scientific reports

Check for updates

OPEN

Analysis of interaction dynamics and rogue wave localization in modulation instability using data-driven dominant balance

Andrei V. Ermolaev[1], Mehdi Mabed[1], Christophe Finot[2], Goëry Genty[3] & John M. Dudley[1]✉

We analyze the dynamics of modulation instability in optical fiber (or any other nonlinear Schrödinger equation system) using the machine-learning technique of data-driven dominant balance. We aim to automate the identification of which particular physical processes drive propagation in different regimes, a task usually performed using intuition and comparison with asymptotic limits. We first apply the method to interpret known analytic results describing Akhmediev breather, Kuznetsov-Ma, and Peregrine soliton (rogue wave) structures, and show how we can automatically distinguish regions of dominant nonlinear propagation from regions where nonlinearity and dispersion combine to drive the observed spatio-temporal localization. Using numerical simulations, we then apply the technique to the more complex case of noise-driven spontaneous modulation instability, and show that we can readily isolate different regimes of dominant physical interactions, even within the dynamics of chaotic propagation.

Modulation instability (MI) of the nonlinear Schrödinger equation (NLSE) describes the process whereby a weak perturbation experiences exponential growth at the expense of a strong input wave[1,2]. MI (sometimes called the Benjamin-Feir or Bespalov-Talanov instability[3,4]) leads to complex spatio-temporal pattern formation, and is one of the fundamental nonlinear dynamical processes of nature. It has been observed in many different systems including hydrodynamics, plasmas, Bose-Einstein condensates, and fiber optics. Despite this large body of work over many years, its centrality to nonlinear science is such that it continues to be extensively studied from both experimental and theoretical perspectives. Recent work, for example, has explored its description in terms of integrable turbulence[5,6], its relationship with computational complexity[7], its thermodynamic link to the soliton-gas concept[8], and its intrinsic association with Fermi-Pasta-Ulam-Tsingou recurrence[9]; to cite only a small number of examples.

The dynamics of MI leads to the spontaneous emergence of localized structures that possess different spatial and/or temporal periodicities[10,11]. These structures are intimately connected with known analytic solutions to the NLSE (including Peregrine soliton and Akhmediev and Kuznetsov-Ma breathers[12–14]), and understanding this correspondence has allowed experimentalists to excite a wide range of soliton and breather solutions in both optics and hydrodynamics[15–18]. Moreover, even under conditions where modulation instability is excited from noise, it has been shown that the random peaks developing from noise possess the expected characteristics of these analytic solutions[19–21]. It is these nonlinear localization dynamics in particular that have attracted great interest as potentially underpinning the growth and decay of destructive rogue waves on the ocean[22,23].

These various studies have yielded significant insights into the properties of MI under diverse conditions, and in different systems. Somewhat surprisingly, however, although some aspects of MI localization can be interpreted precisely using mathematical methods such as the inverse scattering transform[24], the physics of nonlinear and dispersive interactions in MI is more often discussed in qualitative terms by a comparison with specific limiting cases or characteristic nonlinear and dispersive length scales[25]. It would be highly desirable to have a means of interpreting the physics of MI that went beyond such a qualitative description, and yet avoided the formalism of the inverse scattering method.

In this paper, we show that the machine-learning technique of data-driven dominant balance can address this problem. Machine learning methods are currently of great interest in all areas of physics[26–28], and in the

[1]Université de Franche-Comté, Institut FEMTO-ST, CNRS UMR 6174, 25000 Besançon, France. [2]Laboratoire Interdisciplinaire Carnot de Bourgogne, CNRS UMR 6303, Université de Bourgogne, 21078 Dijon, France. [3]Photonics Laboratory, Tampere University, 33104 Tampere, Finland. ✉email: john.dudley@univ-fcomte.fr

particular field of nonlinear optics, have been applied to the study of various NLSE propagation scenarios[29–32]. The technique of dominant balance aims to automatically determine the contributing dominant physical processes at each step of propagation. As a subset of unsupervised learning techniques, it has been successfully applied to interpret the physics of a number of nonlinear propagation scenarios in hydrodynamics, as well as the more challenging case of broadband supercontinuum generation[33].

In this paper, we use a dominant balance approach to analyse modulation instability of the NLSE. We first apply the method to interpret known analytic solutions for Akhmediev breather, Kuznetsov-Ma, and Peregrine soliton structures, and for these spatio-temporal dynamics, we show how we can distinguish background regions of dominant nonlinear propagation from regions where nonlinearity and dispersion interact to drive localization. This is especially important in showing how dominant balance can provide complementary insights into the dynamics, because associating the nonlinear stage of evolution with the background may seem counter-intuitive as this is a region of low intensity. Following these studies of analytic SFB solutions we then use numerical simulations to study the more complex propagation case of noise-driven chaotic MI, and find again that we can automatically identify these different regimes of physical interaction.

NLSE solutions

We consider MI occurring in the focusing NLSE which is written in normalised form as follows:

$$i\frac{\partial \psi}{\partial \xi} + \frac{\partial^2 \psi}{\partial \tau^2} + |\psi|^2\psi = 0. \tag{1}$$

Here $\psi(\xi,\tau)$ is a field envelope evolving in distance ξ and co-moving time τ. Dimensionless variables ξ and τ are related to the usual notation of nonlinear optics by $\xi = z/L_{\mathrm{NL}}$ and $\tau = t/\sqrt{L_{\mathrm{NL}}|\beta_2|/2}$, where $L_{\mathrm{NL}} = (\gamma P_0)^{-1}$. Here z and t are dimensional distance and time, P_0 is power (usually that of the input continuous wave), and β_2 and γ are the usual dimensional fiber group velocity dispersion and nonlinearity parameters respectively[25]. The field envelope $\psi(\xi,\tau)$ is normalized with respect to $P_0^{1/2}$.

The NLSE possesses a number of known analytic solutions[11,34]. Those associated with MI are the solitons on finite background (SFB), that can be written in compact form as follows:

$$\psi(\xi,\tau) = \left[1 + \frac{2(1-2a)\cosh(b\xi) + ib\sinh(b\xi)}{\sqrt{2a}\cos(\omega_m\tau) - \cosh(b\xi)}\right]\exp(i\xi), \tag{2}$$

The physical behaviour of the solution is determined by the single governing parameter a through arguments $b = [8a(1-2a)]^{1/2}$ and $\omega_m = [2(1-2a)]^{1/2}$. When $a = 1/2$, $\omega_m = b = 0$ and the solution is the limiting rational Peregrine soliton, double-localized in ξ and τ[14]. For $a < 1/2$, ω_m and b are real, and we obtain the τ -periodic Akhmediev breather, with ω_m and b taking on physical significance of a modulation frequency and exponential growth/decay rate respectively. When $a > 1/2$, ω_m and b become imaginary, and we obtain the ξ -periodic Kuznetsov-Ma solution. These various SFB structures are well known, and have been observed in a range of experiments since 2010[35–37].

Implementing the dominant balance technique

In this section, we give a general overview of how the dominant balance technique and algorithm are applied to nonlinear propagation in the NLSE. Further details and references are given in the "Methods" section. The dominant balance technique aims to automate the process of identifying the key interacting physical processes associated with different spatio-temporal regions of evolution. The technique involves several steps. The first is to determine the evolution of the field $\psi(\xi,\tau)$, and this is straightforward here as we have access to the analytic result in Eq. (2). However, as we see below for noise-driven MI, the evolution can also be obtained using numerical integration of the NLSE. Indeed, in the most general case, this could also involve analysis of experimental data when access to full field information is available[38,39].

The second step analyses the evolution $\psi(\xi,\tau)$ in its associated "equation space," where each coordinate axis corresponds to a physical process defined by one of the terms in the governing NLSE (see "Methods"). Specifically, for each point (ξ,τ), the NLSE terms $\{i\psi_\xi, \psi_{\tau\tau}, \psi|\psi|^2\}$ are separately computed, and we search for a "dominant balance" regime where the NLSE is approximately satisfied by only a subset of terms (the other terms contributing only negligibly.) As shown in Ref.[33], machine learning tools can automate this search, using cluster detection (Gaussian mixture modelling) and sparse regularization to identify regions where different combinations of terms drive the dynamics. These are standard tools of unsupervised learning and optimization, and allow robust detection of clusters even when they overlap (see "Methods")[28,40]. When different clusters are found to possess the same sparcity pattern (significantly reduced variance in the same directions of equation space), these are grouped together to form a particular candidate "balance model." In the case of the NLSE with three possible interacting terms, this process has a simple geometric interpretation: two interacting terms will be associated with a cluster falling on a line in the three-dimensional equation space, three interacting terms will be associated with a cluster in a plane.

When the data is fully grouped into balance models, the final step is to re-map the clusters back onto the (ξ,τ) space for comparison with the standard evolution dynamics. Visually, we do this by segmenting the original domain using a color key describing each balance model. In our analysis, we used the code package described in Ref.[33], and available at the online repository[41]. We also note that since we are dealing with complex fields, we stacked real and imaginary components as input to allow grouping of regions of significant variance irrespective whether identified in the real or imaginary components[41].

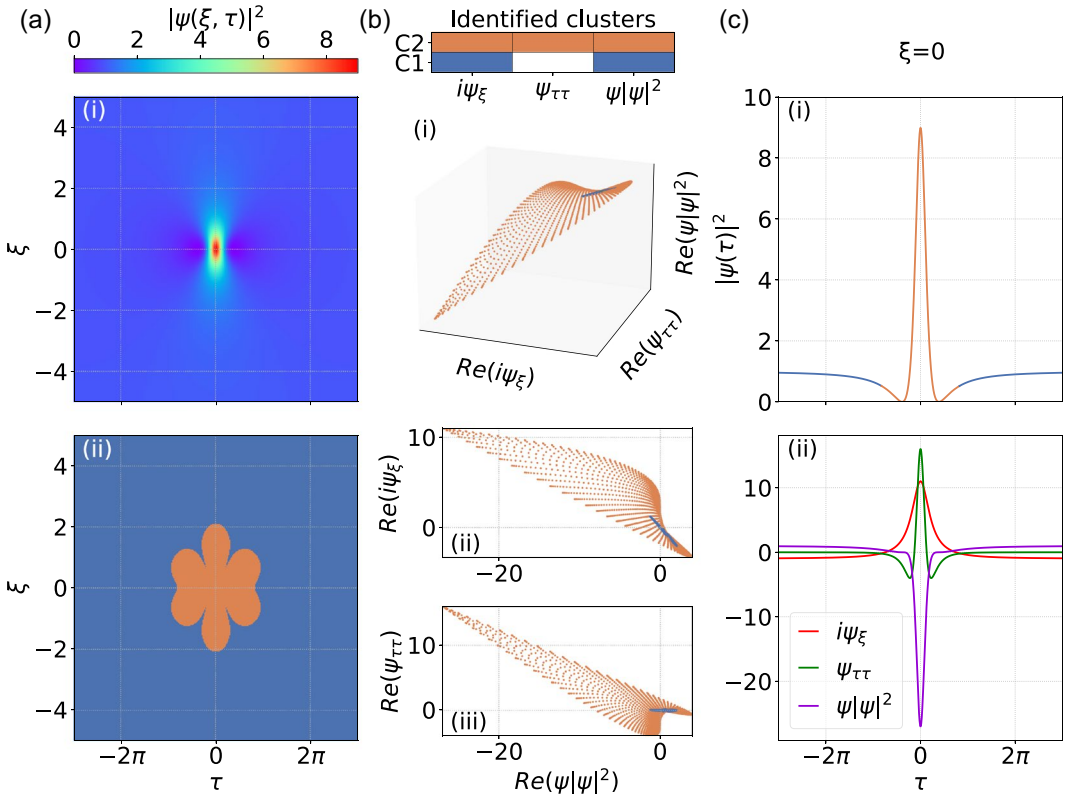

Figure 1. Dominant balance method applied to the Peregrine soliton. (**a**-i) Spatio-temporal evolution of $|\psi(\xi,\tau)|^2$. (**a**-ii) Segmented map of the evolution space where the color key describes: only nonlinear and propagation terms (blue), and all NLSE terms (orange). Using the same color key, (**b**) shows cluster identification for: (i) real parts of $\{i\psi_\xi, \psi_{\tau\tau}, \psi|\psi|^2\}$; (ii) real parts of $\{\psi|\psi|^2, i\psi_\xi\}$; (iii) real parts of $\{\psi|\psi|^2, \psi_{\tau\tau}\}$. (**c**) Using the same color key, (i) shows the intensity profile at $\xi = 0$; (ii) Individual contributing terms in the NLSE at $\xi = 0$ as indicated in the legend.

Results

We first apply this technique to identify locally-dominant interactions during the evolution of the three classes of SFB described above. Figure 1 shows results for the Peregrine soliton. Specifically, Fig. 1a-i shows the spatio-temporal evolution $|\psi(\xi,\tau)|^2$ which reveals the expected double-localization. The results of the dominant balance procedure are shown in Fig. 1b. Here Fig. 1b-i plots the identified clusters in the three-dimensional space of the real parts of coordinates $\{i\psi_\xi, \psi_{\tau\tau}, \psi|\psi|^2\}$, whereas Fig. 1b-ii and b-iii show two projections as indicated. The color key corresponds to two different dominant balance models that are found: one where only the nonlinear and propagation terms contribute (blue) and another where all NLSE terms contribute (orange). No cluster is found that involves only the dispersive and propagation terms. Note that for convenience we plot dependencies only for the real field components, but similar results are found for the imaginary components. The results in Fig. 1b show that all the points assigned to the blue cluster (nonlinear and propagation terms) are strongly localised in the equation space forming a dense distribution that manifests nearly zero variation with respect to the $\psi_{\tau\tau}$ axis (see particularly Fig. 1b-iii). In contrast, the orange cluster (all terms) is distributed throughout the equation space with no reduced variance with respect to any of three axes. This illustrates the geometrical interpretation lying behind the dominant balance approach.

The color-coded clusters are then mapped back onto a segmented dominant balance plot shown in Fig. 1a-ii, and the particular intensity profile at $\xi = 0$ is also plotted in Fig. 1c-i using the same color key. At $\xi = 0$, it is also instructive to plot the different contributions of each term of the equation space as shown in Fig. 1c-ii, clearly revealing how different combinations of terms contribute to satisfy the NLSE (i.e. add to zero) in different regions. Note that at $\xi = 0$ all the three terms $\{i\psi_\xi, \psi_{\tau\tau}, \psi|\psi|^2\}$ corresponding to the SFB solution are purely real.

These results reveal the key physical features of NLSE dynamics. For example, considering the Peregrine soliton and comparing Fig. 1a-i and a-ii, the orange region reveals how the strong spatio-temporal localization around $(\xi = 0, \tau = 0)$ arises from the interaction between all terms in the NLSE, as both nonlinearity and dispersion combine to drive spatio-temporal compression. In contrast, the surrounding background region (blue) is dominated only by nonlinear evolution, and whilst this might be considered counter-intuitive since the background is where the intensity is lowest, this result actually highlights how interpreting NLSE physics requires comparison of the relative contributions of dispersion and nonlinearity. Specifically, a plane wave with no τ-structure can not "experience" dispersion, and thus it is only nonlinear self-focussing that can initially

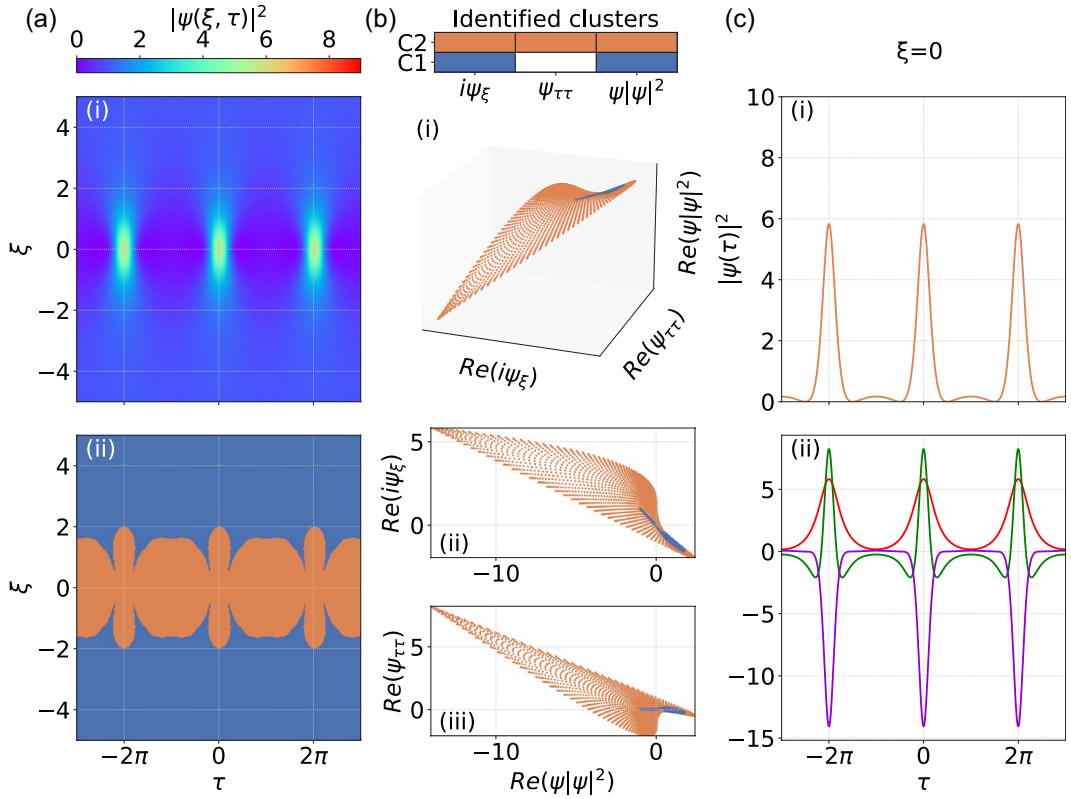

Figure 2. Dominant balance method applied to the Akhmediev breather. (**a**-i) Spatio-temporal evolution of $|\psi(\xi,\tau)|^2$. (**a**-ii) Segmented map of the evolution space where the color key describes: only nonlinear and propagation terms (blue), and all NLSE terms (orange). Using the same color key, (**b**) shows cluster identification for: (i) real parts of $\{i\psi_\xi, \psi_{\tau\tau}, \psi|\psi|^2\}$; (ii) real parts of $\{\psi|\psi|^2, i\psi_\xi\}$; (iii) real parts of $\{\psi|\psi|^2, \psi_{\tau\tau}\}$. (**c**) Using the same color key, (i) shows the intensity profile at $\xi = 0$; (ii) Individual contributing terms in the NLSE at $\xi = 0$ as indicated in the legend of Fig. 1c-ii.

influence the evolution of the background. It is only after temporal structure develops from this nonlinear stage of evolution that dispersion and nonlinearity interact. In fact, this approach to visualizing the evolution very clearly illustrates the well-known "nonlinear" stage of the instability[34,42]. The ability of the dominant balance analysis to identify this nonlinear stage explicitly (even though perhaps counter-intuitive from a naive perspective) is an example of how it can yield important insights into nonlinear evolution.

The results in Figs. 2 and 3 for the Akhmediev and Kuznetsov-Ma breathers respectively have similar interpretation. Here we see again how regions of background associated only with dominant nonlinearity (blue) have been clearly identified, but we also clearly see how the contributions of all terms (orange) leads to the expected spatio-temporal localization characteristics. We also note how for the particular case of the Akhmediev breather, the $\xi = 0$ profile plot in Fig. 2c shows how all terms contribute to the dynamics in the lower amplitude regions between the localized peaks. These analytical SFB solutions, of course, do not exhaust the full variety of localised structures appearing in MI such as higher-order solutions[43], breather or soliton collisions[44], ghost interactions[45] etc. However, these key examples provide a clear indication of how the dominant balance approach can complement existing techniques such as inverse scattering transform[24,34,46] in interpreting NLSE dynamics.

We now apply the dominant balance approach to interpret the more complex dynamics of noise-driven MI. For this case, the NLSE is solved numerically for a plane wave input with an imposed low level broadband noise background. We used a common optical noise model corresponding to a one photon per mode background[47], but in fact similar chaotic dynamics in MI can be seen with essentially any class of random amplitude and/or phase fluctuation on the input[25]. The spatio-temporal intensity dynamics of $|\psi(\xi,\tau)|^2$ for this case are shown in Fig. 4a and for completeness, we also show in Fig. 4b the associated spectral evolution[19].

We clearly see how the input plane wave evolves into a series of localized peaks, displaying both random temporal (transverse) and spatial (longitudinal) periodicity. Maximum gain for the spontaneous instability is at sideband frequency $\Omega = 1$ and is associated with the initial emergence of Akhmediev breathers with temporal periodicity of $\Delta\tau = 2\pi$. After this initial stage, subsequent evolution is plotted up to $\xi = 120$. We also see how the incoherent temporal evolution is reflected in the frequency domain with chaotic spectral expansion and contraction, as the random emergence of particularly high intensity temporal peaks of ultrashort duration is associated with broader spectra.

Analyzing the evolution in terms of dominant balance yields the results shown in Fig. 4c. The color scale is the same as the previous figures. Comparing these results with the analytic SFB structures above allows us to

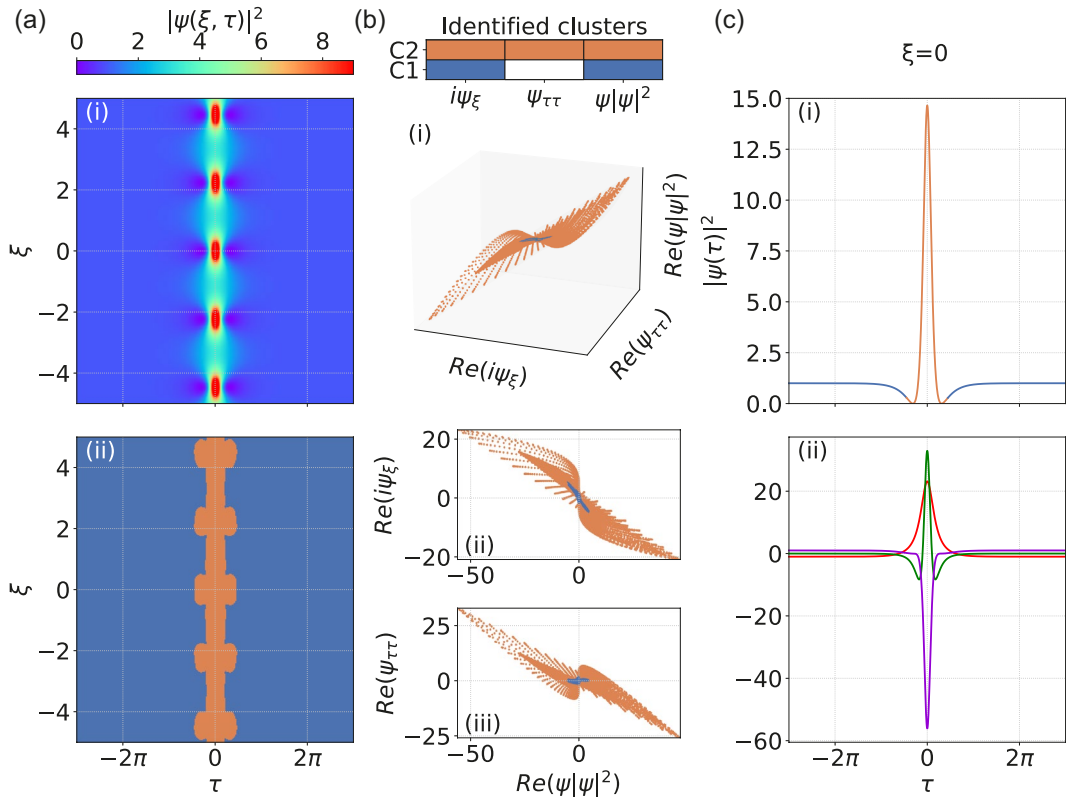

Figure 3. Dominant balance method applied to the Kuznetsov-Ma breather. (**a**-i) Spatio-temporal evolution of $|\psi(\xi,\tau)|^2$. (**a**-ii) Segmented map of the evolution space where the color key describes: only nonlinear and propagation terms (blue), and all NLSE terms (orange). Using the same color key, (**b**) shows cluster identification for: (i) real parts of $\{i\psi_\xi, \psi_{\tau\tau}, \psi|\psi|^2\}$; (ii) real parts of $\{\psi|\psi|^2, i\psi_\xi\}$; (iii) real parts of $\{\psi|\psi|^2, \psi_{\tau\tau}\}$. (**c**) Using the same color key, (i) shows the intensity profile at $\xi = 0$. (ii) Individual contributing terms in the NLSE at $\xi = 0$, as indicated in the legend of Fig. 1c-ii.

distinguish the emerging localised structures. Indeed, even in this case of highly random MI dynamics data-driven dominant balance successfully finds the Akhmediev breathers with period $\Delta\tau \approx 2\pi$ (for example, at $\xi \approx 11$ and $\xi \approx 38$), and we also see how propagation is associated with various ξ-periodic structures, breather collisions and Peregrine soliton-like rogue wave structures (e.g. the isolated feature in Fig. 4c at $\xi \approx 93$). Being based on unsupervised clustering of contributing terms to the evolution equation rather then simple intensity thresholding, the technique successfully identifies developing localised structures even in low-intensity regions. This suggests the further application of the method in automated identification of emerging rogue wave structures[48].

Discussion and conclusions

In conclusion, these results have clearly shown how the dominant balance approach provides a powerful tool for studying the interactions between dispersion and nonlinearity in the context of breather and modulation instability dynamics. In particular, even though these processes have been the subject of much previous study, visualising the dynamics with dominant balance segmentation clearly provides valuable insights into the relative contributions of different physical processes at different points in the evolution map.

We stress here, however, that data driven methods are not designed to replace existing techniques of analysing nonlinear dynamics, but should be seen as complementary tools to assist the use of physical considerations. For example, of particular interest is the way in which the dominant balance technique correctly associates the evolution of the plane wave background with a nonlinear stage of propagation. This illustrates how simplistic interpretations such as associating nonlinear evolution with intensity thresholding could be misleading, and it is always necessary to consider the relative contributions of nonlinearity and dispersion in discussing the dynamics of the NLSE.

Moreover, whilst with experience, inspection of spectral and temporal evolution maps of the NLSE can allow some processes (such as collisions) to be readily associated with combined nonlinear and dispersive interactions, such interpretations can sometimes be misleading. This is particularly the case with generalized forms of the NLSE where multiple processes combine, as previously studied in Ref.[33] for the case of optical fibre supercontinuum generation. The strength of the dominant balance approach is that it provides additional information in an unsupervised manner (i.e. not based on intuition or experience). When applied in parallel with other analysis

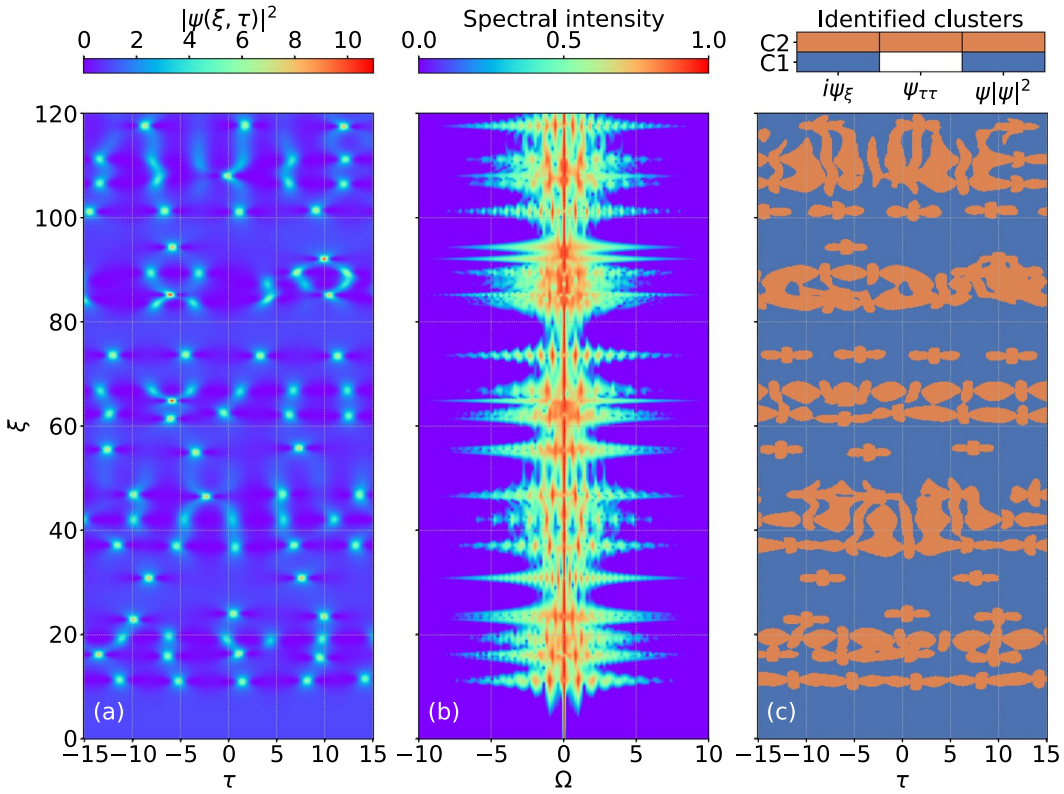

Figure 4. (**a**) Spatio-temporal evolution of the normalized intensity for spontaneous modulation instability. (**b**) The associated spectral evolution. (**c**) The results of dominant balance revealing the different interaction regions according to the colormap shown (same as in previous figures).

techniques, this provides important complementary information to yield the best possible physical interpretation of complex evolution.

Finally, we note that the NLSE describes propagation in many systems other than optical fiber, and there has been a strong recent focus on studying novel NLSE dynamics in deep-water hydrodynamics[22]. In this context, we anticipate an important area of future application will be the case of MI induced by localized perturbations[49], and the associated emergence of rogue wave statistics[50,51]. There is clearly much potential for data-driven discovery methods to be applied in NLSE-related systems[25].

Methods

Equation space representation. The methodology of identifying a dominant balance model for a physical system at a particular stage of propagation aims to find a subset of terms of a more broadly applicable propagation model that locally dominates the dynamics. Following the approach and notation of Ref.[33], we consider a general evolution equation on a domain (ξ, τ) written as follows:

$$\sum_{i=1}^{K} f_i\left(\psi, \psi_\xi, \psi_\tau \ldots, \psi^2, \psi\psi_\xi, \psi\psi_\tau, \ldots, \psi_{\xi\xi}, \psi_{\tau\tau}, \ldots\right) = 0, \tag{3}$$

where K is the number of terms, and the terms f_i can be constructed in various ways from the spatio-temporal field $\psi(\xi, \tau)$. As discussed in Ref.[33] (and its accompanying Supplementary Information), the advantage of this implicit form of the propagation equation is that it stresses the balance that must be present to satisfy the equality: the sum of all the terms must be zero. "Dominant balance" describes the situation when only a subset p of the K terms dominate the equality such that the contributions from the other $K - p$ terms are small or negligible. Geometrically, the equation space is described by a vector: $\mathbf{f}(\xi, \tau) = \left[f_1[\psi(\xi_n, \tau_m), \ldots], \ldots, f_K[\psi(\xi_n, \tau_m), \ldots]\right]^T$ where each of the dimensions (directions) corresponds to a specific term in the evolution equation (here indices $n \in [1, N]$ and $m \in [1, M]$ represent the discretization of $\psi(\xi, \tau)$, where N and M are the number of points in the ξ and τ directions respectively). A dominant balance regime then has a direct geometrical interpretation - dynamical points attributed to a certain dominant balance regime will be restricted to p directions of the full K-dimensional space. In other words, when plotting the different terms in the equation space, the points associated with the dominant p terms will have significantly reduced variance with respect to other $K - p$ directions.

In the case of the NLSE, the dimensionality $K = 3$ and each dynamical point $\psi(\xi_n, \tau_m)$ is associated with a vector $\left[i\psi_\xi(\xi_n, \tau_m), \psi_{\tau\tau}(\xi_n, \tau_m), |\psi(\xi_n, \tau_m)|^2\psi(\xi_n, \tau_m)\right]^T$. In geometrical terms, dominant balance between the propagation and nonlinear Kerr terms $\{i\psi_\xi, |\psi|^2\psi\}$ will be represented by an ensemble of points restricted on

a line with near-zero variance with respect to the dispersion term $\psi_{\tau\tau}$ (e.g. the blue clusters in Figs. 1, 2 and 3). In contrast, the ensemble of points distributed throughout the $i\psi_\xi + \psi_{\tau\tau} + |\psi|^2\psi = 0$ plane will represent the full dynamics that involves the interplay of all three dynamical terms (e.g. the orange clusters in Figs. 1, 2 and 3).

Finding dominant balance models through clustering. The search for dominant combinations of terms within a higher-dimensional equation space is an ideal problem for unsupervised clustering algorithms[28,40]. In particular, we use the algorithm and code package described in Refs.[33] and[41] respectively which are based on a probabilistic Gaussian Mixture Model (GMM) framework. GMM seeks to locate clustered subpopulations within an overall population of data, under the assumption that the data consists of a mixture of Gaussian distributions with specified weights, means and covariance matrices (usually denoted $\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ respectively, where k is the cluster index). The covariance matrix here generalizes the usual variance of a one-dimensional Gaussian distribution to higher dimensions. In contrast to simpler techniques such as k-means associated with hard partitions between clusters, GMM describes membership of a clusters in a probabilistic sense, allowing the algorithm to fit and return clusters that overlap. The GMM algorithm is based on the expectation-maximisation technique, a standard approach that is fully described in e.g. Ref.[40]. The particular GMM algorithm used here is `GaussianMixture` from the `scikit-learn` Python package[52], as implemented in Ref.[41].

A key motivation to use the GMM is that the covariance matrices can be interpreted physically to identify combinations of terms that dominate the dynamics. In particular, clusters associated with directions (dimensions) with significant variance correspond to physical terms that contribute actively to the dynamics (see the discussion of the results in Figs. 1, 2 and 3 above). However, there are some important additional factors that need to be considered to apply this approach successfully. In particular, since the data points in the equation space may not actually approximate a mixture of Gaussian distributions, the algorithm will usually return a number of clusters greater that the number of physical balance regimes. As described in detail in Ref.[33], this problem can be overcome using Sparse Principal Component Analysis (Sparse PCA) which uses l_1-regularisation to determine a sparse approximation to the leading principal component of each cluster[53,54]. In this case, when a particular cluster is associated with a dominant balance regime, it should be well described by the particular direction of its maximum variance. Note that l_1-regularisation in this context is a standard approach in machine learning using the l_1 norm as the penalty in the PCA regression-optimization problem[53].

There are two key parameters that need to be selected to ensure that the returned models correspond as accurately as possible to physical regimes of dominant balance. The first is the particular number of clusters used in the Gaussian Mixture Model. Although in principle we can already anticipate the maximum number of potential clusters based on the number of terms in the propagation model, it is usually advantageous to initially choose a greater number, as the l_1-regularisation step will later group together clusters found to possess the same sparcity patterns (i.e. reduced variances in the same directions of equation space)[33]. The second parameter is associated with the sparse regularisation of the PCA that describes the tradeoff between accuracy and sparsity in the returned models. A procedure for this selection process is described in detail in the Supplementary information of Ref.[33], and is based on considering a returned Pareto-type curve that plots the residual error of the inactive terms (accuracy) against the regularization parameter (sparsity). It is generally straightforward to see from this plot the most suitable parameter to generate the returned balance model. The very last step of the algorithm involves re-mapping the sparse clusters back onto the original spatio-temporal domain, and it is at this point we can directly compare the initial field distribution with the identified cluster map (as in Figs. 1, 2, 3 and 4).

It is useful to give further numerical details for our results. For the three classes of soliton on finite background considered in Figs. 1, 2 and 3, the evolution maps $\psi(\xi, \tau)$ were computed over $(N \times M) = (501 \times 1024)$ in ξ and τ respectively. For the noise-driven map considered in Fig. 4, evolution was computed over $(N \times M) = (5001 \times 1024)$ in ξ and τ respectively. The GMM search was based on an initial selection of up to 5 clusters and the sparse regularisation parameter α (used in the Python function `SparsePCA`[41]) was in the range 50–100. We also note the computation time associated with the GMM clustering and SPCA analysis, which was typically 6 and 21 min respectively for solitons on finite background and noise-driven MI, running on a standard Windows PC with 3.00 GHz 6 MB cache double-core CPU.

Data availibility

The data underlying the results presented in this paper are available from the corresponding author J.M.D. upon reasonable request.

Received: 16 March 2023; Accepted: 14 June 2023
Published online: 28 June 2023

References

1. Zakharov, V. E. Stability of periodic waves of finite amplitude on the surface of a deep fluid. *J. Appl. Mech. Tech. Phys.* **9**, 190–194 (1972).
2. Zakharov, V. E. & Ostrovsky, L. A. Modulation instability: The beginning. *Phys. D: Nonlinear Phenom.* **238**, 540–548 (2009).
3. Benjamin, T. B. & Feir, J. E. The disintegration of wave trains on deep water. Part I. Theory. *J. Fluid Mech.* **27**, 417–430 (1967).
4. Bespalov, V. I. & Talanov, V. I. Filamentary structure of light beams in nonlinear liquids. *JETP Lett.* **3**, 307–310 (1966).
5. Randoux, S., Walczak, P., Onorato, M. & Suret, P. Nonlinear random optical waves: Integrable turbulence, rogue waves and intermittency. *Phys. D: Nonlinear Phenom.* **333**, 323–335 (2016).
6. Walczak, P., Randoux, S. & Suret, P. Optical rogue waves in integrable turbulence. *Phys. Rev. Lett.* **114**, 143903 (2015).
7. Perego, A. M., Bessin, F. & Mussot, A. Complexity of modulation instability. *Phys. Rev. Res.* **4**, 1022057 (2022).
8. Gelash, A. *et al.* Bound state soliton gas dynamics underlying the spontaneous modulational instability. *Phys. Rev. Lett.* **123**, 234102 (2019).

9. Mussot, A. *et al.* Fibre multi-wave mixing combs reveal the broken symmetry of fermi–pasta–ulam recurrence. *Nat. Photon.* **12**, 303–308 (2018).
10. Akhmediev, N. N. & Korneev, V. I. Modulation instability and periodic solutions of the nonlinear Schrödinger equation. *Theor. Math. Phys.* **69**, 1089–1093 (1986).
11. Akhmediev, N. & Ankiewicz, A. *Solitons: Nonlinear Pulses and Beams* (Chapman and Hall, 1997).
12. Kuznetsov, E. Solitons in a parametrically unstable plasma. *Soviet Phys. Doklady* **22**, 507–508 (1977).
13. Ma, Y. C. The perturbed plane-wave solutions of the cubic Schrödinger equation. *Stud. Appl. Math.* **60**, 43–58 (1979).
14. Peregrine, D. H. Water waves, nonlinear Schrödinger equations and their solutions. *J. Austr. Math. Soc. Ser. B Appl. Math.* **25**, 16–43 (1983).
15. Dudley, J. M., Dias, F., Erkintalo, M. & Genty, G. Instabilities, breathers and rogue waves in optics. *Nat. Photon.* **8**, 755–764 (2014).
16. Kibler, B., Chabchoub, A. & Bailung, H. Editorial: Peregrine soliton and breathers in wave physics: Achievements and perspectives. *Front. Phys.* **9**, 795983 (2021).
17. Chabchoub, A. & Akhmediev, N. Observation of rogue wave triplets in water waves. *Phys. Lett. A* **377**, 2590–2593 (2013).
18. Chabchoub, A., Hoffmann, N. P. & Akhmediev, N. Rogue wave observation in a water wave tank. *Phys. Rev. Lett.* **106**, 204502 (2011).
19. Dudley, J. M., Genty, G., Dias, F., Kibler, B. & Akhmediev, N. Modulation instability, akhmediev breathers and continuous wave supercontinuum generation. *Opt. Express* **17**, 21497–21508 (2009).
20. Toenger, S. *et al.* Emergent rogue wave structures and statistics in spontaneous modulation instability. *Sci. Rep.* **5**, 10380 (2015).
21. Närhi, M. *et al.* Machine learning analysis of extreme events in optical fibre modulation instability. *Nat. Commun.* **9**, 4923 (2018).
22. Dudley, J. M., Genty, G., Mussot, A., Chabchoub, A. & Dias, F. Rogue waves and analogies in optics and oceanography. *Nat. Rev. Phys.* **1**, 675–689 (2019).
23. Chen, S. *et al.* Modulation instability—rogue wave correspondence hidden in integrable systems. *Commun. Phys.* **5**, 297 (2022).
24. Randoux, S., Suret, P. & El, G. Inverse scattering transform analysis of rogue waves using local periodization procedure. *Sci. Rep.* **6**, 29238 (2016).
25. Agrawal, G. P. *Nonlinear Fiber Optics* (Elsevier Science & Techn, 2019).
26. Jordan, M. I. & Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. *Science* **349**, 255–260 (2015).
27. Brunton, S. L., Proctor, J. L. & Kutz, J. N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl. Acad. Sci.* **113**, 3932–3937 (2016).
28. Brunton, S. L. & Kutz, J. N. *Data-Driven Science and Engineering* (Cambridge University Press, 2022).
29. Genty, G. *et al.* Machine learning and applications in ultrafast photonics. *Nat. Photon.* **15**, 91–101 (2020).
30. Salmela, L. *et al.* Predicting ultrafast nonlinear dynamics in fibre optics with a recurrent neural network. *Nat. Mach. Intell.* **3**, 344–354 (2021).
31. Ermolaev, A. V., Sheveleva, A., Genty, G., Finot, C. & Dudley, J. M. Data-driven model discovery of ideal four-wave mixing in nonlinear fibre optics. *Sci. Rep.* **12**, 12711 (2022).
32. Mabed, M. *et al.* Machine learning analysis of instabilities in noise-like pulse lasers. *Opt. Express* **30**, 15060–15072 (2022).
33. Callaham, J. L., Koch, J. V., Brunton, B. W., Kutz, J. N. & Brunton, S. L. Learning dominant physical processes with data-driven balance models. *Nat. Commun.* **12**, 1016 (2021).
34. Sulem, C. & Sulem, P. *The Nonlinear Schrödinger Equation. Self-Focusing and Wave Collapse* (Springer, 1999).
35. Kibler, B. *et al.* The Peregrine soliton in nonlinear fibre optics. *Nat. Phys.* **6**, 790–795 (2010).
36. Kibler, B. *et al.* Observation of Kuznetsov-Ma soliton dynamics in optical fibre. *Sci. Rep.* **2**, 463 (2012).
37. Frisquet, B., Kibler, B. & Millot, G. Collision of akhmediev breathers in nonlinear fiber optics. *Phys. Rev. X* **3**, 041032 (2013).
38. Ryczkowski, P. *et al.* Real-time full-field characterization of transient dissipative soliton dynamics in a mode-locked laser. *Nat. Photon.* **12**, 221–227 (2018).
39. Tikan, A., Bielawski, S., Szwaj, C., Randoux, S. & Suret, P. Single-shot measurement of phase and amplitude by using a heterodyne time-lens system and ultrafast digital time-holography. *Nat. Photon.* **12**, 228–234 (2018).
40. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, 2008).
41. Callaham, J. L., Koch, J. V., Brunton, B. W., Kutz, J. N. & Brunton, S. L. Learning dominant physical processes with data-driven balance models. *Methods Codes Release Accomp. Publ.*https://doi.org/10.5281/zenodo.4428904 *(2021).*
42. Zakharov, V. E. & Gelash, A. A. Nonlinear stage of modulation instability. *Phys. Rev. Lett.* **111**, 054101 (2013).
43. Akhmediev, N., Ankiewicz, A. & Soto-Crespo, J. M. Rogue waves and rational solutions of the nonlinear schrödinger equation. *Phys. Rev. E* **80**, 026601 (2009).
44. Agafontsev, D. S. & Gelash, A. A. Rogue waves with rational profiles in unstable condensate and its solitonic model. *Front. Phys.* **9**, 610896 (2021).
45. Xu, G., Gelash, A., Chabchoub, A., Zakharov, V. & Kibler, B. Ghost interaction of breathers. *Front. Phys.* **8**, 608894 (2020).
46. Gelash, A., Xu, G. & Kibler, B. Management of breather interactions. *Phys. Rev. Res.* **4**, 033197 (2022).
47. Dudley, J. M., Genty, G. & Coen, S. Supercontinuum generation in photonic crystal fiber. *Rev. Modern Phys.* **78**, 1135–1184 (2006).
48. Zou, L., Luo, X., Zeng, D., Ling, L. & Zhao, L.-C. Measuring the rogue wave pattern triggered from Gaussian perturbations by deep learning. *Phys. Rev. E* **105**, 054202 (2022).
49. Gelash, A. A. & Zakharov, V. E. Superregular solitonic solutions: A novel scenario for the nonlinear stage of modulation instability. *Nonlinearity* **27**, R1-38 (2014).
50. Kraych, A. E., Agafontsev, D., Randoux, S. & Suret, P. Statistical properties of the nonlinear stage of modulation instability in fiber optics. *Phys. Rev. Lett.* **123**, 093902 (2019).
51. Gelash, A., Agafontsev, D., Suret, P. & Randoux, S. Solitonic model of the condensate. *Phys. Rev. E* **104**, 044213 (2021).
52. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
53. Zou, H., Hastie, T. & Tibshirani, R. Sparse principal component analysis. *J. Comput. Graph. Stat.* **15**, 265–286 (2006).
54. Jolliffe, I. T. & Cadima, J. Principal component analysis: A review and recent developments. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **374**, 20150202 (2016).

Author contributions

Simulations and analysis were performed by A.V.E. with guidance from J.M.D. All authors were jointly involved in validation and interpretation of the results obtained, and in the writing of the article. J.M.D. provided overall project supervision.

Funding

Academy of Finland (318082, 320165 Flagship PREIN, 333949); Centre National de la Recherche Scientifique (MITI Evènements Rares 2022); Agence Nationale de la Recherche (ANR-15-IDEX-0003, ANR-17-EURE-0002, ANR-20-CE30-0004). We thank Daniel Brunner and Pierre Colman for stimulating discussions.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-37039-7.

Correspondence and requests for materials should be addressed to J.M.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023