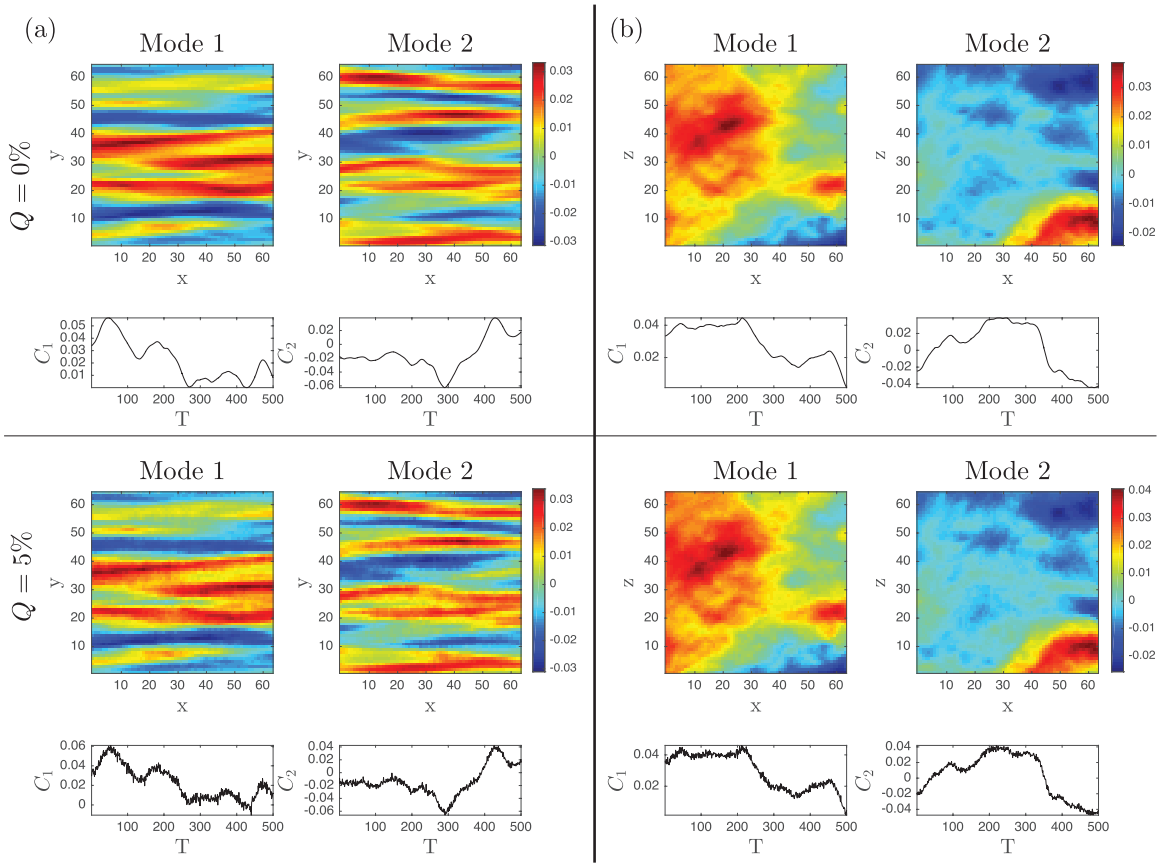

Figure 2. Example of the POD decomposition on the original and contaminated time series. The left column shows the results for the channel flow case. The right column corresponds to the isotropic case flow. The first row shows the results for the original time series. The second row shows the results for a contaminated time series with $Q = 5\%$ and $N_c = 3$. Even though differences in the spatial structures of the first two modes can be observed, the most evident differences can be seen in the noisier structure of the temporal coefficients associated with these modes.

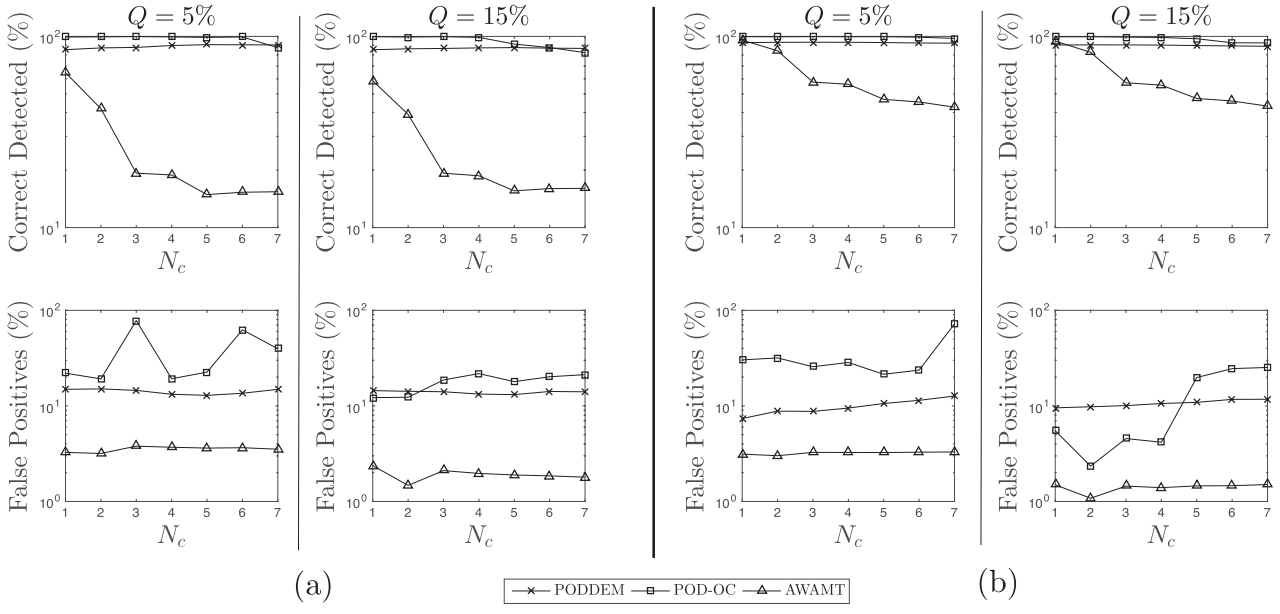

Figure 3. Assessment of the detection capabilities of the outliers introduced in the channel and isotropic flow time series. The plot shows the performance results for PODDEM and benchmark methods. The top row shows the percentage of correct detected outliers as a function of total number of introduced synthetic outliers, for $Q = 5\%$ and $Q = 15\%$. The bottom row shows the number of false positives, similarly expressed as a function of total number of introduced synthetic outliers. (a) Channel flow, (b) isotropic flow.

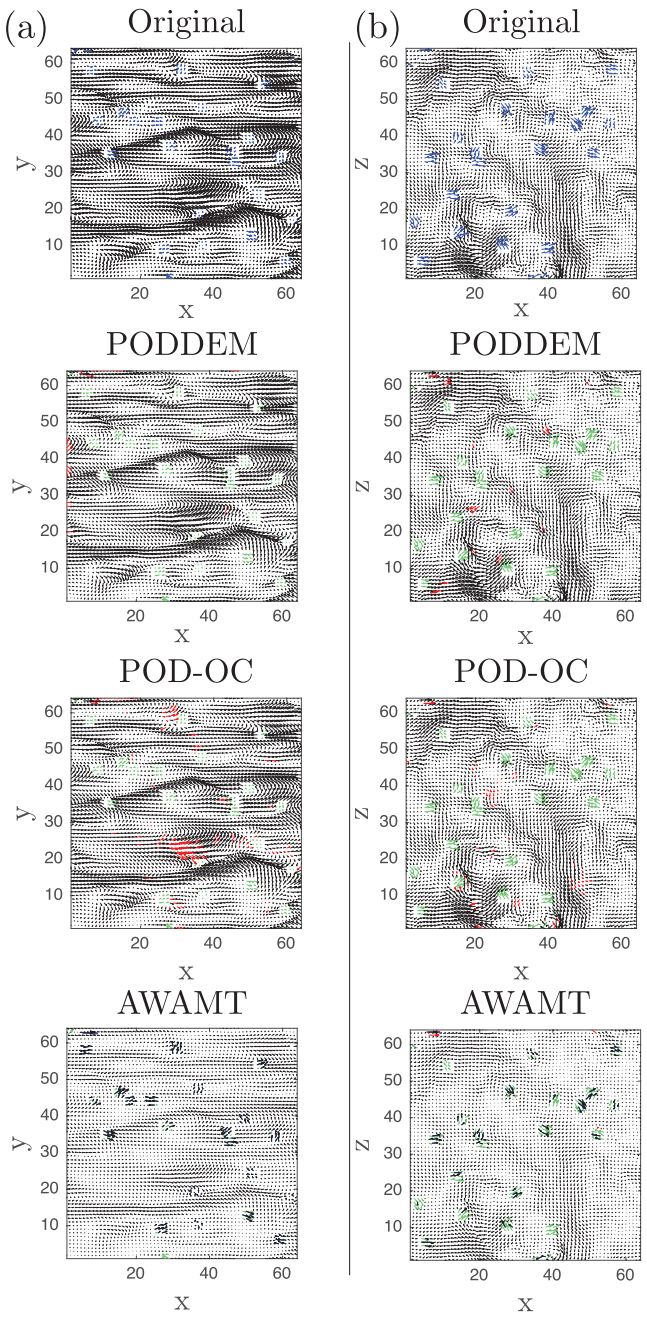

Figure 4. Example of the detection on a single vector field using the benchmarked detection methods. For this, a $Q = 5\%$, $N_c = 3$ have been used. (a) Channel flow (b) isotropic flow. The black vectors show the original flow, blue show the applied synthetic outliers, green show the correct detections and red show the false positive detections. The 500th vector field in the sequence is shown.

transforms and the generalised cross-validation method. Whilst the aim of the method is to reduce the influence of the outliers, Wang *et al* (2015) notes this method can weaken instantaneous velocity fluctuations and gradients.

A recent iterative, spatio-temporal statistical method, which couples the detection and estimation of outliers using a POD, is the POD-outlier correction method or POD-OC (Wang *et al* 2015). This method assumes that outliers do not perturb into the low-order POD spatial modes of a flow decomposition. It detects outliers by comparing each vector field to a residual calculated from the mean, standard deviation and a 'robust parameter' ($a = 3$). The detected points are replaced using a low-order reconstruction and a second residual, this process is repeated until convergence of the POD spectrum. The author applied this method to a sequence of vector fields with good success.

In the present work a novel non-iterative alternative POD based method, termed 'POD detection and estimation', or PODDEM, is proposed. The proposed method is benchmarked using two datasets obtained from the John Hopkins Turbulence Database (JHTDB) (Li *et al* 2008), which have been modified to introduce synthetic outliers, and from real PIV data obtained from Hain and Kähler (2007).

This paper is structured as follows: in section 2, the POD method, the process used to create synthetic outliers, and the PODDEM algorithm are described; in section 3 the PODDEM algorithm is benchmarked in terms of its detection capabilities followed by an assessment of its ability to estimate the detected vectors. These benchmarks are constructed for a time series of vector fields, and a single vector field respectively. Section 4 discusses a number of improvements, and suggestions for POD-based outlier detection methods. Finally in section 5 the main conclusions are presented.

2. Proper orthogonal decomposition detection and estimation

2.1. Proper orthogonal decomposition (POD)

POD is a statistical method commonly used in fluid mechanics for the extraction and analysis of energy meaningful turbulent structures (Aubry 1991, Berkooz *et al* 1993). POD was independently derived by a number of individuals, consequently acquiring a variety of names in different fields including Karhunen–Loève decomposition, singular value decomposition (SVD) and principal components analysis (PCA) (Kosambi 1943, Loève 1945, Karhunen 1946, Pugachev 1953, Obukbov 1954). POD extracts energy relevant structures (modes) from set of a stochastic, statistically steady-state turbulent fields, within a finite time domain, ordering them by their contribution to the total variance of the physical property being analysed, e.g. velocity (Brevis and García-Villalba 2011). A set of $t = 1, 2, \ldots, T$ temporally ordered vector fields, $\mathbf{V}(x, y; t)$, is considered, each of which is of size $X \times Y$. The method requires the construction of a $N \times T$ matrix $\mathbf{W}$ from T columns $\mathbf{w}(t)$ of length $N = XY$, each column corresponding to a column-vector version of a transformed snapshot $\mathbf{V}(x, y; t)$. A POD is obtained by:

$$\mathbf{W} \equiv \mathbf{\Phi}\,\mathbf{S}\,\mathbf{C}^{\mathrm{T}} \tag{1}$$

where $\mathbf{S}$ is a matrix of size $\Omega \times \Omega$, (Ω are the number of modes of the decomposition, and $(\cdot)^T$ represents a transpose matrix operation). The $\lambda = \mathrm{diag}(\mathbf{S})^2/(N-1)$ is the vector containing the contribution to the total variance of each Ω. The elements in λ are ordered in descending rank order, i.e. $(\lambda_1 \geqslant \lambda_2 \geqslant \ldots \lambda_\Omega \geqslant 0)$. In practical terms the matrix $\mathbf{\Phi}$ of size

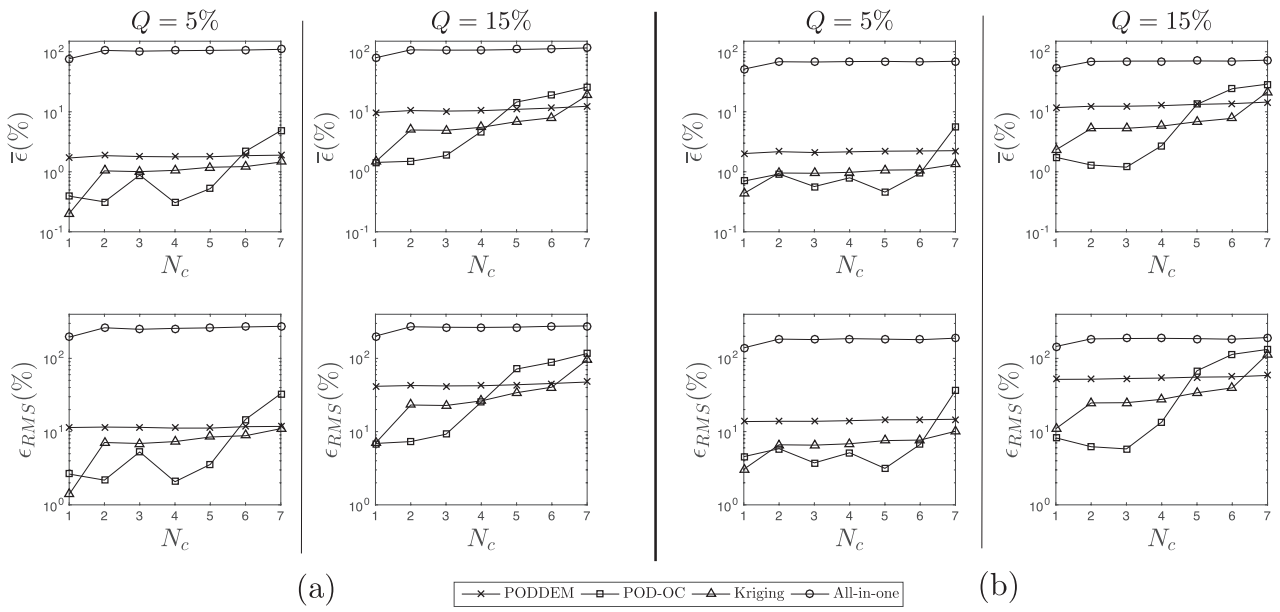

Figure 5. The plots show a comparison of the performance for the estimation of the correct value of outlier vectors between PODDEM, POD-OC, all-in-one and Kriging methods, where all locations of the outlier points are known. The top row shows $\overline{\epsilon}$ (accuracy) and bottom rows show the spatio-temporal ϵ_{RMS} (precision) error. (a) Shows the error obtained with the contaminated channel case. (b) Shows the results for the isotropic case.

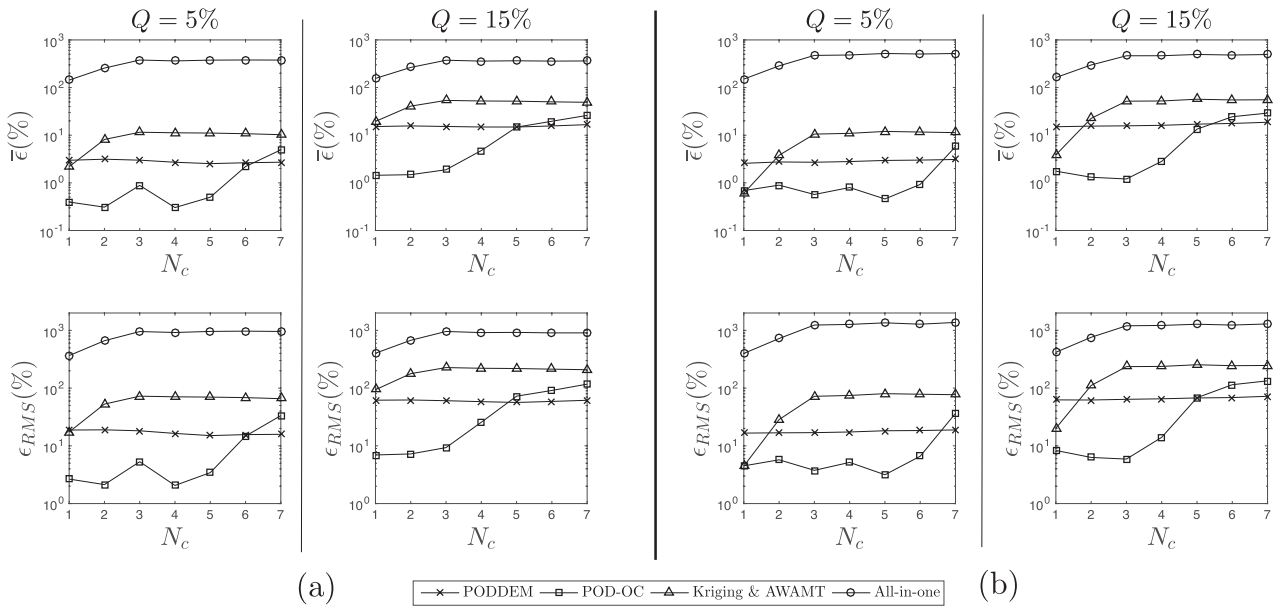

Figure 6. The plots show a comparison of the performance for the estimation of the correct value of outlier vectors between PODDEM, POD-OC, all-in-one and the coupled AWAMT & Kriging methods, where all locations of the outlier points are unknown. The top row shows $\overline{\epsilon}$ (accuracy) and bottom rows show the spatio-temporal ϵ_{RMS} (precision) error. (a) Shows the error obtained with the contaminated channel case. (b) Shows the results for the isotropic case.

$N \times \Omega$ contains the spatial structure of each of the modes and the matrix $\mathbf{C}$ of size $\Omega \times \Omega$ contains the coefficients representing the time evolution of the modes.

2.2. POD outlier detection & estimation method (PODDEM)

The present study suggests a methodology for the detection and estimation of outliers in every vector field of a dataset, through the modification of the results of a POD. Unlike other POD-based methods, the proposed method is non-iterative, and hence less computationally expensive. Alternative POD-based methods are built on modifications of $\mathbf{\Phi}$, while the present one relies on changes to $\mathbf{C}$. The present method is based on the observation that outliers in every vector field in a time series can produce spikes or a noisy evolution of $\mathbf{C}$ (see figure 2). The hypothesis of this work is that a suitable correction of $\mathbf{C}$ can be

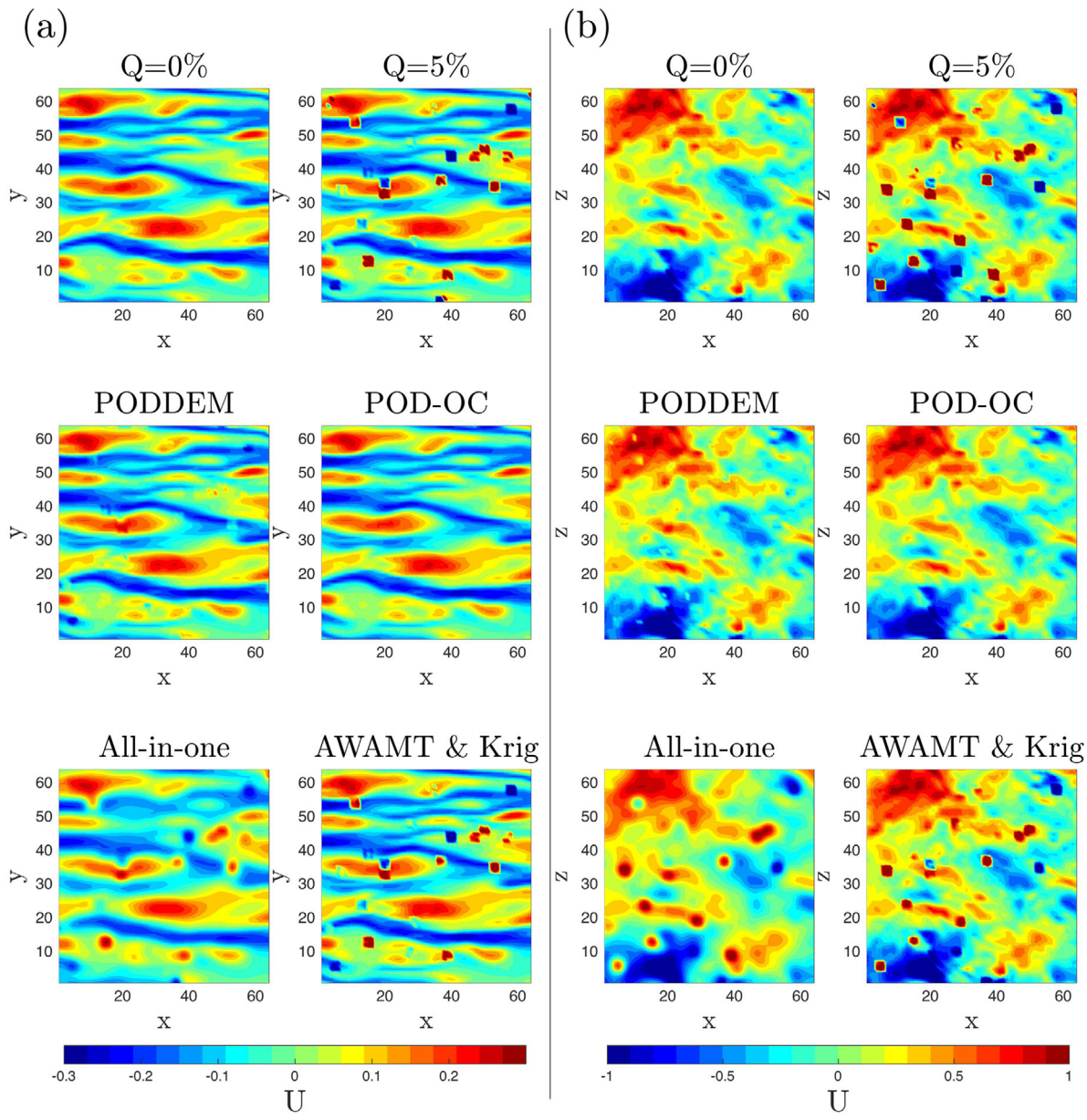

Figure 7. Example of the estimated instantaneous longitudinal velocity component after the application of the benchmarked estimation methods. For this, a $Q = 5\%$, $N_c = 3$ have been used. (a) Channel flow (b) isotropic flow. The original snapshot, with no outliers, i.e. $Q = 0\%$, and with $Q = 5\%$ are shown in the top row for reference, where U is the streamwise component. The 500th vector field in the sequence is shown.

Table 1. Comparison of computing time between PODDEM and benchmark algorithms.

	PODDEM	POD-OC	AWAMT & Krig	All-in-one
Channel	1.0	55	4073.6	1.1
Isotropic	1.0	68	2905.8	0.8

Note: The values are normalised with respect to the PODDEM calculation time.

used to reduce the influence of outliers in the time series. As summarised in algorithm 1, this is achieved as follows:

- A POD, as shown in equation (1), is performed on the input matrix $\mathbf{W}$ (in this study only two velocity components are used).
- A moving average is performed on each POD coefficient vector $\mathbf{c}_n$, where $n = 1 \ldots \Omega$. These vectors correspond to column components of $\mathbf{C}$. In this work a convolution kernel size of 0.01 of the average integral time scale, $0.01\tau_I$, was used during the moving averaging procedure. $0.01\tau_I$ for a kernel size was found to be effective in the test cases presented in the present study, as the kernel was large enough to remove the smaller scale noise, but not large enough to affect the temporal evolution of $\mathbf{C}$, a sensitivity analysis can be found in the appendix (figures A1 and A2). The resulting vectors are stored in $\mathbf{C}^E$, an estimated version of $\mathbf{C}$.
- A new $\mathbf{W}^E$ is created, using equation (1), $\mathbf{W}^E = \mathbf{\Phi S}(\mathbf{C}^E)^T$.
- A matrix $\mathbf{W}'$ is created from $|\mathbf{W} - \mathbf{W}^E|$, where $|\cdot|$ represents the absolute value operation.

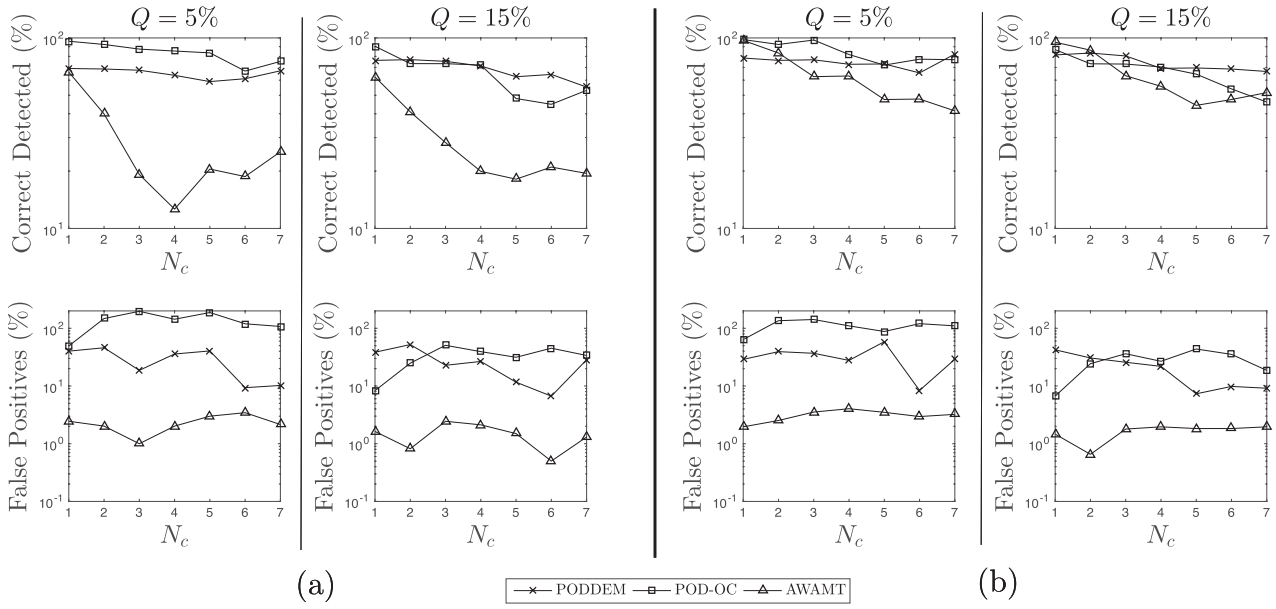

Figure 8. Performance comparison between PODDEM, POD-OC and AWAMT for the detection of outliers in case of a single contaminated vector field (500th). The top row shows the correct outliers detected. The bottom row shows the percentage of false outliers detected. (a) Results for channel flow. (b) Results for isotropic case. The snapshot has been transformed in an ensemble by using sub-fields of size $n_b \times m_b = 16 \times 16$.

Algorithm 1. PODDEM

Require: A sequence of T vector fields transformed into a matrix $\mathbf{W}$ and a user defined percentage t_r%.

Output: A matrix $\mathbf{W^c}$ with outliers removed. The columns of $\mathbf{W^c}$ can be reshaped to obtain a filtered version of the sequence of vector fields.

$\mathbf{W} \leftarrow \{w_1, w_2 \ldots w_T\}$,

$[\mathbf{\Phi}, \mathbf{S}, \mathbf{C}] \leftarrow \mathrm{SVD}(\mathbf{W})$.

$\mathbf{C}^E \leftarrow \mathbf{c}_n^E \leftarrow \mathrm{movingaverage}(\mathbf{c}_n, 0.01\tau_I)$; where $n = 1 \ldots \Omega$,

$\mathbf{W}^E \leftarrow \mathbf{\Phi S C}^{E^T}$,

$\mathbf{W}' \leftarrow |\mathbf{W} - \mathbf{W}^E|$,

$\mathbf{M_{ij}} \leftarrow 1$,

$\mathbf{M_{ij}} \leftarrow 0$, corresponding to locations of top t_r% of sort($\mathbf{W}_{ij}$),

$\mathbf{W}^c \leftarrow \mathbf{W} \cdot \mathbf{M} + \mathbf{W}^E \cdot (1 - \mathbf{M})$.

- Similar to previous approaches, a mask matrix $\mathbf{M}$ of the same size as $\mathbf{W}$ is introduced, in which each element is assigned the value 1.
- The elements of $\mathbf{W}'$ are sorted in descending order. The locations of $\mathbf{W}'$ corresponding to the first t_r% (user defined percentage, relating to the ratio of the number of outliers to total number of vectors in the dataset) of the sorted $\mathbf{W}'$ are assigned a 0 in $\mathbf{M}$.
- Using a simple operation a corrected version $\mathbf{W}^c$ of $\mathbf{W}$ is obtained: $\mathbf{W}^c = \mathbf{W} \cdot \mathbf{M} + \mathbf{W}^E \cdot (1 - \mathbf{M})$, where $\cdot$ corresponds to the inner product operation. More simply: the valid data, i.e. those with elements of $\mathbf{M}$ with value 1, are retained, while the detected outliers are replaced by those calculated in $\mathbf{W}^E$.

2.3. Selection of test cases

In this study two datasets from the JHTDB are used for a quantitive assessment. These data are chosen due to the availability of long time series. 1000 vector fields are selected for each case, each of them containing 64×64 grid points. The first dataset selected is a subset of a direct numerical simulation (DNS) of a channel flow (Graham *et al* 2016). The origin of the selected section is located at $x = 18.2$, $y = -0.99$, and $z = 6.6$. From that point, 64 points are taken in the x and y positive direction, at a spacing of 0.01. The selected domain size is equal to $8\pi \times 2\pi \times 3\pi$. For the construction of the time series, this region was sampled with a $\delta t = 0.012$; on average the dataset contains 9 integral time scales, $\tau_I = 9$. The second dataset is a subset from the DNS of a forced homogenous isotropic turbulence. The origin of the selected region was located at $x = 0$, $y = 0$, and $z = 0$. From the origin, 64 points are taken in the x and z positive direction, at a spacing of 0.015. The total size of the sampled region is $2\pi \times 2\pi \times 2\pi$. The temporal sampling is performed with a $\delta t = 0.012$ and on average $\tau_I = 6$.

According to Shinneeb *et al* (2004), PIV measurements can contain two types of outliers: single spurious vectors, and clusters of spurious vectors, the latter of the two being more common. As the datasets obtained from JHTDB are outlier-free, synthetic outliers were introduced in the time series. For comparison purposes the same method of synthesising outliers developed by Wang *et al* (2015) is used to benchmark the proposed method. An outlier rate is introduced, Q, defined as the percentage of outliers in each vector field. In the case of single distributed outliers, a random location is obtained from a uniform random function. Similarly to previous works, the magnitude of the x and y components of these outlier

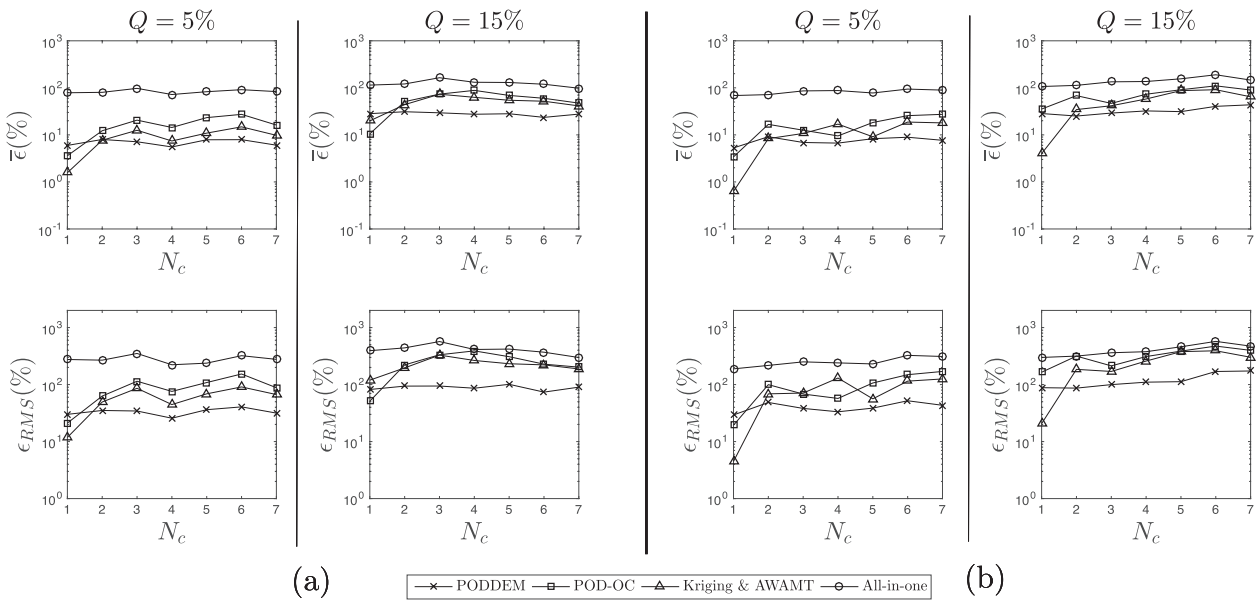

Figure 9. $\overline{\epsilon}$ (accuracy) and ϵ_{RMS} (precision) of PODDEM and POD-OC vector estimations for a single contaminated vector field (500th). The top row shows $\overline{\epsilon}$ (accuracy) and bottom rows show the spatio-temporal ϵ_{RMS} (precision) error. (a) Shows the error obtained with the contaminated channel case (b) shows the results for the isotropic case. The snapshot has been transformed in an ensemble by using sub-fields of size $n_b \times m_b = 16 \times 16$.

vectors are drawn independently from a uniform distribution $(-u_{\max}, u_{\max})$, where $u_{\max}$ is the maximum magnitude of velocity in the entire data. In the case of clusters of outliers, an analogous approach to the one presented by Shinneeb *et al* (2004) is adopted, in which, a parameter N_c is also introduced. This parameter defines the number of vectors involved in a certain cluster of size $f(N_c)$; however, the total number of outlier vectors in each snapshot remains defined by Q. A distribution similar to the one used by Garcia (2011) is adopted for the determination of the size of the clusters:

$$f(N_c) = A \cdot \exp(-N_c^2/\sigma^2) \tag{2}$$

where σ is the standard deviation of the size distribution, and A is a parameter defining the size of a cluster corresponding to the mean number of elements. Different sizes of outliers are distributed throughout the datasets. As in Wang *et al* (2015), the vectors within a cluster also are of a similar magnitude and values of $A = 0.4$ and $\sigma = 2.8$ are used. Several cases are tested in this work, involving $Q = 5\%$ and $Q = 15\%$. For these outlier rates, outlier clusters in the range $1 \leqslant N_c \leqslant 7$ are analysed. In figure 1, an example of a generated synthetic vector field is presented, where $Q = 5\%$ and $N_c = 1$.

Figure 2 shows the spatial and temporal structure of the two leading POD modes for both test cases. It also shows the changes introduced by the outliers on the modes structure when $Q = 5\%$ and $N_c = 3$ are introduced. In both cases, the general patterns in the leading spatial modes remain as in the original time series, but with a grainier structure. Qualitatively speaking, a more obvious effect of the outliers in the POD can be observed in the temporal POD coefficients. In both time series, it is observed that temporal behaviour can be effected by noise. While it is clear that, noise reduction in the two-dimensional spatial structure is possible, the strategy of correcting the temporal behaviour of the modes is followed in this work, leading to the development of PODDEM.

To supplement the quantitive assessment, a third 'real' experimental dataset is used, namely that of of the turbulent flow over periodic hills (Hain and Kähler 2007). This data set contains single frame particle images acquired in the central plane of the channel, using hollow glass particles of $d = 10$ μm illuminated with a 5 W Nd:YAG cw-laser and recorded by means of a Phantom v12 camera. PIV is undertaken on 1000 sequential images using PIVLab (Thielicke and Stamhuis 2014); two passes are undertaken using interrogation windows of size 64×64 and 32×32 respectively, each with a 50% overlap. It is found that, on average, the data set contains $\tau_I = 6$. No synthetic outliers are introduced to the dataset.

2.4. Quantification of algorithm performance

An assessment of the algorithm's performance requires the introduction of criteria for error quantification. All elements are considered to establish the effect of false positive detections and following estimations on the error statistics. Following the criteria defined by Wang *et al* (2015), the relative error ϵ_i between an unmodified element (obtained prior to the application of synthetic outliers) of the matrix $\mathbf{W}$, w_i, and its estimated value w_i', can be defined as:

$$\epsilon_i = \frac{|w_i' - w_i|}{|w_i|}, \tag{3}$$

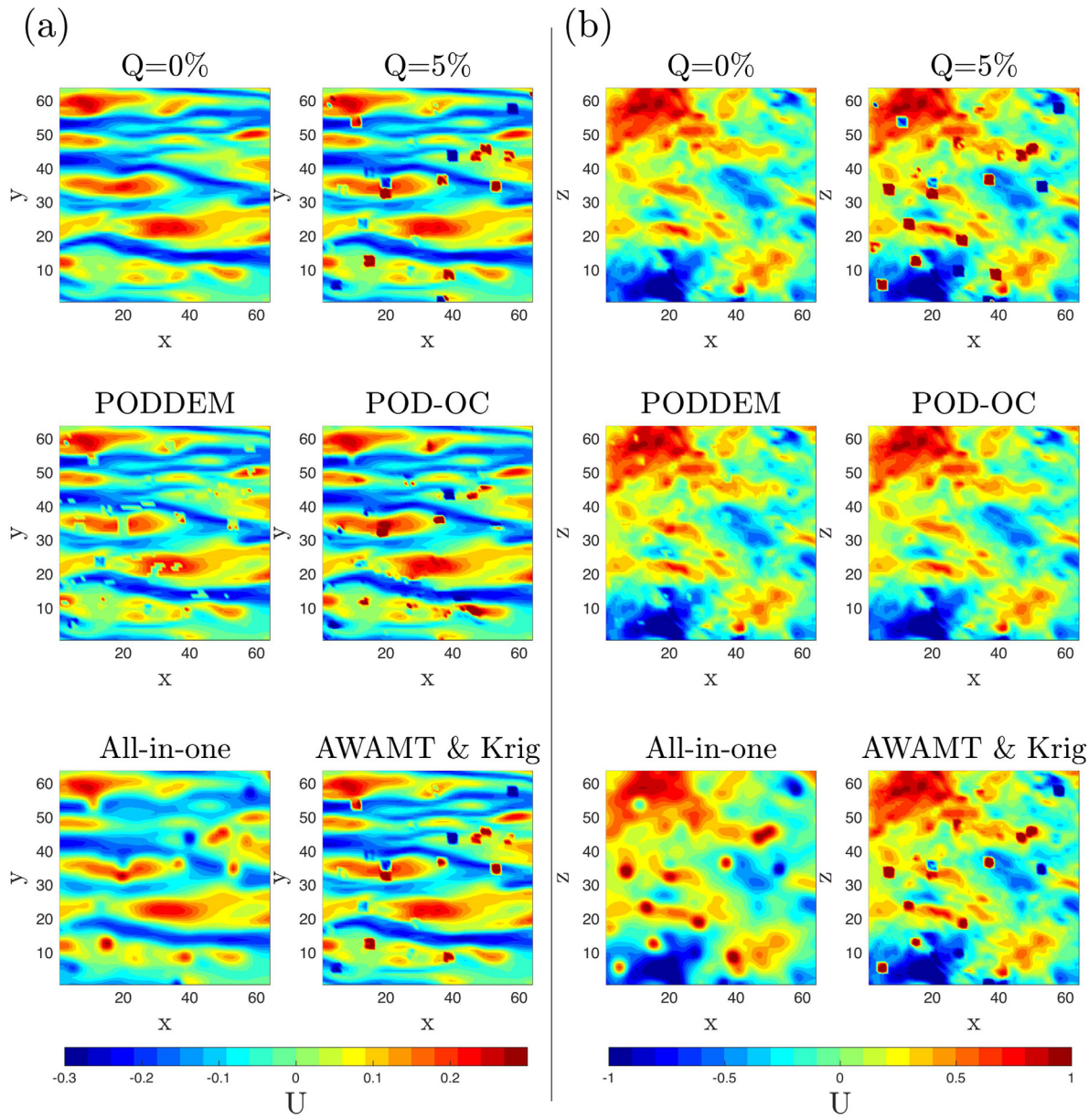

Figure 10. Estimation of the vector field associated to a single contaminated snapshot. The top rows show the original frame, $Q = 0\%$ and the contaminated frames with $Q = 5\%$ of outliers and $N_c = 3$. The bottom row shows the estimation obtained using the PODDEM and POD-OC. (a) Results for channel flow. (b) Results for the isotropic flow. The vector field has been transformed in an ensemble by using sub-fields of size $n_b \times m_b = 16 \times 16$, where U is the streamwise component. The 500th vector field in the sequence is shown.

where the sub-index $i = 1...NT$ represents individual elements of $\mathbf{W}$. This means that the double-averaged error, i.e. spatial and temporal averaged relative error, $\overline{\epsilon_i}$, can be calculated as:

$$\overline{\epsilon_i} = \frac{1}{NT}\sum_{i=1}^{NT}\epsilon_i, \tag{4}$$

Using this definition the spatio-temporal root mean square (RMS) of the relative error can be calculated as:

$$\epsilon_{\mathrm{RMS}} = \sqrt{\frac{1}{NT}\sum_{i=1}^{NT}(\epsilon_i - \overline{\epsilon_i})^2}. \tag{5}$$

Hence equation (4) is a means of characterising the accuracy of the various methods, whilst equation (5) is a measure of precision. A number of methods are chosen in order to benchmark the estimation functionality of PODDEM. The first method is the so-called POD-OC (Wang *et al* 2015). As POD-OC has shown an increased accuracy in comparison

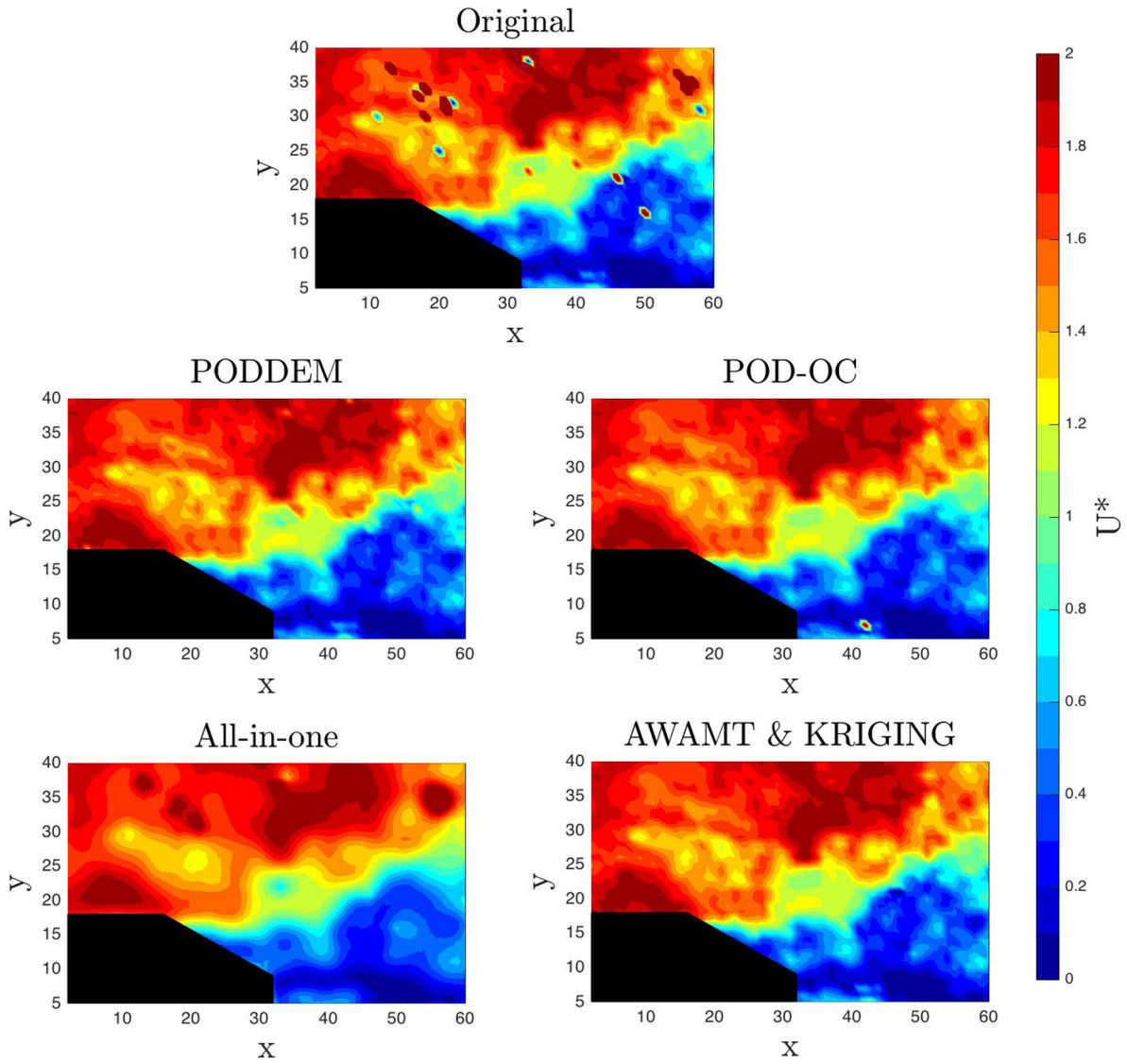

Figure 11. An example of the application of PODDEM, POD-OC, all-in-one and AWAMT & Kriging to real PIV data to a time series of data. Where U^* is defined at the velocity magnitude. As previous the 500th vector field is presented.

with standard statistical methods, e.g. global-mean and linear interpolation, these latter methods are omitted from further consideration. The second comparative method is the all-in-one smoothing function of Garcia (2010), which is implemented using the MATLAB function 'smoothn'. Kriging has also been used for benchmarking, as this method has shown good performance in some of the tests presented by Wang *et al* (2015) and Gunes *et al* (2006). This method has been implemented in the DACE toolbox for MATLAB (Nielsen *et al* 2002), with a second-order polynomial regression and a Gaussian correlation model (Raben *et al* 2012). The detection performance of the PODDEM is quantified by benchmarking the result with the POD-OC method (Wang *et al* 2015) and with the AWAMT method introduced by Masullo and Theunissen (2016). As Kriging is solely an interpolation method and AWAMT is purely a detection method, these two methods are coupled when examining estimation and detection.

N.B. the comparisons of POD-OC and AWAMT are computed using algorithms obtained from the authors. For AWAMT the user defined options are set to the default settings, as outlined in Masullo and Theunissen (2016).

3. Results

3.1. Detection ability

Figure 3 shows a comparison of the methods outlined above when used to identify the location of the synthetic outliers introduced in the time series. In this work, a correct detection is defined as the detection of a velocity vector belonging to the introduced list of synthetic outliers, while the performance is measured as a percentage of the total number of introduced outliers. A false positive is defined as a velocity vector detected as outlier, but not belonging to the original outlier list; similarly, the performance is measured as a

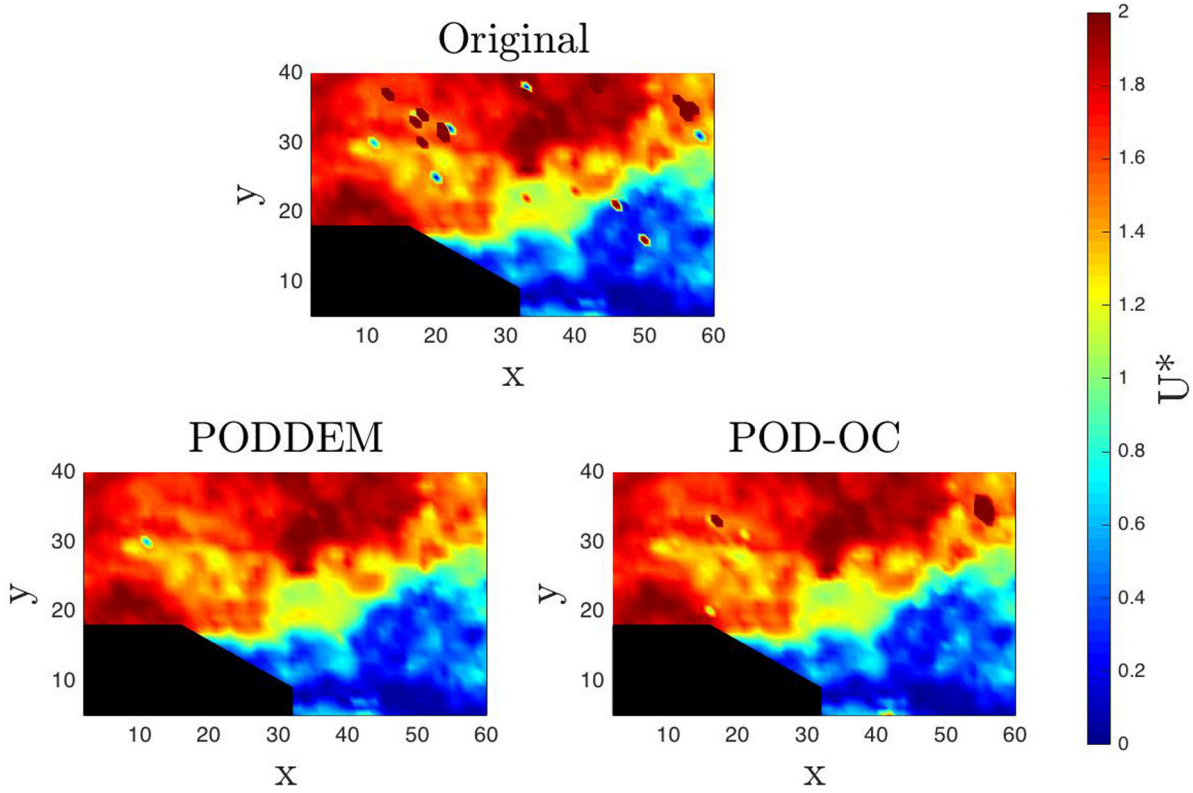

Figure 12. An example of the application of PODDEM, POD-OC, single frame of real PIV data (vector field 500th). Where U^* is defined at the velocity magnitude. The vector field has been transformed in an ensemble by using sub-fields of size $n_b \times m_b = 16 \times 16$.

percentage of the total number of introduced synthetic outliers. Only a subset of the estimation methods also have the capability to detect outliers, and thus only PODDEM, POD-OC and the AWAMT methods are benchmarked in this section. As shown in figures 3 and 4, PODDEM performs similarly to the POD-OC for the detection of correct outliers positions. However PODDEM shows a higher reliability as it has a lower rate of detection of false positives. Of all the benchmarked algorithms, AWAMT detects the least false positives, but as the size of N_c (the size of the cluster), is increased AWAMT becomes less effective in detection. A noticeable benefit of PODDEM is its constant performance in detection, as only a minor difference in its detection ability is seen between outlier rates and test cases.

3.2. Estimation ability

In order to gain a rich perspective on the different methods, they are all are examined twice. Firstly, the methods are examined purely on the basis of their estimation ability i.e where all of locations of the outlier points are known. Secondly, the methods are examined on their coupled estimation and detection ability i.e. where the locations of the outlier points are unknown. The accuracy ($\overline{\epsilon}$) and precision (ϵ_{RMS}) of the methods are presented in figures 5 and 6. Figure 5 shows that when all of the locations of outliers are known and the methods are used solely for interpolation, POD-OC and Kriging for clusters $N_c \leqslant 4$, are the most accurate and precise methods. However, with a higher outlier rate ($Q = 15\%$), for clusters i.e. $N_c > 4$, PODDEM is the more accurate and precise. For detection, the accuracy of PODDEM remains constant, regardless of the size of N_c. Figure 6 demonstrates that, even when coupled with the detection functionality the PODDEM's error remains constant. Between the test cases the results for PODDEM are similar, unlike any of the other methods; this suggests that the accuracy of PODDEM could be independent of the test case, and only dependent on the outlier rate Q. A qualitative comparison of the spatial characteristics of the estimation by the different methods is presented in figure 7. It is clear from the figure that the small scale details of the flow are retained by both POD based methods. The AWAMT method has struggled to detect all of the outlier clusters, resulting in a the vector field which still contains a number of errors. The all-in-one method clearly filters small scale structures thus producing a blurred estimation of the vector field. (N.B. the all-in-one method can be used to interpolate missing values, as shown or to remove influences of outliers making a new estimate of the whole field, as shown in figure 7.)

An estimated computational efficiency of the calculation under the current implementation is shown in table 1. Of course, a computational performance assessment depends on many factors, such as the programming technique and programming language. So as to exclude such variables, the computations were all undertaken on the same computer, using MATLAB R2015b, and restricted to a single core. The results are normalised with respect to the PODDEM method. It is

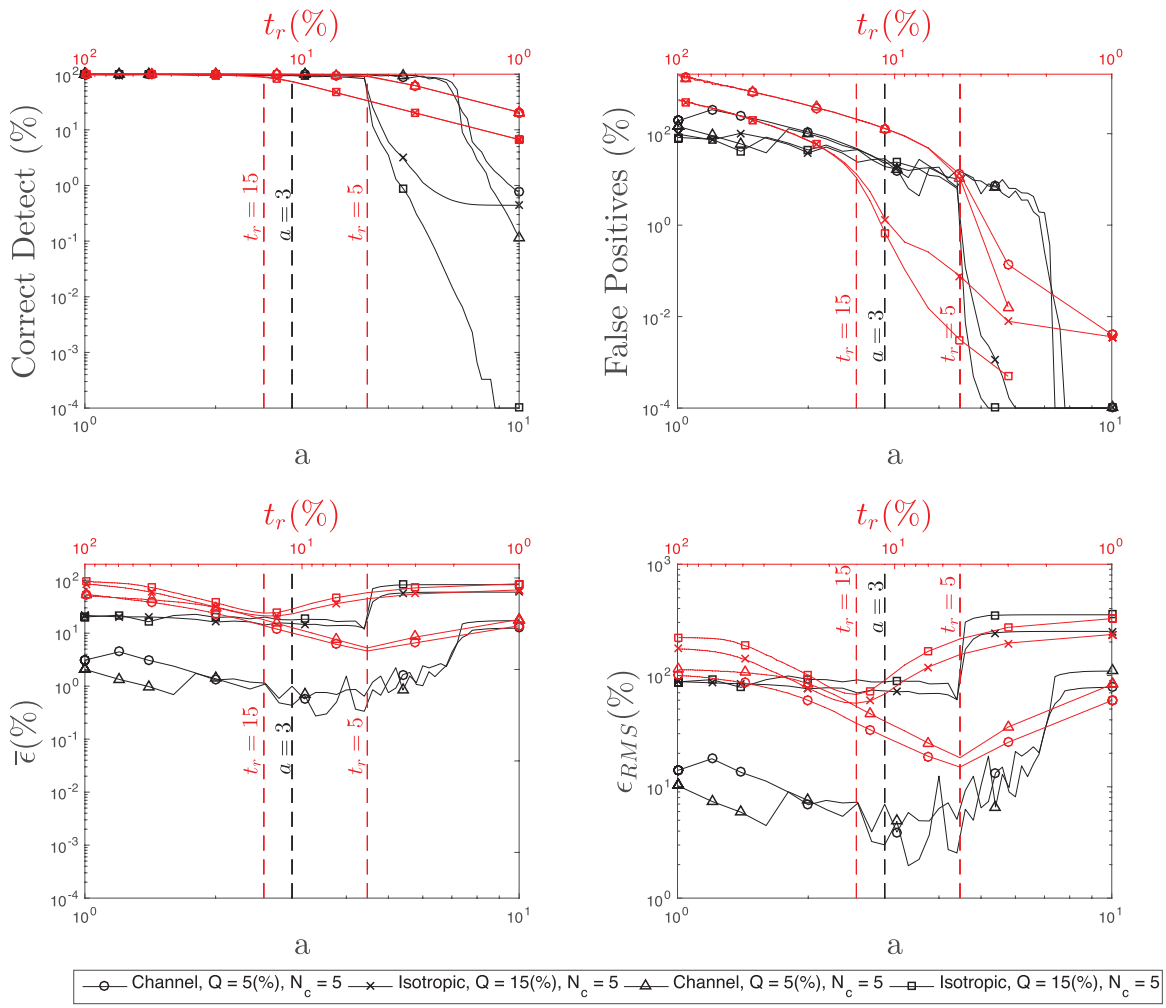

Figure 13. Assessment of varying the 'robust parameter', a, in the POD-OC algorithm (black), and the user defined threshold, t_r, in PODDEM (red). Top panels show the percentage of correction detections and false positives. Bottom panels show $\overline{\epsilon}$ (accuracy) and ϵ_{RMS} (precision).

found that under these conditions, the PODDEM's time efficiency is comparable to that for the all-in-one method and far superior to that for other methods.

The SVD which is at the core of the PODDEM and POD-OC methods is memory intensive. As illustrated by the test cases where only between 6–9 integral time scales are used, the dataset could be temporally partitioned (assuming it is statically converged) if memory is limited. PODDEM offers a substantial time benefit compared with POD-OC, which requires a minimum of two SVDs while PODDEM only ever requires one.

3.3. Detection and estimation on a single vector field

In this section the detection and estimation capabilities of the proposed algorithm are tested not for a sequence of vector fields, but instead for a single contaminated field. This case has been selected to demonstrate that PODDEM can still be used if time resolute data is not available. PODDEM is benchmarked in a manner analogous to the benchmark performed by Wang *et al* (2015). A single vector field (500th) is sub-divided into multiple sub-fields, which are used to build an ensemble of observations. Wang *et al* (2015) showed that the size of the number of sub-fields is critical: a larger size of the sub-fields will offer more spatial information, but will reduce the number of ensemble components. This means the total number of modes involved in the decomposition of the ensemble will be reduced. According to Wang *et al* (2015), the ensemble construction is more effective for filtering when the ratio between the sub-fields size, $n_b \times m_b$, and the size of the original snapshot ($N \times M$), R_B is between 0.2 and 0.5. Wang *et al* (2015) also recomends creating the ensemble for overlapping sub-fields i.e. a one vector element shift along both x and y, thereby increasing number of fields; accordingly, a sub-field of size $n_b \times m_b = 16 \times 16$ was chosen.

The results of the detection assessment are shown in figure 8. Much as in the results in previous section, PODDEM shows a better detection performance than POD-OC in terms of the percentage of correct outliers identified. Whilst

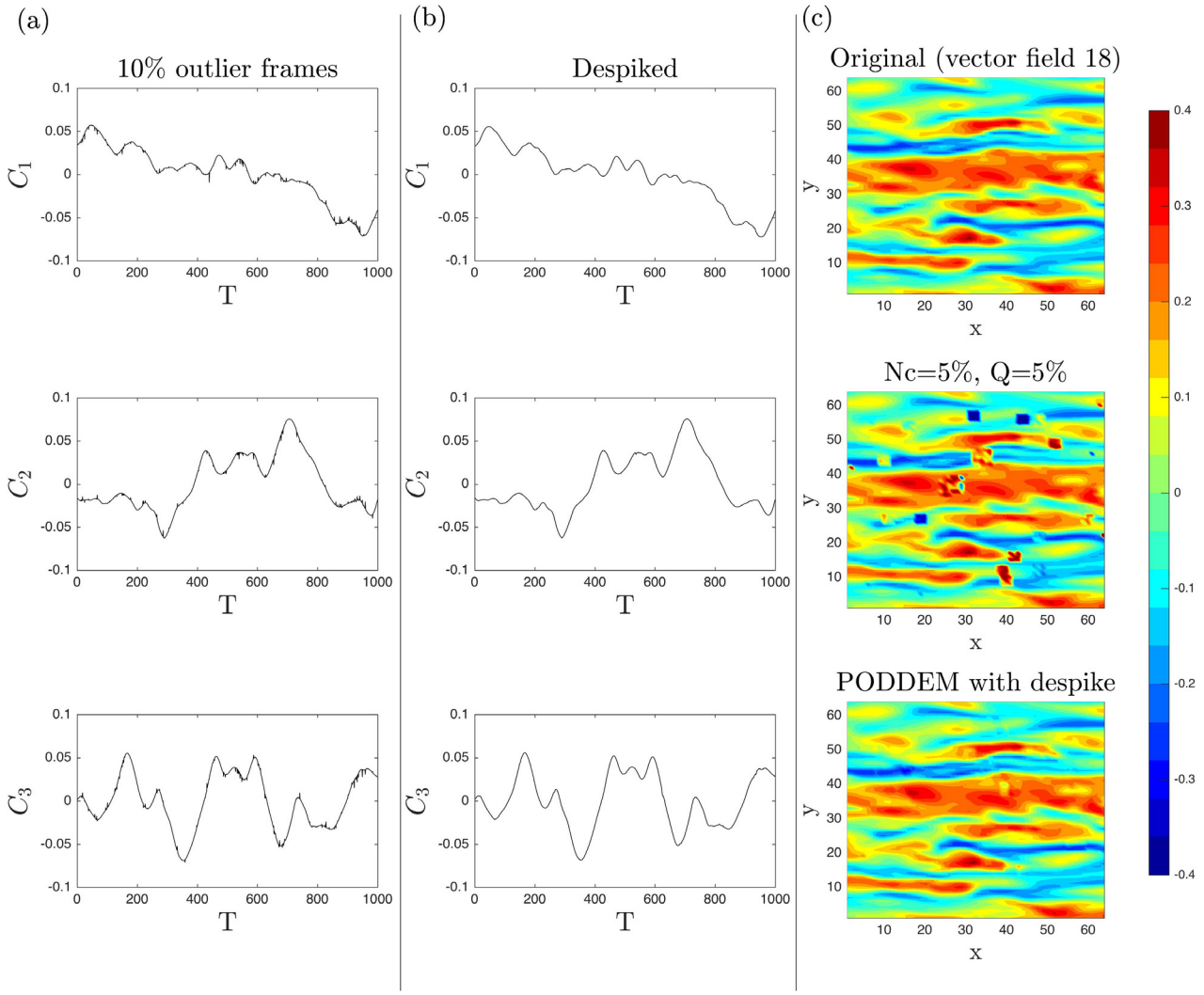

Figure 14. An example of outliers, $Nc = 3$ & $Q = 5\%$, applied to 100 random frames within the time series (vector field 18 shown). It is seen their locations perturb into the temporal coefficients.

the AWAMT method does not detect many false positives, its detection ability decreases as the cluster size increases again. Between the POD-based methods, a difference can be observed when the percentages of false positives are compared. The higher reliability of PODDEM in this regard is evident from the results for both channel and isotropic case flow.

The errors in estimating the detected outliers for the single contaminated vector field are shown in figure 9. It is observed from the figure that PODDEM offers the most robust, accurate and precise estimation of the vector field. In particular, the improvement in the precision statistics when using PODDEM are clear. A qualitative comparison of the spatial characteristics is presented in figure 10. The results of the estimation highlight some of the limitations of the POD based methods for the estimation of a single frame. However, the quantitative and qualitative results show that PODDEM improves estimates compared to those obtained using POD-OC, AWAMT & Kriging and the all-in-one method.

3.4. PIV data

To supplement the quantitative analysis, the same methods are applied to real PIV data containing real outliers. As the locations of outlier vectors are not known, a formal analysis is not possible. As shown in figure 11, qualitatively speaking all of the methods perform well apart from the all-in-one method, which again removes/blurs the smaller spatial scale. Unfortunately, the PIV data contained no large clusters of outliers which may have highlighted PODDEM's ability. From figure 11 it is clear that the AWAMT method and Kriging is favourable. However, PODDEM detects all of the outlier points, especially those which could have a statistical impact, which POD-OC does not. If the data had contained large groups of outlier points, as demonstrated earlier, the results for AWAMT method and Kriging may not have been as favourable.

To further qualitatively demonstrate the detection and estimation capabilities of the PODDEM on a single field, it is applied to the real PIV data. In figure 12 vector field (500th)

of the real PIV data is selected and the PODDEM is compared with POD-OC, where $n_b \times m_b = 16 \times 16$. From figure 12 it is clear that the PODDEM out performs the POD-OC in both detection and estimation.

4. Discussions

The results in section 3 show that there are clear advantages to using spatio-temporal information for the detection and estimation of outliers. As demonstrated by the present study, a POD-based technique can be approached using either a modification of the spatial modes (POD-OC) or a modification of the temporal coefficients (PODDEM). Figures 3 and 4 in section 3 further demonstrates that there are clear benefits to modifying the temporal coefficients (PODDEM) for detection, especially in the case of large clusters of outliers. Figures 5–7, show the estimation ability of PODDEM may not always be the optimal choice for smaller clusters of outliers, a user could opt to use a hybrid of a Kriging based method for small scale estimations and PODDEM for large scale estimations. This may be especially beneficial in the case of single vector fields. Furthermore, if time is not a limiting factor as user may opt to use the adaptive Gappy-POD formulation (Raben *et al* 2012), however this method is extremely computationally expensive and impractical for large datasets.

4.1. POD-OC modifications

From the authors' investigations, it is found that the 'robust parameter', $a = 3$, which is proposed by Wang *et al* (2015) for the POD-OC algorithm is not optimal, and that changes to a can improve the performance of POD-OC. A sensitivity analysis of a is shown in figure 13. PODDEM also requires a user defined percentage, t_r, which was previously introduced as dependent on the outlier rate, Q, a sensitivity analysis of t_r is also shown on the same figure, but on different axes. For the sensitivity analysis, a subset of four test cases are selected, two from each dataset (channel flow and isotropic turbulence), using two outlier rates $Q = 5\%$ and $Q = 15\%$, with an $N_c = 5$. Figure 13 demonstrated that POD-OC has an optimal performance for $a \approx 4.5$. If this parameter is used, the correct rate of detection is increased, and rate of false detection minimised. The dependence of t_r with Q is also clear in the results. The optimum value of t_r in PODDEM is defined only by Q, which is a parameter that can be estimated based on a visual inspection of the PIV snapshots.

4.2. Further advancements to the PODDEM algorithm

The proposed PODDEM algorithm is based on the premise of 'smoothing' outliers within the temporal coefficients. This is ideal when every vector field contains an outlier; realistically however, not all vector fields will contain outliers. As shown in figure 14 when only 100 random frames contain outliers (i.e. 10% $Nc = 3$ & $Q = 5\%$), at the temporal locations relating to the vector fields containing outliers, spikes are perturbed in to the temporal coefficients. By imposing a spike detection algorithm, instead of a moving average, such as the 'Nikora–Goring method', typically used to remove spikes from acoustic doppler velocimetry data, Goring and Nikora (2002), the spikes can be removed without effecting other vector fields devoid of outliers. This is particularly beneficial where the temporal resolution of the dataset it low. This will be investigated in future work.

5. Conclusions

The current work proposes a novel, rapid and non-iterative POD method for the detection and estimation of outliers (PODDEM) based on modifications of the temporal coefficients. By introducing synthetic outliers to time series extracted from the John Hopkins Turbulence Database, and to real PIV data, the detection and estimation abilities of PODDEM are benchmarked against state-of-the-art spatial/spatio-temporal methods, including POD-OC. From the results it is observed that there are clear advantages from using the POD (spatio-temporal) methods for the detection and estimation of outliers. As the method is non-iterative substantial time benefits are observed by comparison with other POD based methods. A sensitivity analysis reveals that a modification of the temporal coefficients is beneficial in robustness for the detection of outliers compared with modifications of spatial modes, as in POD-OC. Furthermore, for cases which are not time resolved, PODDEM can be applied to a single vector field. Compared with state-of-the-art spatial estimation and detection methods, PODDEM is able to improve the detection of outliers for single frames without decreasing the estimation accuracy.

Acknowledgments

Johns Hopkins Turbulence Database (JHTDB) and Hain and Kähler for the data. The UK Natural Environment Research Council for the PhD studentship of the first author and Engineering and Physical Sciences Research Council for the second author. Furthermore, special thanks to Dr Hongping Wang (POD-OC) and Mr Alessandro Masullo for providing their algorithms (AWAMT), and to Mr Paul Raven for editorial courtesy.

Appendix

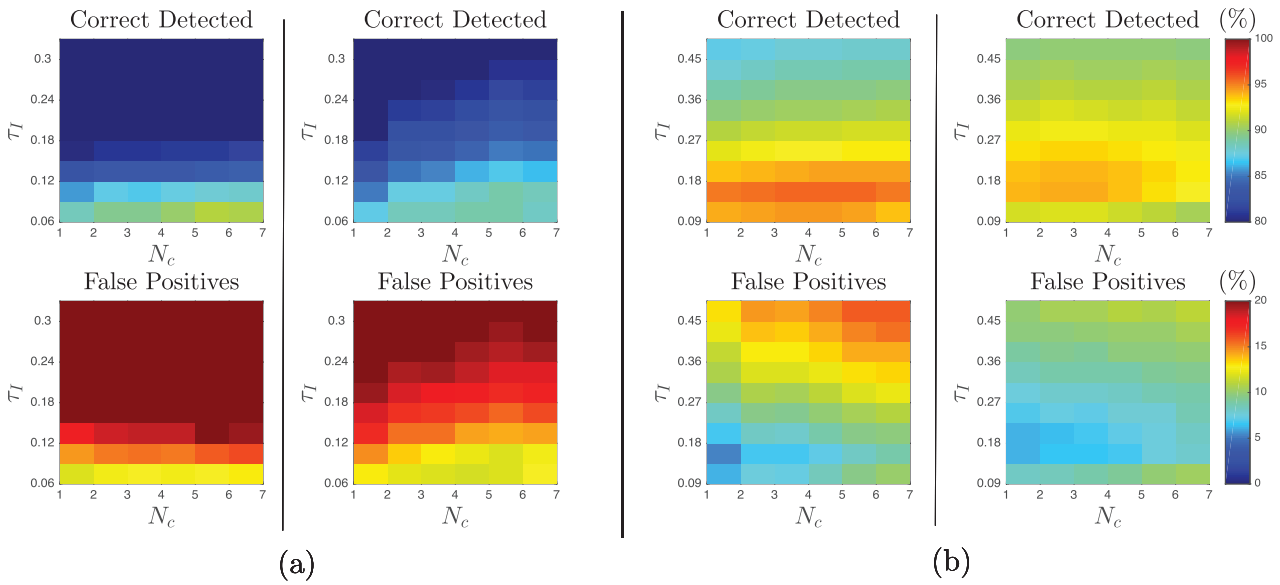

Figure A1. Sensitivity analysis of kernel sizes used in moving average step of the PODDEM, where τ_I is the kernel size based on the average integral time scale. The top row shows the percentage of correct outliers detected. The bottom row shows the percentage of false outliers detected. (a) Results for channel flow. (b) Results for isotropic case.

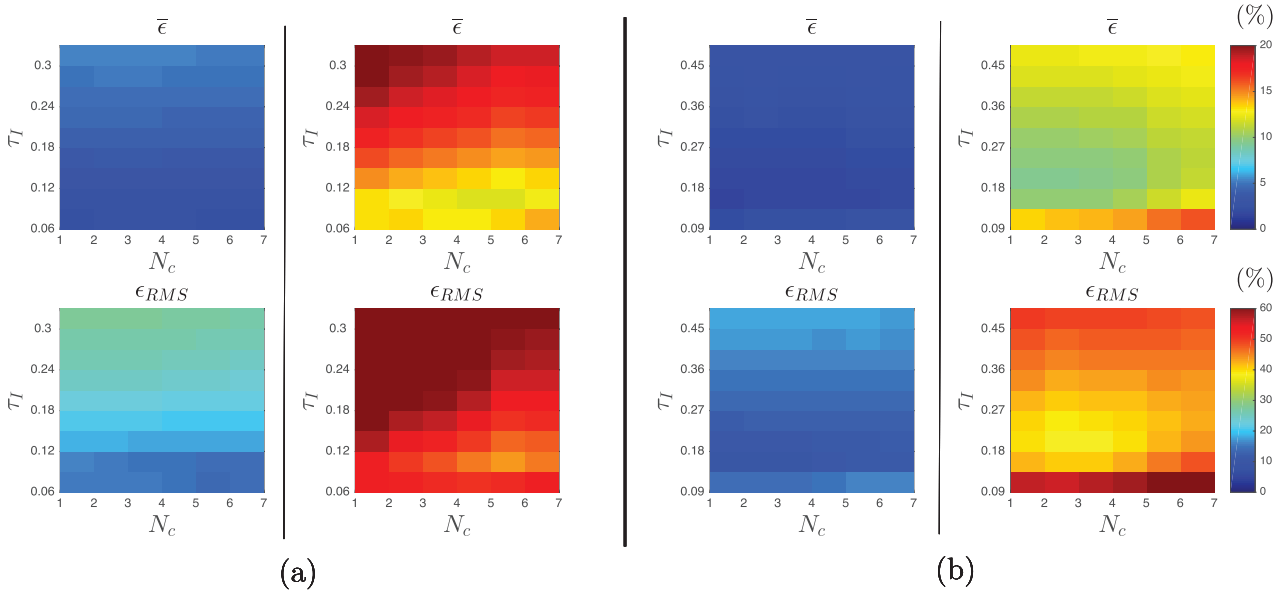

Figure A2. Sensitivity analysis of kernel sizes used in moving average step of the PODDEM, where τ_I is the kernel size based on the average integral time scale. The top row shows $\overline{\epsilon}$ (accuracy) and bottom rows show the spatio-temporal ϵ_{RMS} (precision) error. (a) shows the error obtained with the contaminated channel case. (b) shows the results for the isotropic case.

References

Adrian R J and Westerweel J 2011 *Particle Image Velocimetry* (Cambridge: Cambridge University Press)

Aubry N 1991 On the hidden beauty of the proper orthogonal decomposition *Theor. Comput. Fluid Dyn.* **2** 339–52

Berkooz G, Holmes P and Lumley J L 1993 The proper orthogonal decomposition in the analysis of turbulent flows *Annu. Rev. Fluid Mech.* **25** 539–75

Brevis W and García-Villalba M 2011 Shallow-flow visualization analysis by proper orthogonal decomposition *J. Hydraul. Res.* **49** 586–94

Everson R and Sirovich L 1995 Karhunen–Loève procedure for gappy data *J. Opt. Soc. Am.* **12** 1657–64

Garcia D 2010 Robust smoothing of gridded data in one and higher dimensions with missing values *Comput. Stat. Data Anal.* **54** 1167–78

Garcia D 2011 A fast all-in-one method for automated post-processing of PIV data *Exp. Fluids* **50** 1247–59

Goring D and Nikora I 2002 Despiking acoustic Doppler velocimeter data *J. Hydraul. Eng.* **128** 117–26
Graham J *et al* 2016 A web services accessible database of turbulent channel flow and its use for testing a new integral wall model for LES *J. Turbul.* **17** 181–215
Gunes H, Sirisup S and Karniadakis G E 2006 Gappy data: to krig or not to krig? *J. Comput. Phys.* **212** 358–82
Hain R and Kähler C 2007 Fundamentals of multi-frame particle image velocimetry (PIV) *Exp. Fluids* **42** 575–87
Karhunen K 1946 Zur spektral theorie stochastischer prozesse *Ann. Acad. Sci. Fenn. A1: Math.—Phys.* **34**
Kosambi D 1943 Statistics in function space *J. Indian Math. Soc.* **7** 76–88
Li Y, Perlman E, Wan M, Yang Y, Meneveau C, Burns R, Chen S, Szalay A and Eyink G 2008 A public turbulence database cluster and applications to study Lagrangian evolution of velocity increments in turbulence *J. Turbul.* **9** N31
Liang D, Jiang C and Li Y 2003 Cellular neural network to detect spurious vectors in PIV data *Exp. Fluids* **34** 52–62
Loève M 1945 Functions aléatoire de second ordre *C. R. Acad. Bulg. Sci.* **220**
Masullo A and Theunissen R 2016 Adaptive vector validation in image velocimetry to minimise the influence of outlier clusters *Exp. Fluids* **57** 1–21
Nielsen H, Lophaven S and Søndergaard J 2002 DACE—a MATLAB Kriging toolbox *Technical report* Informatics and Mathematical Modelling, Technical University of Denmark, DTU
Obukbov A M 1954 Statistical description of continuous fields *Tr. Gos. Inst., Akad. Nauk SSSR* **24** 3–42
Pugachev V S 1953 The general theory of correlation of random functions *Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya* **17** 401–20
Raben S G, Charonko J J and Vlachos P P 2012 Adaptive gappy proper orthogonal decomposition for particle image velocimetry data reconstruction *Meas. Sci. Technol.* **23** 025303
Shinneeb A, Bugg J and Balachandar R 2004 Variable threshold outlier identification in PIV data *Meas. Sci. Technol.* **15** 1722
Thielicke W and Stamhuis E 2014 PIVLab–towards user-friendly, affordable and accurate digital particle image velocimetry in MATLAB *J. Open Res. Software* **2**
Wang H, Gao Q, Feng L, Wei R and Wang J 2015 Proper orthogonal decomposition based outlier correction for PIV data *Exp. Fluids* **56** 1–15
Westerweel J 1994 Efficient detection of spurious vectors in particle image velocimetry data *Exp. Fluids* **16** 236–47
Westerweel J and Scarano F 2005 Universal outlier detection for PIV data *Exp. Fluids* **39** 1096–100