Elsevier Editorial System(tm) for Journal Of Computational Physics

Manuscript Draft

Manuscript Number:

Title: Nonlinear Laplacian Spectral Analysis of Rayleigh-Benard convection

Article Type: Regular Article

Keywords: rayleigh-benard convection; nonlinear laplacian spectral analysis; machine learning

Corresponding Author: Mr. Noah Domino Brenowitz, B.S.

Corresponding Author's Institution: New York University

First Author: Noah Domino Brenowitz, B.S.

Order of Authors: Noah Domino Brenowitz, B.S.; Dimitris Giannakis, BA, MSci, MPhil, PhD; Andrew J Majda, B.S., M.S., Ph.D.

Abstract: The analysis of physical datasets using modern methods developed in machine learning presents unique challenges and opportunities. These datasets typically feature many degrees of freedom, which tends to increase the computational cost of statistical methods and complicate interpretation. In addition, physical systems frequently exhibit a high degree of symmetry that should be exploited by any data analysis technique. The classic problem of Rayleigh Benard convection in a periodic domain is an example of such a physical system with trivial symmetries. This article presents a technique for analyzing the time variability of numerical simulations of two-dimensional Rayleigh-Benard convection at large aspect ratio and intermediate Rayleigh number. The simulated dynamics are highly unsteady and consist of several convective rolls that are distributed across the domain and oscillate with a preferred frequency. Intermittent extreme events in the net heat transfer, as quantified by the time-weighted probability distribution function of the Nusselt number, are a hallmark of these simulations.

Nonlinear Laplacian Spectral Analysis (NLSA) is a data-driven method which is ideally suited for the study of such highly nonlinear and intermittent dynamics, but the trivial symmetries of the Rayleigh-Benard problem such as horizontal shift-invariance can mask the interesting dynamics. To overcome this issue, the vertical velocity is averaged over parcels of similar temperature and height, which substantially compresses the size of the dataset and removes trivial horizontal symmetries. This isothermally averaged dataset, which is shown to preserve the net convective heat-flux across horizontal surfaces, is then used as an input to NLSA. The analysis generates a small number of orthogonal modes which describe the spatiotemporal variability of the heat transfer. A regression analysis shows that the extreme events of the net heat transfer are primarily associated with a family of modes with fat tailed probability distributions and low frequency temporal power spectra. On the other hand, the regular oscillation of the heat transfer is associated with a pair of modes with nearly uniform probability distributions. Physical mechanisms for the regular oscillation and the

extreme heat transfer events are hypothesized based on an analysis of the spatio-temporal structure of these modes. Finally, proposals are made for this approach to be applied to the study of other problems in turbulent convection, including three-dimensional Rayleigh-Benard convection and moist atmospheric convection.

Suggested Reviewers: Jorg Schumacher
joerg.schumacher@tu-ilmenau.de
Expert in Rayleigh Benard convection

Nadine Aubry
n.aubry@neu.edu
Expert in the data analysis of dynamical systems

Daan Crommelin
Daan.Crommelin@cwi.nl

**Significance and Novelty of this paper**

This manuscript presents an algorithm for the analysis of a numerical simulation of Rayleigh Benard convection using state of the art methods adapted from the machine learning community. There are two primary contributions:

1) The transient dynamics of two dimensional Rayleigh Benard convection are analyzed with a special focus on intermittency in the bulk heat transfer, which is quantified by the probability distribution of the Nusselt number.

2) An algorithm which combines nonlinear Laplacian spectral analysis (NLSA) and an isothermal averaging procedure is developed. This algorithm is specifically designed to take into account the trivial symmetries inherent to fluid dynamics simulations in idealized geometries. This algorithm generates mode families which successfully explain the intermittent aspects of heat transfer. The isothermal averaging procedure has the added benefit of substantially reducing the computational cost of the NLSA algorithm, which suggests that the algorithm is applicable to large three-dimensional simulations.

# Nonlinear Laplacian Spectral Analysis of Rayleigh-Bénard convection

N.D. Brenowitz[a,*], D. Giannakis[a], A.J. Majda[a]

[a] *Center for Atmosphere-Ocean Science, Courant Institute of Mathematical Sciences, New York University, New York, NY, USA 10025*

**Abstract**

The analysis of physical datasets using modern methods developed in machine learning presents unique challenges and opportunities. These datasets typically feature many degrees of freedom, which tends to increase the computational cost of statistical methods and complicate interpretation. In addition, physical systems frequently exhibit a high degree of symmetry that should be exploited by any data analysis technique. The classic problem of Rayleigh Benárd convection in a periodic domain is an example of such a physical system with trivial symmetries. This article presents a technique for analyzing the time variability of numerical simulations of two-dimensional Rayleigh-Bénard convection at large aspect ratio and intermediate Rayleigh number. The simulated dynamics are highly unsteady and consist of several convective rolls that are distributed across the domain and oscillate with a preferred frequency. Intermittent extreme events in the net heat transfer, as quantified by the time-weighted probability distribution function of the Nusselt number, are a hallmark of these simulations.

Nonlinear Laplacian Spectral Analysis (NLSA) is a data-driven method which is ideally suited for the study of such highly nonlinear and intermittent dynamics, but the trivial symmetries of the Rayleigh-Bénard problem such as horizontal shift-invariance can mask the interesting dynamics. To overcome this issue, the vertical velocity is averaged over parcels of similar temperature and height, which substantially compresses the size of the dataset and removes trivial horizontal symmetries. This isothermally averaged dataset, which is shown to preserve the net convective heat-flux across horizontal surfaces, is then used as an input to NLSA. The analysis generates a small number of

*Corresponding Author

orthogonal modes which describe the spatiotemporal variability of the heat transfer. A regression analysis shows that the extreme events of the net heat transfer are primarily associated with a family of modes with fat tailed probability distributions and low frequency temporal power spectra. On the other hand, the regular oscillation of the heat transfer is associated with a pair of modes with nearly uniform probability distributions. Physical mechanisms for the regular oscillation and the extreme heat transfer events are hypothesized based on an analysis of the spatio-temporal structure of these modes. Finally, proposals are made for this approach to be applied to the study of other problems in turbulent convection, including three-dimensional Rayleigh-Bénard convection and moist atmospheric convection.

## 1. Introduction

There are several aspects of physical datasets, be they from simulation, observation, or experiment, that complicate the straightforward application of advanced machine learning techniques. From a practical standpoint, these datasets are often extremely large with three spatial dimensions, and thousands time samples. Despite these thousands of time samples, it is clear that the resultant time series are frequently not fully sampled, and any technique should be able to cope with this. Secondly, physical datasets are frequently heterogeneous, with several relevant variables such as fluid velocity and temperature. Finally, unlike many industrial or biological datasets, the dynamics of physical systems are often governed by partial differential equations which can be simulated with a reasonable degree of accuracy on modern computer systems. This paper presents a physics-aware methodology for the analysis of physical datasets, that addresses each of these three issue in the context of a numerical simulation of Rayleigh-Bénard (RB) convection.

Rayleigh-Bénard convection is an important, but relatively simple, test-bed for the analysis of complex physical datasets. RB convection is the simplest prototype model for turbulent convection and has been extensively studied over the last century. It was an early example of convective instability (Chandrasekhar, 1981), and it forms the basis for Lorenz's famous work on chaos (Lorenz, 1963). A large part of the theoretical, numerical, and experimental work has gone into deriving scaling laws relating the important parameters of the problem at steady state (Ahlers et al., 2009;

Chillà and Schumacher, 2012, and references therein). However, the transient nature of the dynamics are also a topic of heavy interest (Kadanoff, 2001; Xi et al., 2009). Recent work has analyzed the transient aspects of RB convection using techniques such as empirical orthogonal functions (EOF) analysis (Bailon-Cuba et al., 2010)— which is also known by a wide array of names including principal components analysis, proper orthogonal decomposition (Aubry et al., 1991), and the discrete Karhunen-Loève expansion.

EOF analysis, however, has some faults. Because it ranks modes by variance explained, EOF analysis can fail to capture intermittent aspects of the dynamics. This concern is particularly relevant in the study of self-organizing nonlinear systems like RB convection, where the intermittent aspects of the data are frequently the most interesting. Moreover, when the system is undersampled in space, time, or otherwise, the dataset can suffer from loss of Markovianity. One technique for overcoming undersampling in dynamical systems is the classic method of expanding the dimension of the data by delay-embedding (Takens, 1981; Sauer et al., 1991). An important application of this principle is Singular Spectrum Analysis (SSA) (Vautard and Ghil, 1989), which combines Takens' embedding and EOF analysis. However, because SSA still selects modes based on variance explained it often under-emphasizes intermittent modes of variability and extreme events. A particularly striking example of the failure of variance-greedy algorithms is described by Crommelin and Majda (2004, Sec. 3), where the leading five EOFs of a six dimensional chaotic system capture over 99.9% of the energy, but fail to reproduce the chaotic dynamics. Specifically, the reduced dynamics converge towards a single fixed point.

To allow an accurate representation of intermittent and chaotic dynamics, Giannakis and Majda (2012b) recently proposed an algorithm known as Nonlinear Laplacian Spectral Analysis (NLSA). Like SSA, the NLSA algorithm consists of delay-embedding followed by an eigen-decomposition technique. NLSA improves upon SSA by using state of the art nonlinear manifold learning methods (Belkin and Niyogi, 2003; Coifman and Lafon, 2006) instead of linear EOF-based techniques. As a matter of course, NLSA is able to recreate the chaotic dynamics of the six-dimensional system described above using only three modes—the minimal dimensionality required for chaotic systems by the Poincare-Bendixson theorem (Strogatz, 2001). Since its original publication, NLSA has

been successfully applied to the analysis of highly complex datasets in the field of atmosphere-ocean science (AOS). Examples include the study of variability in North Pacific ocean currents (Giannakis and Majda, 2011, 2012a), the coupled dynamics of sea ice, the ocean, and the atmosphere (Bushuk et al., 2014; Bushuk and Giannakis, 2015; Bushuk et al., 2015), and the dynamics of cloud patterns in the tropics (Giannakis et al., 2012; Tung et al., 2014; Chen et al., 2014; Székely et al., 2015). However, unforced systems in simplified geometries like RB convection present unique challenges that are not present in climate datasets, which are strongly forced by effectively external factors such as the diurnal cycle, the seasonal cycle, topography, and anthropogenic sources.

Unlike the climate system, isolated systems in idealized geometries frequently have many trivial symmetries such as shift and Galilean invariance, which from a data analysis perspective can serve mask interesting aspects of the dynamics. For example, the EOFs of a system with shift invariance and periodic boundary conditions are essentially spatial Fourier modes (Aubry et al., 1993), a fact which only reflects the geometry of the system, not the underlying equations of motions. In the context of fluid mechanics, a useful approach to remove the effect of these symmetries can be achieved by averaging in an approximately Lagrangian frame (McIntyre, 1980).

A recent approach that is quasi-Lagrangian in nature has been proposed in Pauluis and Mrowiec (2013) for the study of moist-atmospheric convection. The authors there average various scalar quantities, such as humidity and vertical mass flux, over parcels of constant height and equivalent potential temperature, a temperature-like quantity which is a function of the humidity, temperature, and pressure. Because the equivalent potential temperature is nearly materially conserved in moist convective processes, this isentropic binning approach essentially replaces a spatial dimension such as $x$ with a dynamically significant dimension. Therefore, it is able to represent advective fluxes of scalars in a robust fashion. Taking time averages allows inference of the steady-state mixing and transport properties of the convective circulation. However, in applications, the unsteady dynamics of these convective flux quantities are of great interest.

In this paper, we use an analogous technique to generate a post-processed dataset for dry RB convection in an idealized geometry. Because RB convection is already isentropic by definition, it suffices to use temperature to stratify the data. The intuition is that temperature is nearly

materially conserved in regions with small temperature gradients. Moreover, parcels of the same temperature and height are likely to be similar to each other. Combining this isothermal binning procedure with NLSA analysis homes in on the interesting aspects of the heat transfer dynamics such as extreme events in the net heat transfer.

The outline of the paper follows. Section 2 provides a brief overview of RB convection. NLSA and the proposed isothermal-binning technique are described in Section 3. Section 4 describes the specifics of the numerical simulations performed and places emphasis on the intermittency of the bulk heat transfer dynamics. The integrated data analysis pipeline is performed on this dataset, and the results are discussed in Section 5. Finally, some concluding remarks are given in Section 6.

## 2. Rayleigh-Bénard Convection

Rayleigh-Bénard convection is perhaps the best understood model of turbulent convection (Kadanoff, 2001; Ahlers et al., 2009; Chillà and Schumacher, 2012). The setup consists of a thin layer of Boussinesq fluid suspended between two horizontal plates. The upper plate is held at a colder temperature than the lower plate, which leads to buoyant instability when the temperature difference is sufficiently large. The equations of motion for the layer of fluid can be derived from the full Navier-Stokes equations by assuming that density is a linear function of temperature $\rho = \rho_0 [1 - \alpha(T - T_0)]$, where $\alpha$ is known as the thermal expansion coefficient. Moreover, it is assumed that fluctuations in density are small compared to the background so that $1/\rho \approx 1/\rho_0$. On other hand, gradients in density induce pressure gradients which can drive the flow. Using these approximations, the momentum and continuity equations are given by

$$\frac{D\mathbf{v}}{Dt} = -\nabla p + g\alpha T \mathbf{e}_z + \nu \nabla^2 \mathbf{v}, \tag{1}$$

$$\nabla \cdot \mathbf{v} = 0, \tag{2}$$

where $\mathbf{v} = (u, v, w)$ and the material derivative is, as usual, given by $D/Dt = \partial_t + \mathbf{v} \cdot \nabla \mathbf{v}$. Note that the constant $T_0$ term has been absorbed into the pressure gradient. The reader familiar with atmospheric or oceanic dynamics will notice that these equations are a more general form of *hydrostatic* Boussinesq equation often used for geophysical flows. Indeed, a hydrostatic approximation

is not valid in regimes where the vertical and horizontal velocities are similar in magnitude, which is the case in RB convection. The temperature equation is given by a simple advection-diffusion equation

$$\frac{DT}{Dt} = \kappa \nabla^2 T, \tag{3}$$

where $\kappa$ is the thermal diffusivity. From this relation it is easy to see that temperature is nearly materially conserved when the diffusion term is small. This will typically be the case away from the boundaries, which motivates the isothermal binning procedure to be described in Section 3.3.

In the RB problem, a thin layer of fluid with width $d$ is suspended between an cold upper and warm lower surface. The vertical boundary conditions for the velocity are traditionally taken to be either free-slip or no-slip, and the temperature at the lower plate is fixed at $T(z = 0) = T_0$ and for the upper plate at $T(z = d) = T_0 - \Delta T$. A wide array of horizontal geometries and boundary conditions can be used, but in this paper we simply use periodic conditions.

The equations of motion can be cast in nondimensional form in a number of possible ways, but for the present purposes it is best to choose scales associated with the buoyant force. The typical length ($d$) and temperature scales ($\Delta T$) are provided by the geometry and boundary conditions. A typical velocity scale associated with the buoyant acceleration of the fluid is known as the free fall velocity scale and is given by $U_f = \sqrt{g\alpha d \Delta T}$. The advective free-fall time scale associated with the free-fall velocity is then given by

$$T_f = \frac{d}{U_f} = \sqrt{\frac{d}{g\alpha \Delta T}}. \tag{4}$$

For most interesting physical regimes, the free fall time is much faster than than diffusive or viscous time scales.

Using these scales to non-dimensionalize (1)–(3), leads to the system given by

$$\frac{D\mathbf{v}}{Dt} = -\nabla p + T\mathbf{e}_z + \frac{Pr^{1/2}}{Ra^{1/2}}\nabla^2 \mathbf{v}, \tag{5}$$

$$\frac{DT}{Dt} = \frac{1}{(RaPr)^{1/2}}\nabla^2 T, \tag{6}$$

$$\nabla \cdot \mathbf{v} = 0. \tag{7}$$

This system depends on the Rayleigh and Prandtl nondimensional numbers, which are, respectively,

given by

$$Ra = \frac{g\alpha\Delta T d^3}{\nu\kappa} \text{ and } Pr = \frac{\nu}{\kappa}.$$

The Rayleigh number can be interpreted as the ratio of bouyant and viscous forces or time scales, whereas the Prandlt number is simply the ratio of kinematic viscosity and thermal diffusivity. In both the free-slip and no-slip cases, this state with no fluid motion is linearly unstable when the Rayleigh number, exceeds some critical value $R_{crit} \approx 1000$ (Chandrasekhar, 1981); Rayleigh was the first to show this, which is why RB convection bears his name. The Rayleigh and Prandtl numbers are important controlling parameters, but the actual amount of heat transported by the flow depends on these two parameters in a nontrivial way. The Nusselt number, which quantifies the relative contribution of advective and diffusive heat fluxes is therefore an important diagnostic. It is given by

$$\text{Nu} = \frac{\frac{1}{L}\int_0^H \int_0^L wT - \kappa T_z dx dz}{\kappa\Delta T}, \tag{8}$$

where the integration is performed over the entire domain given by $0 \leq x < L$ and $0 \leq z < H$. The normalizing factor of $\kappa\Delta T$ is horizontally averaged heat flux due to conduction of heat, which can be seen by computing the integral

$$\int_0^H \kappa T_z dz = \kappa\Delta T. \tag{9}$$

A great deal of research has gone into studying the dependence of statistical steady-state relationships between Ra, Pr, and Nu on these two parameters (Kadanoff, 2001; Ahlers et al., 2009, and references therein), but the transient aspects of the relationship are less well understood. More plainly, in a given numerical simulation of RB convection the (8) can be calculated for each sample to generate a time series of Nu. In this paper, it is emphasized that $\text{Nu}(t)$ is a highly intermittent oscillatory time series with rich low frequency content in regimes with suitably high Rayleigh number. Understanding the dynamics of the Nusselt number in all their richness is the key target of the analysis below.

## 3. Spatio-temporal analysis of Rayleigh-Bénard Convection

As described in the introduction, the proposed data analysis pipeline for studying RB convection consists of an isothermal averaging procedure followed by NLSA. We begin by describing the general NLSA algorithm.

### *3.1. Nonlinear Laplacian spectral analysis*

The nonlinear Laplacian spectral analysis (NLSA) technique developed in Giannakis and Majda (2012b) is an extension of singular spectrum analysis (SSA) that targets intermittency and time-scale separation in dynamical systems. NLSA is a combination of ideas from spectral graph theory and dynamical systems, and has been used to analyze several datasets in field of atmosphere-ocean science (Bushuk et al., 2014; Székely et al., 2015).

NLSA, as well as SSA, are based on Takens' time-lagged embedding (Takens, 1981; Sauer et al., 1991). To describe the algorithm, assume the input data has $n$ degrees of freedom and $s$ time samples. For example, the degrees of freedom can be fluid temperature and velocity sampled over a regular grid. Let the data for a time sample $i$ be given by the $n$ dimensional vector $\mathbf{x}_i$. The common step between SSA and NLSA is called the lagged embedding step, wherein the dimension of this dataset is compounded by constructing a new dataset for which each sample includes $q$ lags of the given spatial pattern. Specifically, the $i^{\text{th}}$ sample of the $q$-lagged embedded data is given by $nq$-dimensional vector

$$\mathbf{x}_i^q = \begin{bmatrix} \mathbf{x}_i \\ \mathbf{x}_{i-1} \\ \vdots \\ \mathbf{x}_{i-(q-1)} \end{bmatrix}.$$

For convenience, we denote the phase-space of these lagged embedded vectors as $M = \mathbb{R}^{nq}$. These lagged-embeded samples are then collated in data matrix given by

$$X = \left[\mathbf{x}_1^q\, \mathbf{x}_2^q\, \ldots\, \mathbf{x}_{s-q}^q\right]. \tag{10}$$

At this point in the algorithm, NLSA fundamentally differs from SSA. SSA proceeds simply by performing a singular value decomposition (SVD) of $X$ and returning the associated empirical

orthogonal functions (EOFs) and principal components (PCs). The EOFs represent the spatial-temporal patterns over $q$-lags, and the principal components describe the temporal variability of these spatio-temporal patterns. Here, we note that the SVD is equivalent to solving for the eigenvalues and eigenvectors of the temporal covariance matrix $C = (s-q)^{-1}X^TX$. NLSA also returns EOFs and PCs, but does so by solving a different, dynamically-aware, eigenvalue problem.

Laplacian eigenmaps are a nonlinear dimension reduction procedure that are particular suited to the empirical analysis of nonlinear systems (Belkin and Niyogi, 2003). In machine learning, this and other techniques are known by the label "nonlinear manifold learning". In an extension of this work, Coifman and Lafon (2006) introduced a theory for Diffusion maps (DM) that forms the basis of the eigenvalue problem solved by NLSA. The formulation is a closely related to that of the covariance matrix, and is a form of kernel EOF analysis. In the formulation described in Giannakis and Majda (2012b), a kernel matrix between time samples is defined by

$$K_{ij} = \exp\left(\frac{\left\|\mathbf{x}_i^q - \mathbf{x}_j^q\right\|^2}{\epsilon\xi_i\xi_j}\right).$$

In addition to the Takens embedding, the dynamical systems point of view further enters through the normalizing values, $\xi_i = \left\|\mathbf{x}_{i+1}^q - \mathbf{x}_{i-1}^q\right\|$, which are the approximations to the velocity in phase space (Giannakis and Majda, 2012b). It has been shown both experimentally (Giannakis and Majda, 2012b) and theoretically (Berry et al., 2013) that using Takens embedding in conjunction with kernel methods enhances time scale separation in the computed eigenfunctions. The parameter $\epsilon$ is used to alter the preferred width of this kernel in phase space, and is typically chosen so that $\epsilon = O(1)$.

Following the DM machinery of Coifman and Lafon (2006), this kernel matrix, which is symmetric, is then normalized by a function of its row (or column) sums, $Q_i = \sum_j K_{ij}$. This gives an adjusted matrix

$$\tilde{K}_{ij} = \frac{K_{ij}}{Q_i^\alpha Q_j^\alpha},$$

where $\alpha$ is a parameter of the analysis. Finally, this kernel is turned into a probability matrix given

by

$$P_{ij} = \frac{\tilde{K}_{ij}}{\sum_j \tilde{K}_{ij}},$$

which represents a Markov chain defined on the underlying state space of the dynamical system.

The eigenvalues of the graph Laplacian of this probability matrix, which is given by $L = I - P$ are then found, typically using a computationally inexpensive iterative method. These eigenfunction/eigenvalue pairs, which are given by $L\phi_i = \lambda\phi_i$ for $i \in \mathbb{N}$ can be used as a new set of coordinates on the manifold. A more practical interpretation of these eigenfunctions as time series will be provided below in Section 3.2. The eigenfunctions of the graph Laplacian on the manifold are analogous to Fourier modes on the periodic real line. This machinery also predicts an equilibrium measure $\mu P = \mu$ which induces an inner product defined by

$$(u, v) = \sum_k^{s-q} u(k)v(k)\mu(k) \tag{11}$$

A key point is that the Laplacian eigenfunctions are orthogonal with respect to this inner product, so that $(\phi_i, \phi_j) = \delta_{ij}$, where $\delta_{ij}$ is the Kronecker delta. Therefore, they can be used as a convenient orthogonal basis for original dataset that respects the nonlinear structure of the manifold the data inhabit in phase space.

*3.2. NLSA in practice*

In the previous section, it was shown that the main output of NLSA is a set of Laplacian eigenfunctions $\{\phi_i\}_{i=1,2,\ldots}$ which are orthogonal with respect to the inner product induced by the metric $\mu$. The first interpretation of these eigenfunctions is as functions on the lagged embedded phase space $M$, which in mathematical symbols implies $\phi_i : M \to \mathbb{R}$. However, the NLSA analysis only returns the values of $\phi_i$ at a discrete set of locations in phase-space which sampled at uniform time intervals. Therefore, $\mu$ and the $\phi_i$ can be interpreted as discrete time series of length $s - q$. This can be written symbolically $\phi_i(t_j) = \phi_i(\mathbf{x}(t_j))$ with $t_j = j\Delta t$ and $q \leq j \leq s$. These time series are analogous to the principal components returned in an SSA or EOF analysis and can be analyzed in a similar fashion to any time series.

The phase-space structure of the Laplacian eigenfunctions can also be analyzed by projecting

any time series of length $s-q$ onto the $\phi_i$, using the inner product defined by (11). In more detail, let a matrix $Y \in \mathbb{R}^{m\times s-q}$, which can be interpreted as a discrete time series of length $s-q$ in am $m$-dimensional phase space, be given by $Y = [\mathbf{y}_1 \dots \mathbf{y}_{s-q}]$. The lag-embedded data matrix $X$ defined in (10) is an example. Then, each eigenfunction $\phi_i$ can be associated with a spatio-temporal pattern $\psi_i^Y \in \mathbb{R}^m$ by a simple weighted projection

$$\psi_i^Y = Y(\phi_i \odot \mu) \tag{12}$$

where "$\odot$" denotes elementwise multiplication of the column vectors $\phi_i$ and $\mu$. If the lagged-embedded data matrix, $X$, is used, then the column vectors $\psi_i^X$ can be interpreted as spatio-temporal patterns consisting of $q$ lags of the original non-embedded data. These spatial patterns are typically visualized as movies of a two or three dimensional field such as temperature or velocity.

The lagged embedding framework has numerous advantages and theoretical motivations (Takens, 1981; Sauer et al., 1991), but it is often easier to interpret modes in the original space. Here, a procedure is described that essentially allows the entire NLSA procedure to be treated as blind source separation that returns dynamically significant modes of variability. The typical algorithm in SSA and NLSA to accomplish this is known as reconstruction (Giannakis and Majda, 2012b; Vautard and Ghil, 1989; Vautard et al., 1992). The outer product of a spatio-temporal pattern and and its corresponding Laplacian eigenfunction, $M_i^X = \psi_i^X \otimes \phi_i$, can be regarded as function $M_i^X(k,j,\ell) = \psi_i^X(k+n\ell)\phi_i(j\Delta t)$ over the spatial index $k$, time index $j$, and lag index $\ell$. The redundancy between the time index $j$ and lag index $\ell$ means that any element of the set $\{M_i^X(\cdot,j,\ell)|t=(j-\ell)\Delta t\}$ can be identified as a spatial pattern at time $t$. A simple way to resolve this redundancy is to take an average over each of these sets. In particular, the "reconstructed" function associated with a Laplacian eigenfunction $\phi_i$ is given by

$$\tilde{M}_i^X(\cdot,t_j) = \frac{1}{q}\sum_{\ell=0}^{q-1} M_i^X(\cdot,j+\ell,\ell) \tag{13}$$

This reconstructed mode $\tilde{M}_i^X(k,t_j)$ can be interpreted as an anomaly due to the $i^{th}$ Laplacian eigenfunction, and can be visualized and analyzed in an identical fashion as the original non-embedded data. Summing the reconstructed modes allows for a convenient summary of the aggregate behavior

of different Laplacian eigenfunctions. The anomaly due to a set of Laplacian eigenfunctions $S$, is therefore given by

$$\tilde{M}_S^X = \sum_{i \in S} \tilde{M}_i^X. \tag{14}$$

Grouping modes is qualitative in nature, and typical methods include grouping modes with similar power spectra or statistical characteristics. If all the eigenfunctions are used (e.g. $S = \{1, \ldots, s-q\}$), then the original dataset, $X$, is recovered exactly by (14).

### *3.3. Isothermal binning*

There are, however, some caveats when any machine learning algorithm is applied directly to a large physical dataset with trivial and uninteresting symmetries. The approach described in Section 3.1 does not account for these symmetries. In the present context, the most basic of these is the spatial shift-invariance of two dimensional periodic RB problem. Practically speaking, the leading modes of an EOF or NLSA analysis of the full temperature or velocity fields consist of low frequency spatial shifts, which do not effect integral diagnostics such as bulk heat transfer and Nusselt number. These modes are, for our purposes dynamically uninteresting, and they mask the underlying convective dynamics. Therefore, we first develop a technique to remove these uninteresting symmetries. In addition, this technique saves significant computational expense. Both NLSA and SSA require computing the pairwise distances between every time sample, which is an operation that scales quadratically with the time-lagged dimension. Practically speaking solving the eigenvalue problem is comparatively cheap due to the use of cheap iterative eigenvalue solvers. While NLSA has been shown to scale to systems with millions of degrees of freedom, there are still practical limits when analyzing systems with billions of degrees of freedom such as the output of a three-dimensional fluid dynamics simulation.

Prior to analysis with NLSA, the RB data are compressed by using temperature rather than $x$ as a horizontal coordinate. As described in the introduction, this procedure is quite similar to the isentropic stream function analysis of Pauluis and Mrowiec (2013). However, since Boussinesq flows are already isentropic, the compression for this dataset works in temperature coordinates, $T$, rather than equivalent potential temperature. Essentially, the net effect of the procedure is to transform

the vertical velocity $w$, which is a function of $(x, z, t)$, as a function of $(T, z, t)$. In so doing, the Eulerian horizontal coordinate $x$ is replaced by the dynamically significant coordinate $T$.

This coordinate change is accomplished by averaging the vertical velocity over parcels of the same height and temperature. In the continuum setting, the formula which performs this is given by

$$w_T(T_0, z_0) \quad = \quad \iint \delta(T(x,z) - T_0)\delta(z - z_0)w(x,z)dxdz, \tag{15}$$

where $\delta(\cdot)$ is the Dirac delta function. A convenience of this averaging procedure is that, in the continuum limit, the convective heat flux can be given by

$$\int w_T(T', z)T'dT' = \int T\delta(T(x,z) - T')wdxdT' = \int wTdx, \tag{16}$$

so that integrated convective heat flux through a particular level can be perfectly recovered. In practice, however, (15) is implemented by dividing the temperature range into a finite number of bins, and averaging the vertical velocity within each bin; therefore, (16) is only approximately true in practice. On the other hand, diffusive heat fluxes cannot be captured using this approach, but as shown in (9), the bulk conductive transport of heat is constant.

In other words, the Nusselt number (c.f. Eq. 8), which is the primary diagnostic quantity of interest here is preserved under the isothermal averaging procedure. The benefit is that the convective heat flux is a linear functional of $w_T$, whereas it is a quadratic one of $w$ and $T$. Therefore, if $w_T$ can be expressed as a linear superposition of modes, which will be generated here by NLSA, then each mode can be associated with an unique heat-flux profile. Again, the isothermal binning procedure does not effect the overall heat transfer dynamics, and it removes the obfuscating presence of trivial horizontal symmetries. However, the averaging approach implicitly relies on the assumption that parcels with similar temperatures are somehow dynamically similar. In a highly organized system, like RB convection, there is a mechanism for the interaction of spatially distant parcels of the same temperature; however, in unorganized systems, the summation formulas above would likely lead to the cancellation of the interesting dynamics. This means that the non-locality of the underlying physical dataset should be considered when applying this technique.

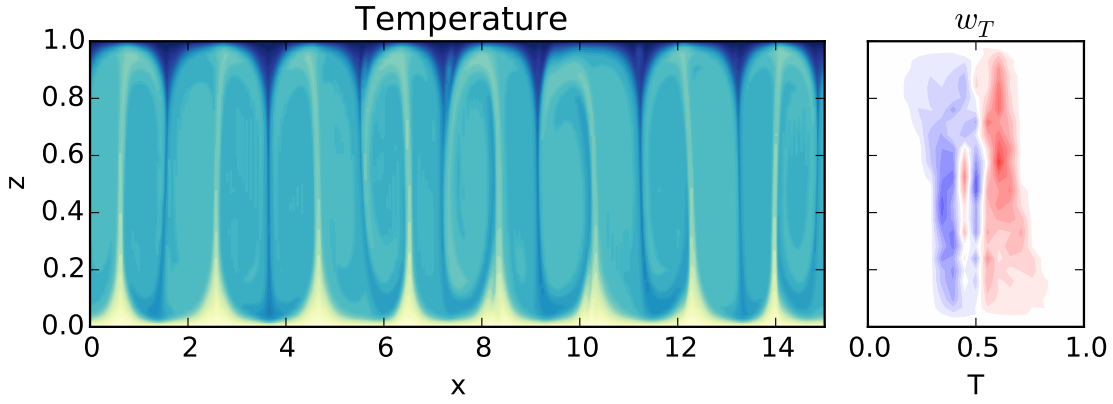


Figure 1: A single snapshot of the convection system at $t = 0$ (recall the first 10000 samples have been excluded). The raw temperature data are plotted (left) with temperature in color ranging between 0 (blue) and 1 (yellow). The isothermally averaged velocity is plotted on the right with color ranging between $-2$ (blue) and 2 (red). The vertical axis is z, and the horizontal axes are x (left) and $T$ (right). The flow consists of several convective cells (left), which are condensed into a single cell in $(T, z)$ space (right).

## 4. Simulation of two dimensional Rayleigh-Bénard Convection

The isothermal NLSA data analysis pipeline described in the previous section is now applied to a numerical simulation of two-dimensional RB convection for intermediate Rayleigh number in a large aspect ratio domain. As mentioned in Section 2, the geometry consists of a horizontal layer of fluid suspended between two plates which are separated by $d = 1$, and the temperature difference between the plates is $\Delta T = 1$. The vertical boundary conditions are no-slip, and the horizontal boundary conditions are periodic. To study the dynamics of multiple interacting convection cells, the horizontal extent of the domain is fixed at 15 times the width of the fluid layer, and periodic boundary conditions are used. The Rayleigh number is fixed at linearly unstable value, but non-turbulent, value of $Ra = 5 \cdot 10^6$, to allow for the emergence of intermittency and low-frequency oscillation in the resultant circulation. The Prandtl number is fixed at $Pr = 1$.

The simulations are performed using a spectral element code known as `Nek5000` (Fischer, 1997; nek5000, 2015). The numerical grid used is non-uniform but orthogonal in the $(x, z)$ plane. The domain is discretized using 15 square spectral elements distributed horizontally, which each contain $32 \times 32$ grid points. Therefore, the entire computational grid consists of 32 and $480 = 32 \cdot 15$ grid points in the vertical and horizontal directions, respectively. A time step of $1/600T_f$ is used, and

output is saved every 600 time steps, so that the output time step is the same as the free-fall time scale for this problem, which is the fastest intrinsic time scale of the flow. To reach statistical equilibrium, a long time series of 20000 samples is generated, and the first 10000 samples are excluded from the analysis pipeline. The isothermal binning procedure described in Section 3.3 is performed using 21 equally spaced bins between $T = 0$, which is the temperature of the top plate, and $T = 1$, the temperature of the bottom plate.

A single snapshot of the raw model output is shown in Figure 1. The flow consists of several convective cells with aspect ratios close to unity. The ascent (descent) occurs in narrow regions of warm (cold) fluid. When the isothermal averaging procedure described in Section 3.3 is carried out, the multiple cells in $(x, z)$ space are collapsed into a single cell in $(T, z)$ space. This cell is associated with the ascent (descent) of warm (cold) fluid.

The Nusselt number defined in (8) is the key number quantifying both the total heat transfer of the system as well as its dynamical character. A portion of the Nu time series is plotted in the first row of Figure 2 along with its power spectrum and probability density function (pdf). It is marked by a characteristic high frequency oscillation, which corresponds to the "wobble" of the circulation cells. The Nusselt number also experiences epochs of extremely low heat transfers which occur intermittently. As can be seen in the power spectrum, the Nusselt number has a spectral peak at a period of around $5T_f$, which is relatively high frequency. However, the time series also shows significant power for periods longer than $100T_f$. On a logarithmic axis, a gaussian pdf will appear as a parabola. By comparing the pdf of Nu, with a Gaussian pdf of the same mean and variance, it is clear the Nu pdf has a fat left tail. Also, the skewness in the pdf reflects the asymmetry in the time series of Nu.

As mentioned in Section 3.3, the binning procedure can introduce some errors in the calculated net heat transfer (e.g. Nusselt number). Figure 3 contains a scatter plot of Nu calculated using the original and isothermally binned data. The near perfect correspondence shows that the number of bins (21) chosen here is sufficiently accurate to resolve the heat transfer dynamics.

## 5. NLSA of the isothermally averaged vertical velocity

Having given a brief overview of the spatial, temporal, and statistical characteristics of the simulations, we perform NLSA on the isothermally averaged dataset. First, the temporal mean of the isothermally averaged vertical velocity $w_T(T, z)$ is calculated, and then NLSA is performed on the anomalies from this mean. A lagged embedding window of 20 is chosen to capture patterns of variability associated with the high frequency oscillation of Nu, which has a period of 5 $T_f$. The Laplacian eigenfunctions are generated for $\alpha = .5$ and $\epsilon = .3$. The results are fairly robust to changes in these parameters, and this particular set was chosen because it yields the cleanest results. For more information about how to choose the parameters $\alpha$ and $\epsilon$, we refer the interested reader to Giannakis and Majda (2012b).

A subset of the resulting Laplacian eigenfunctions ($\phi_i$), are visualized as time series in the left column of Figure 2. The corresponding spectral and statistical characteristics are shown in second and third columns, respectively. The first two eigenfunctions, $\phi_1$ and $\phi_2$, are a pair of modes which comprise the regular high frequency oscillation of Nu. They have a spectral peak at the same frequency as the Nusselt number, and have nearly uniform pdfs. In other words, these two modes are nearly equally active in all samples of the time series.

While the first two eigenfunctions capture the regular high-frequency behavior of the data, there are other eigenfunctions which capture the intermittent high frequency behavior, as well as a family of low frequency modes. For example, $\phi_{12}$ and $\phi_{13}$ have similar spectral characteristics to $\phi_1$ and $\phi_2$, but their distributions have symmetric fat tails. Most interestingly, visual comparison reveals that extreme events in the Nusselt number are associated with increased amplitude of these high frequency intermittent eigenfunctions. For instance compare the event around $t = 190$, where the Nusselt number drops to nearly 6. Finally, there are a family of eigenfunctions (e.g. $\phi_3$, $\phi_{10}$ and $\phi_{11}$) which describe the low-frequency variability of the data. It will be shown in the next section that these low-frequency eigenfunctions are related to the envelope of the Nusselt number.

In summary, NLSA yields an interesting decomposition of the data into eigenfunctions with different spectral characteristics. This efficient time scale separation is an advantage of NLSA that has been observed in a wide array datasets (Bushuk et al., 2014; Tung et al., 2014; Székely et al.,

2015). It is also striking that even within sets of eigenfunctions with similar spectral characteristics, NLSA separates signals which have nearly constant amplitude (e.g. 1 and 2) from highly intermittent modes such (e.g. 12 and 13). While the former effect is seen in a conventional SSA analysis (Vautard and Ghil, 1989), the separation of intermittent from regular components is a special strength of the NLSA algorithm (Giannakis and Majda, 2012b).

*5.1. Regression Analysis*

In the previous section, it was hinted that the Laplacian eigenfunctions are able to capture certain aspects of the Nusselt number behavior. We expand upon this here in quantitative fashion by performing two related linear regression analyses. Specifically, it is shown that a small subset of the Laplacian eigenfunctions plotted above are successful in capturing both the spectral and statistical features of Nusselt number. To this effect, a weighted linear regression analysis is performed by regressing the Nusselt number time series onto the Laplacian eigenfunctions. The first 20 Laplacian eigenfunctions and a constant are used, and the model can be written as a weighted least squares regression model

$$\mathrm{Nu}(t) = \alpha + \sum_{i=1}^{20} \beta_i \phi_i(t) + \sigma(t)\eta(t)$$

where $\eta(t)$ is a noise term and the standard deviation process $\sigma(t) = \mu(t)^{-1/2}$ is given by the invariant measure of the graph Laplacian. The particular statistical nature of $\eta(t)$ is not important here. It is only necessary that $\eta(t)$ be orthogonal to $\phi_i(t)$ for $i \in \{1, \ldots, 20\}$ according to the inner product defined by (11). Typically, the invariant measure, $\mu$, is close to a constant, and it is used here for consistency with the procedure described in Section 3.1. To highlight the low frequency character of the extreme events we also perform this regression on the rolling minimum time series

$$\mathrm{Nu}_{20}(t) = \min\left\{\mathrm{Nu}(t-q) : \ 0 \le q \le 20\right\}. \tag{17}$$

This time series is in essence the lower envelope of the Nusselt number. Moreover, using this time series resolves the ambiguity in how the eigenfunction time series should be aligned to the Nusselt number time series temporally. After all, the eigenfunction time series are $q$ samples shorter than the Nusselt number time series, and the first available sample of the eigenfunctions is associated

with all of the $q$ lags of the data.

Each target time series, $\mathrm{Nu}(t)$ and $\mathrm{Nu}_{20}(t)$, is regressed onto the first 20 eigenfunctions returned by NLSA. The resultant regression coefficients $\beta_1$–$\beta_{20}$ are plotted in Figures 4 and 6 for the Nu and $\mathrm{Nu}_{20}$ analyses, respectively. Then the magnitude $|\beta_i|$ is used to rank the eigenfunctions, and the improvements from including more eigenfunctions in each regression model is analyzed. A convenient visualization of the performance of a regression analysis is obtained by plotting the fitted values and the dependent variable together. This is shown for $\mathrm{Nu}(t)$ and $\mathrm{Nu}_{20}(t)$ in Figures 5 and 7, respectively. The log-densities of the fitted values are used to analyze the ability of the regression models to capture the intermittency and extreme events of the Nusselt number.

The raw Nusselt number regressions show the ability of a small number of Laplacian eigenfunctions to capture the primary oscillatory behavior of the heat transfer. The $R^2$ for the raw Nusselt number regression with 20 eigenfunctions is 0.833, and it is clear from Figure 5, that regression model consisting of only three eigenfunctions (e.g. 1,7, and 2) captures a remarkable amount of the variability. That said, the pdfs of the regression estimates do not share the fat left tail of the Nusselt number. It appears that the weighted least squares (WLS) minimization problem is masking the role of intermittent eigenfunctions. This makes sense because the larger share of the variability comes from the regular oscillation of the Nusselt number.

On the other hand, the regression for the rolling minimum Nusselt number (c.f. Figure 7) reveals the important role that the low-frequency components have in modulating the extreme heat transfer events. The $R^2$ for the rolling minimum analysis is .795. The most important regressor is $\phi_3$, which has a Gaussian or sub-Gaussian tails, but including the second ranked regressor, $\phi_{10}$, allows for high-fidelity tracking of extreme events.

For both cases, we have shown that a small number of eigenfuctions are able to quantitatively capture the variance and intermittency of the Nusselt number. In particular, the bulk of the variance of the Nusselt number is captured using a set of high-frequency Laplacian eigenfunctions with nearly constant amplitude, while the intermittent aspects of the signal can be described using fat tailed low frequency eigenfunctions.

*5.2. Reconstructed mode family composites*

Having demonstrated that the Laplacian eigenfunctions adequately reconstitute the temporal behavior of the Nusselt number in a statistical sense, it is important to get a sense of what the modes look like spatially. Using the regression analysis, as well as the statistical and spectral characteristics of the eigenfunction time series as guides, three families of modes are been selected. These are:

- "oscil": A family consisting of the first two eigenfunctions, $S = \{1, 2\}$, which are a pair of modes the capture the bulk uniform oscillation of the Nusselt number (c.f. Figure 5).
- "high": A family with $S = \{12, 13\}$, which have high temporal frequency, but more fat-tailed distributions than the $\phi_1$ and $\phi_2$.
- "low": A family consisting of the low frequency modes, $S = \{3, 10, 11\}$, that explain the variance of the rolling minimum Nusselt number (c.f. Figure 7).

The choice of these families is not intended to provide a comprehensive survey of the variability. Rather these modes families were largely chosen to provide a physical picture to the modes that were played an important role in the regression analysis of the previous section.

The reconstruction procedure described in Section 3.2 is used to calculate reconstructed anomaly fields for each mode family described above. This reconstruction is performed for two different fields. The first is the isothermally averaged data $w_T(T, z)$ that was the input to the NLSA algorithm. This provides a convenient summary of each mode family's effect on the heat transfer. For a given mode family $S$, the $w_T(T, z, t)$ reconstructions are denoted by $\tilde{M}_S^w(x, T, t)$. These fields can be visualized as movies (see the supplement), but generating averages which are conditional on the Nusselt number provides are more coherent summary of each mode family. In particular, the range of the Nusselt number is divided into two intervals $\mathrm{Nu} < 8$, and $14 \leq \mathrm{Nu}$, which are chosen to represent the left-tail and right-tail of the distribution. The reconstructed fields are then temporally averaged conditional on these intervals. The overall temporal average of the fields is also calculated.

As mentioned in Section 3.3, the isothermal binning allows for straightforward reconstruction of the transient heat flux, and it is possible to associate each NLSA mode with a convective temper-

ature flux profile. This is one of the key benefits of the isothermal binning approach. A centered version of the convective heat flux due to a mode family $S$ is given by

$$\overline{M_S^w \cdot (T - 0.5)} = \int_0^1 M_S^w (T - 0.5) dT. \tag{18}$$

This formula is derived by analogy with (16), and the offset of 0.5 is used to center the resultant anomaly.

The $\tilde{M}_S^w$ composites are shown in Figure 8, and the associate heat flux composites for each vertical level $z$ are contained in Figure 9. The "oscil" mode family has a simple structure and amplitude relationship. The behavior is rather simple given that left-tail of the Nusselt number is associated with decreased heat transfer. For low Nusselt number, the "oscil" family is associated with the upward (downward) motion of moderately low (high) temperature fluid, which would tend to depress the heat transfer. The composite for $Nu > 14$ has a very similar structure in $(T, z)$ space, but it has an opposite sign. Finally, the overall temporal average is near zero, which indicates that the "oscil" mode family represents a regular symmetric oscillation. The convective heat flux plots in Figure 9 confirm that the "oscil" family causes no net heat-flux, and that the $Nu < 8$ conditional average is nearly equal-and-opposite the $Nu \geq 14$ average. An additional detail is the composites show that boundary layer and mid-domain heat-fluxes are of opposite sign, although the mid-domain heat flux dominates.

The "high" mode family is also has a high frequency oscillatory behavior, but the composites show that it is preferentially active when the Nusselt number is reduced. Its structure in $(T, z)$ space is similar to that of the "oscil" family in the center of the domain, but different near $z = 0$ and $z = 1$. The pattern also appears to tilt to the left. The vertical structure and asymmetric response to the Nusselt number are particularly clear from the convective heat-fluxes in Figure 9. As seen in Figure 2, the eigenfunctions which make up the "high" family ($\phi_{12}$ and $\phi_{13}$) have extremely fat tailed but symmetric distributions. This provides a demonstration that nearly symmetric eigenfunctions can have an asymmetric impact on transient aspects of the convective heat transfer. The intermittency of these modes is potentially crucial in provoking this asymmetric response.

To summarize, the high-frequency eigenfunctions played a prominent role in the regression analysis of the raw Nusselt number in Figure 5, but the low-frequency modes were vital in reproducing extreme events in the rolling minimum time series defined by (17). Like the "high" family, the "low" family appears to be preferentially active when the the net heat transfer (i.e. Nusselt number) is weak, with opposite sign when $Nu \geq 14$. On the other hand, the "low" family has a simpler vertical structure without any sign changes. In $(x, z)$ space, the pattern appears to tilt to the right. Of the three modes, the "low" family is responsible for the largest anomalies anomalies when $Nu < 8$, which further indicates that extreme events in the Nusselt number are more strongly associated with low-frequency changes than high frequency ones.

*5.3. Spatial visualisation and discussion*

Having provided a brief survey of the effect of the three different mode families in the isothermally binned $(T, z)$ space, it is illuminating to study the structure in the original $(x, z)$ space. This may appear somewhat odd given the original motivation for using the isothermally binning procedure, but viewing the practical NLSA procedure described in Section 3.2 as a black-box provides this flexibility, However, the spatial visualizations contain horizontal drift, so it is not feasible to visualize it using averaged composites. Instead, we analyze the dynamics of the system before, during, and after a particular extreme heat transfer event which happens between $t = 4950$ and $t = 5150$. For the rest of this section, we will use the time from the beginning of this interval to refer to specific time samples.

The portion of the Nusselt number time series for this section is plotted in Figure 10. From $t = 0 - 50$, there is a quiescent period where the oscillations of the heat transfer are damped substantially. This transitions to a period of strong oscillations between $t = 75 - 150$, which account for both high and low extreme events. The spatial reconstructions for the "oscil" and "low" mode families (e.g. $\tilde{M}_S^T(x, z, t)$), along with the original data, during four representative times are shown in Figure 11, and a full video of the event is available as supplementary material. The discussion here will focus on surveying the high frequency regular oscillation and intermittent low-frequency dynamics of the Nusselt number, and the dynamics of the "high" mode family will not be discussed. The reconstructed NLSA mode families will be used to provide plausible physical

hypotheses.

*5.3.1. High frequency oscillation*

The dominant aspect of the visualization is the regular back-and-forth oscillation of the original data. The period of the oscillation is close to 10 $T_f$, and essentially consists of the coordinated tilting to the left or right of all the ascent/descent plumes. Since the observed value of velocity is close to $U = 0.4$, and the diameter of the cells is near $d = 1$, the turnover time scale of the circulation is given by $U(d\pi)^{-1} \approx 10$. This connection between the turnover time and the oscillation period indicates that the oscillation could be caused by advection of temperature and momentum anomalies by the convection rolls.

The impact of this oscillation on the Nusselt number is apparent. The heat transfer is maximized when the ascent/descent plumes are nearly vertical, and minimized when the plumes are tilted from the vertical. Therefore, the Nusselt number oscillates with half the period of the regular oscillation as the plumes tilt from left to right. These dynamics also hint at the reason why there is a fat left tail, but no fat right tail, in the distribution of the Nusselt number (c.f. Figure 2, panel 1). Basically, a tilted circulation can become more strongly tilted, but the un-tilted circulation is tightly constrained. Therefore, the minima of the Nusselt number, which are associated with the tilted states, can take more extreme values than the maxima. The degree to which the convection rolls are tilted depends on the low frequency state of the system, as will be discussed in the next section.

As was seen in Section 5.1, this oscillation is strongly associated with the "oscil" family of reconstructed modes. An interesting aspect of this mode family is its near horizontal homogeneity. More specifically, for $t = 43$, the temperature anomalies near $z = .15$ and $z = 85$ are of opposite sign, but nearly constant in $x$. Moreover, there is a sign change in the boundary layers that corresponds to the interesting vertical structure seen in the heat-flux composites seen in Figure 9. This horizontal homogeneity is not shared by the corresponding velocity anomalies, which tend to follow the roll-like structure of the circulation. This is an indication that it is the advection of momentum, not temperature, that is essential to the dynamics of the regular oscillation. In other words, it is likely a coordination of the velocity field which causes the regular oscillation, rather than stochastic plume dynamics.

This conclusion leads to a possible physical mechanism. Consider the case with two cells in a periodic domain. The left cell rotates counterclockwise and the right cell rotates clockwise. A positive $u$ anomaly near the lower left boundary will be advected around the circulation to the upper boundary in about 5 time units. If the anomaly is significant, it will push the upper circulation to the right. However, if pushed too far to the right, the circulation will be gravitationally unstable, and it will attempt to restore itself by sloshing back to the left. This restoration coincides with the arrival of negative $u$ from the lower half of the right cell, which completes the cycle. Therefore, this oscillation represents the combined effect of buoyant dynamics within each circulation cell, which is coordinated between cells by the momentum transfer. Essentially, the buoyancy generates momentum, and the regular oscillation tosses momentum back and forth between the cells. These pairwise connections lead the oscillation to synchronize throughout the domain. This physical mechanism essentially slaves the temperature anomalies to the velocity field—a conclusion that has been demonstrated for the analogous "sloshing" mode in an experiments of three-dimensional RB convection (Xi et al., 2009).

*5.3.2. Low frequency dynamics and extreme events*

While the high-frequency dynamics can explain the regular oscillation of Nu, the extreme events are associated with longer temporal scales. Qualitatively, the extreme heat transfer events seems to occur when the circulation is strongly tilted and more disorganized. This can be seen by comparing the original dataset visualizations available in Figure 11 for $t = 43$ and $t = 83$, respectively.

The low frequency dynamics of the Nusselt number are associated with the "low" mode family, as was seen in Section 5.1. Visualizing this mode family allows us to isolate the low-frequency intermittent aspects of the signal. In terms of the single heat transfer event plotted in Figure 10, the "low" mode family proceeds from a quiescent phase ($t = 0 - 50$) to a more active phase ($t \geq 50$). Interestingly, the appearance of large reconstructed temperature and velocity anomalies due to "low" family coincides with the increase in the magnitude of the Nusselt number oscillation from $t = 50$ to $t = 100$.

During this time period, the "low" mode family reconstruction passes through two phases with distinct spatial patterns. Both patterns are characterized by anomalously cold/warm patches near

the boundary layers, but, unlike the "oscil" mode family, the temperature anomalies vary greatly in $x$, and are aligned with the individual convection rolls. As might be expected, the cold (warm) patches are associated with downward (upward) motion. The first phase of "low" is typified by $t = 54$, and consists of larger cold (warm) anomalies near the lower (upper) boundary. While this how the stronger anomalies are organized, there are weaker temperature anomalies inside the circulation cell (see $(x, z, t) = (5.5, .8, 54)$). When this phase is active, the regular high frequency oscillation described above is largely suppressed. The second phase of the "low" mode family is typified by $t = 83$. In this case, the temperature anomalies near the upper and lower boundaries have the same sign and magnitude. Typically, this is associated with greater disorder in the full solution and more extreme oscillation in the Nusselt number.

The fact that the appearance of the first phase of the "low" family occurs before the extreme heat transfer event suggests a possible mechanism for the low frequency oscillation and fat left tail of the Nusselt number. The asymmetry of the first phase suggests that process is initiated with plumes that penetrate the boundary layer especially well. This is further supported by the fact that the largest cold (warm) anomalies occur beneath (above) the descending (ascending) regions, and could potentially explain the differences in vertical heat-flux structure seen in Figure 9. When the plume penetrates deeper into the boundary layer, it eventually causes enhanced re-circulation of cold fluid to the upper level by advection, which tends to decrease the efficiency of the circulation. A parallel perspective, is that deep penetrating plumes cause stronger horizontal gradients in temperature, which increase the potential energy available to the system. This potential energy increases until the the roll-like structure of the flow destabilizes as seen around $t = 100$. The exact mechanism for this destabilization is not clear. This analysis of a single heat transfer event suggests that abnormally deep and strong plumes in a few places in the domain cause the appearance of the heat transfer events. But it is uncertain what initiates the plumes and coordinates the actions between non-adjacent circulation cells.

## 6. Conclusions and future work

This paper presents a method for analyzing the temporal variability of two-dimensional Rayleigh-Bénard convection—a prototype model for high-dimensional physical datasets. This method reveals interesting aspects of the dynamics which would be otherwise hidden by trivial symmetries of the data that account for much of the variance of the dataset. It does so by employing a quasi-Lagrangian approach which averages the vertical velocity over parcels of constant temperature and height. This dataset, which is substantially reduced in size, allows for a practically lossless reconstruction of the convective heat-flux, a quadratic quantity.

The intermittent aspects of the isothermally averaged dataset are then revealed using nonlinear Laplacian spectral analysis. In particular, the dynamics are separated into a families of orthogonal modes which each capture distinct aspects of the dynamics. These modes can be divided by the temporal spectral characteristics and statistical features. Namely, there are high/low frequency modes and intermittent/non-intermittent modes. The Nusselt number, which quantifies the bulk averaged heat transfer, is used as a guide in forming these families. The regular oscillation of the Nusselt number is captured by a pair of modes with relatively uniform amplitude in time, and entirely symmetric heat transfer characteristics in vertical. On the other hand, it is shown that the intermittent aspects of the bulk heat transfer are primarily associated with a family of low-frequency modes, which are preferentially active during the low heat transfer events. Notably, the variability of these modes occurs at slower time scales than the uniform oscillation of the heat transfer, and they are responsible for negative heat transfer at all vertical levels. There is also an intermittent family of modes which is preferentially active during the low heat transfer events, but has the same temporal frequency as the regular oscillation.

In summary, it appears that the extreme events of the heat transfer are primarily associated with the low and high frequency modes with intermittent temporal characteristics. When visualized in physical space these modes, both the temperature and velocity anomalies aligns with the individual convection rolls and are non-homogeneous in $x$. This notably contrasts with the non-intermittent high-frequency mode which has a nearly uniform horizontal structure. On the other hand, the high-frequency modes share a similar vertical structure with sign changes in convective heat flux

near the vertical boundaries.

In addition to providing some physical insight into the dynamics of two-dimensional Rayleigh-Bénard convection, this paper is intended to demonstrate a clear path forward for the temporal analysis of turbulent convection. The two-dimensional Rayleigh-Bénard dataset is used primarily as a toy problem demonstrating the utility of our approach, and it is crucial to extend this to more complicated systems. As a next step, the isothermal averaging approach can be readily applied to three-dimensional convection Rayleigh-Bénard datasets. The computational expense of temporal analysis techniques like EOFs, SSA, and NLSA is increased greatly by including an additional spatial direction, so the compression obtained by isothermal averaging will be greatly helpful in conquering this "curse of dimension". Finally, we hope to employ an analogous technique for the study of moist atmospheric convection, a problem of great importance to climate and weather prediction. This can be accomplished relatively straightforwardly by using an isentropic rather than isothermal binning procedure as indicated by (Pauluis and Mrowiec, 2013).

**Acknowledgement**

The research of A. J. M. is partially supported by the Office of Naval Research MURI award grant ONR-MURI N-000-1412-10912, and N.D.B. is supported as a graduate student on the MURI award. D.G. is supported by the National Science Foundation grant DMS-1521775 and the Office of Naval Research grant N-000-1414-10150.

**References**

Ahlers, G., Grossmann, S., Lohse, D., Apr. 2009. Heat transfer and large scale dynamics in turbulent Rayleigh-B\'enard convection. Reviews of Modern Physics 81 (2), 503–537.

Aubry, N., Guyonnet, R., Lima, R., Aug. 1991. Spatiotemporal analysis of complex signals: Theory and applications. Journal of Statistical Physics 64 (3-4), 683–739.

Aubry, N., Lian, W., Titi, E., Mar. 1993. Preserving Symmetries in the Proper Orthogonal Decomposition. SIAM Journal on Scientific Computing 14 (2), 483–505.

Bailon-Cuba, J., Emran, M. S., Schumacher, J., Jul. 2010. Aspect ratio dependence of heat transfer and large-scale flow in turbulent convection. Journal of Fluid Mechanics 655, 152–173.

Belkin, M., Niyogi, P., Jun. 2003. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. Neural Computation 15 (6), 1373–1396.

Berry, T., Cressman, J., Gregurić Ferenček, Z., Sauer, T., Jan. 2013. Time-Scale Separation from Diffusion-Mapped Delay Coordinates. SIAM Journal on Applied Dynamical Systems 12 (2), 618–649.

Bushuk, M., Giannakis, D., Jul. 2015. Sea-ice reemergence in a model hierarchy. Geophysical Research Letters 42 (13), 5337–5345.

Bushuk, M., Giannakis, D., Majda, A. J., May 2014. Reemergence Mechanisms for North Pacific Sea Ice Revealed through Nonlinear Laplacian Spectral Analysis. Journal of Climate 27 (16), 6265–6287.

Bushuk, M., Giannakis, D., Majda, A. J., Feb. 2015. Arctic Sea Ice Reemergence: The Role of Large-Scale Oceanic and Atmospheric Variability. Journal of Climate 28 (14), 5477–5509.

Chandrasekhar, S., Feb. 1981. Hydrodynamic and Hydromagnetic Stability, dover edition Edition. Dover Publications, New York.

Chen, N., Majda, A. J., Giannakis, D., Aug. 2014. Predicting the cloud patterns of the Madden-Julian Oscillation through a low-order nonlinear stochastic model. Geophysical Research Letters, n/a–n/a.

Chillà, F., Schumacher, J., Jul. 2012. New perspectives in turbulent Rayleigh-Bénard convection. The European Physical Journal E 35 (7).

Coifman, R. R., Lafon, S., Jul. 2006. Diffusion maps. Applied and Computational Harmonic Analysis 21 (1), 5–30.

Crommelin, D. T., Majda, A. J., Sep. 2004. Strategies for Model Reduction: Comparing Different Optimal Bases. Journal of the Atmospheric Sciences 61 (17), 2206–2217.

Fischer, P. F., May 1997. An Overlapping Schwarz Method for Spectral Element Solution of the Incompressible Navier–Stokes Equations. Journal of Computational Physics 133 (1), 84–101.

Giannakis, D., Majda, A. J., 2011. Time series reconstruction via machine learning: Revealing decadal variability and intermittency in the North Pacific sector of a coupled climate model. Mountain View, CA, pp. 107–117.

Giannakis, D., Majda, A. J., Dec. 2012a. Limits of predictability in the North Pacific sector of a comprehensive climate model. Geophysical Research Letters 39 (24), n/a–n/a.

Giannakis, D., Majda, A. J., 2012b. Nonlinear Laplacian spectral analysis for time series with intermittency and low-frequency variability. Proceedings of the National Academy of Sciences 109 (7), 2222–2227.

Giannakis, D., Tung, W.-w., Majda, A. J., 2012. Hierarchical structure of the Madden-Julian oscillation in infrared brightness temperature revealed through nonlinear laplacian spectral analysis. In: Intelligent Data Understanding (CIDU), 2012 Conference on. IEEE, pp. 55–62.

Kadanoff, L. P., 2001. Turbulent Heat Flow: Structures and Scaling. Physics Today 54 (8), 34–39.

Lorenz, E. N., Mar. 1963. Deterministic Nonperiodic Flow. Journal of the Atmospheric Sciences 20 (2), 130–141.

McIntyre, M. E., Mar. 1980. An introduction to the generalized Lagrangian-mean description of wave, mean-flow interaction. pure and applied geophysics 118 (1), 152–176.

nek5000, 2015. Nek5000 | A Spectral Element code for CFD.
URL `http://nek5000.mcs.anl.gov/`

Pauluis, O. M., Mrowiec, A. A., Jul. 2013. Isentropic Analysis of Convective Motions. Journal of the Atmospheric Sciences 70 (11), 3673–3688.

Sauer, T., Yorke, J. A., Casdagli, M., Nov. 1991. Embedology. Journal of Statistical Physics 65 (3-4), 579–616.

Strogatz, S. H., 2001. Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering, 2nd Edition. Studies in Nonlinearity. Perseus Books, Cambridge, Mass.

Székely, E., Giannakis, D., Majda, A. J., May 2015. Extraction and predictability of coherent intraseasonal signals in infrared brightness temperature data. Climate Dynamics, 1–30.

Takens, F., 1981. Detecting strange attractors in turbulence. In: Rand, D., Young, L.-S. (Eds.), Dynamical Systems and Turbulence, Warwick 1980. No. 898 in Lecture Notes in Mathematics. Springer Berlin Heidelberg, pp. 366–381.
URL http://link.springer.com/chapter/10.1007/BFb0091924

Tung, W.-w., Giannakis, D., Majda, A. J., Jul. 2014. Symmetric and Antisymmetric Convection Signals in the Madden–Julian Oscillation. Part I: Basic Modes in Infrared Brightness Temperature. Journal of the Atmospheric Sciences 71 (9), 3302–3326.

Vautard, R., Ghil, M., 1989. Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series. Physica D: Nonlinear Phenomena 35 (3), 395–424.
URL http://www.sciencedirect.com/science/article/pii/0167278989900778

Vautard, R., Yiou, P., Ghil, M., 1992. Singular-spectrum analysis: A toolkit for short, noisy chaotic signals. Physica D: Nonlinear Phenomena 58 (1), 95–126.
URL http://www.sciencedirect.com/science/article/pii/016727899290103T

Xi, H.-D., Zhou, S.-Q., Zhou, Q., Chan, T.-S., Xia, K.-Q., Jan. 2009. Origin of the Temperature Oscillation in Turbulent Thermal Convection. Physical Review Letters 102 (4), 044503.

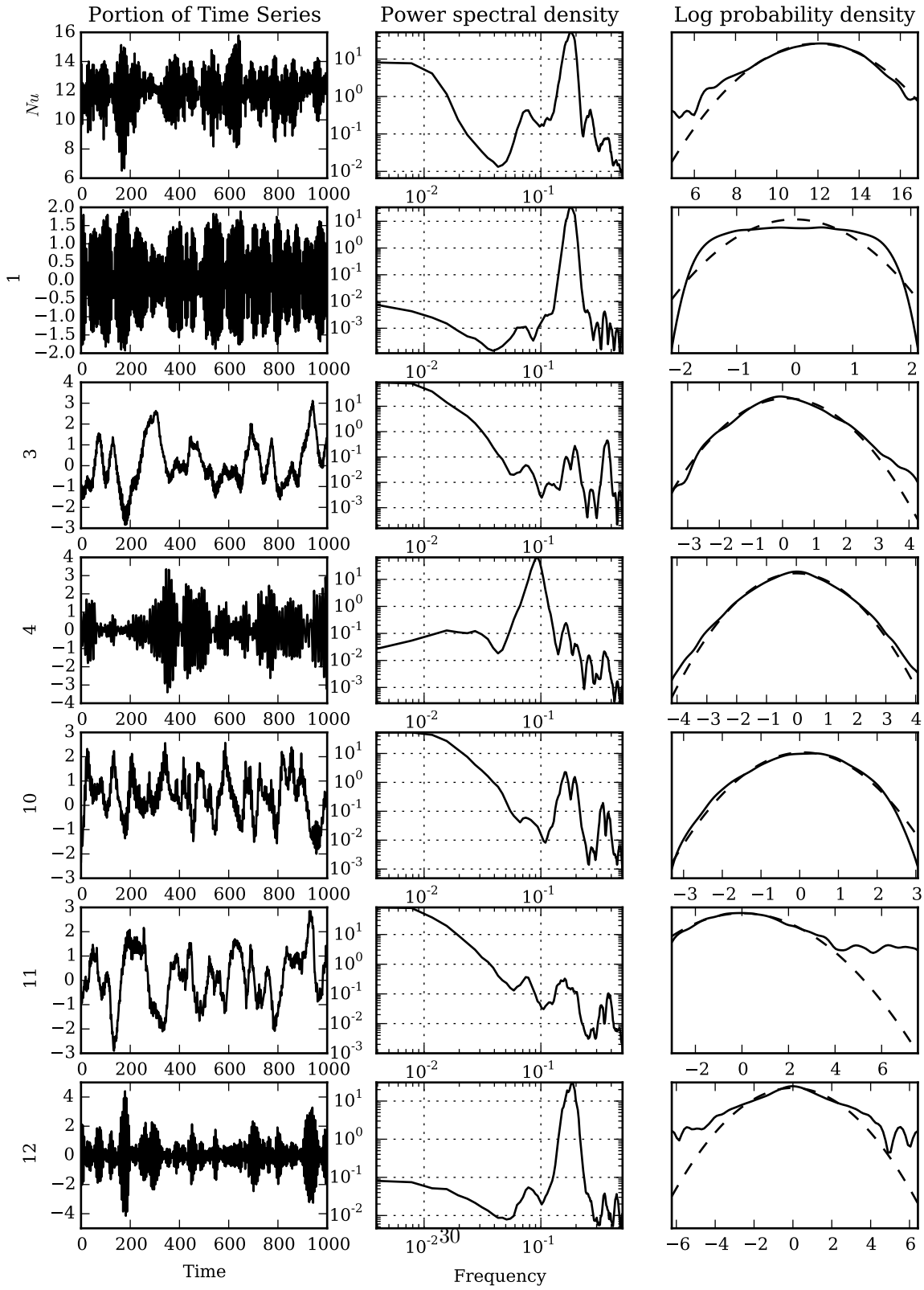


Figure 2: Visualizations of Laplacian eigenfunctions and Nusselt number. Time series portions (left), power spectra densities (middle), and log-probability plots (right) are plotted. The time series portions are only shown for $0 \leq t \leq 1000$, but the full 10000 samples are used to calculate the power spectra and pdfs. The log-probability plots show the kernel density estimate (black) and a Gaussian comparison (dashed). The Nusselt number is plotted in the top row, and in the subsequent rows the labels indicate the index of the eigenfunction.

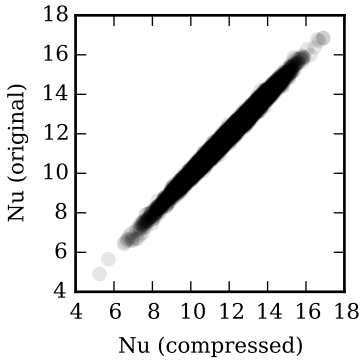


Figure 3: Scatter plot comparing the Nusselt number calculated on the raw and isothermally averaged datasets. The nearly perfect correspondence between the two is even present in the tail of the distribution.

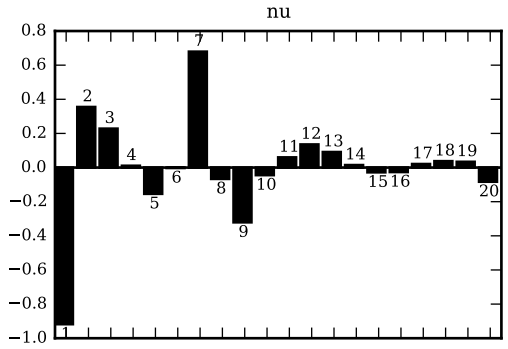


Figure 4: Estimated regression coefficients of the full Nusselt number, Nu, regressed onto the first twenty Laplacian eigenfunctions.

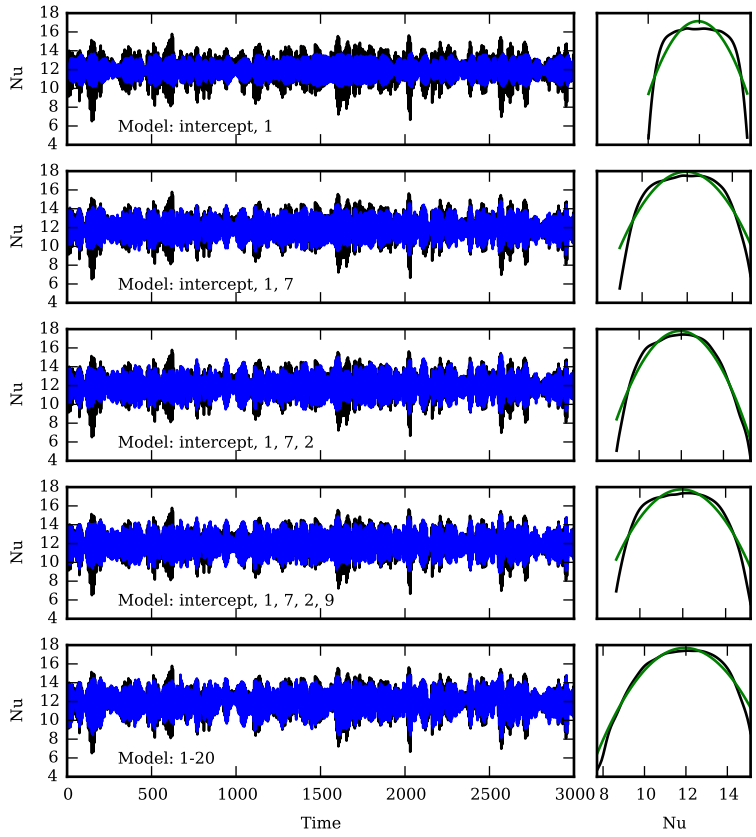


Figure 5: Performance of the regression analysis for the full Nusselt number. Each row includes succesively more regressors in the analysis, and consists of the indicated Laplacian eigenfunctions. The Nusselt number (black) and the estimated model (blue) are shown on the left. The log pdfs (black) are shown on the right with Gaussian comparison (green). The full regression analysis has difficulty capturing the tails of the Nusselt number despite the presence of intermittent eigenfunctions. This occurs because linear regression is not tailored for extreme events, and this analysis could likely be improved using a general linear model. The $R^2$ retaining 20 regressors is .833.

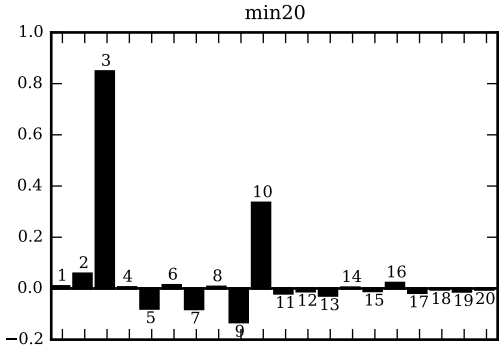


Figure 6: Like Figure 4 but for the coefficients of the rolling minimum nusselt number, $\mathrm{Nu}_{20}$, regressed onto the first twenty Laplacian eigenfunctions.

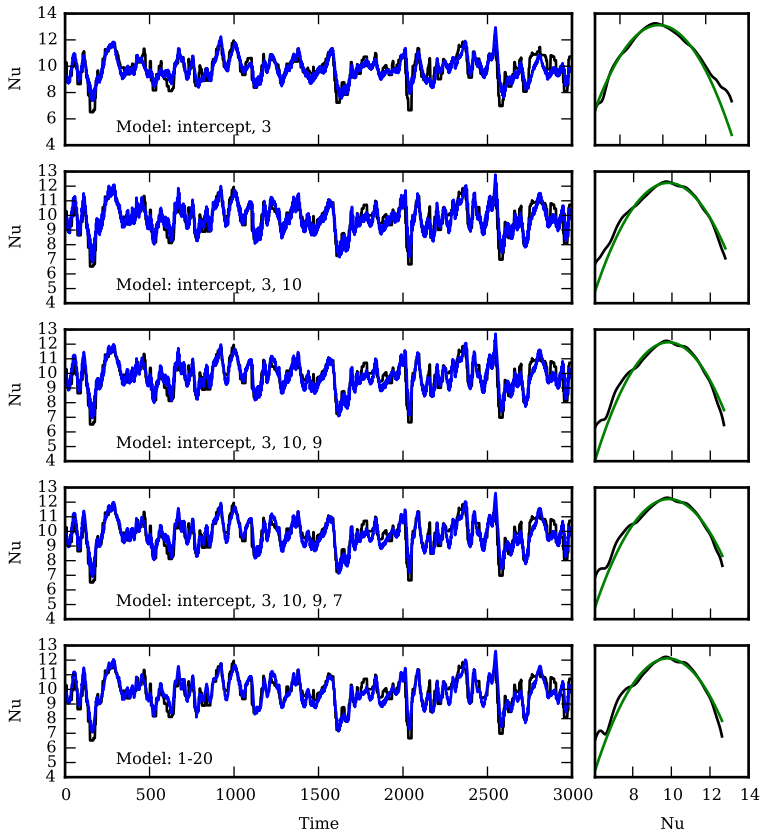


Figure 7: Like Figure 5 but for the regression analysis for the rolling minimum Nusselt number, $\mathrm{Nu}_{20}$. The low frequency eigenfunctions capture the fat tails with high fidelity. The $R^2$ retaining 20 regressors is .795.

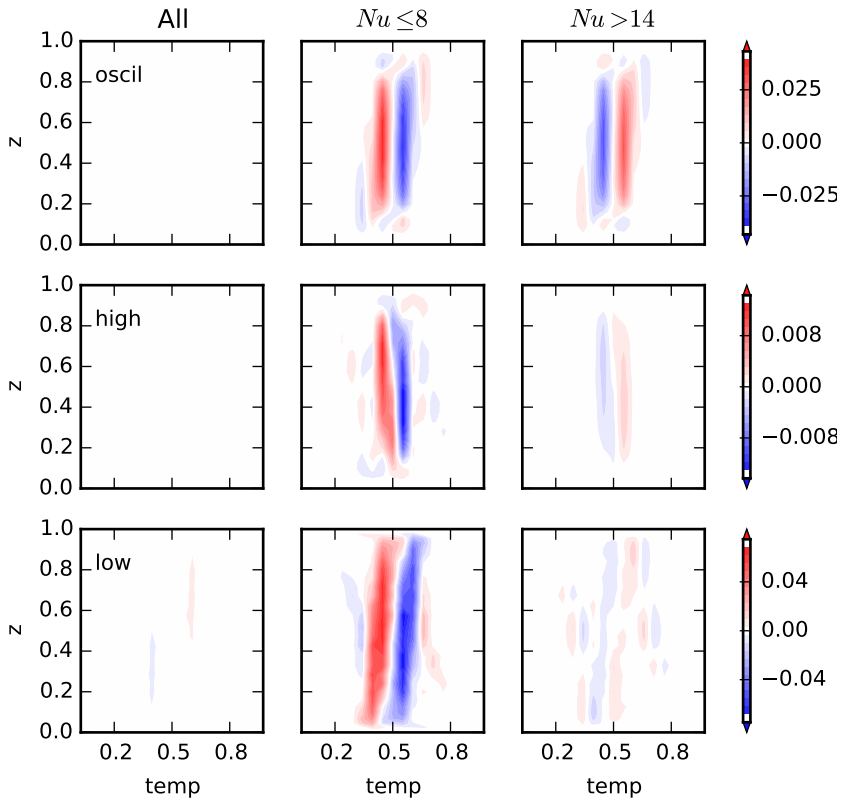


Figure 8: Composites of isothermally binned vertical velocity $w_T$. The colorscale is in non-dimensional velocity units. The (left) mean of the reconstructed mode family and the (middle) temporal mean for all samples with $Nu \leq 8$ and (right) $Nu > 14$ are plotted. For each mode family, the average across all time samples is small, which is a feature of the NLSA analysis assuming the dynamics are ergodic.

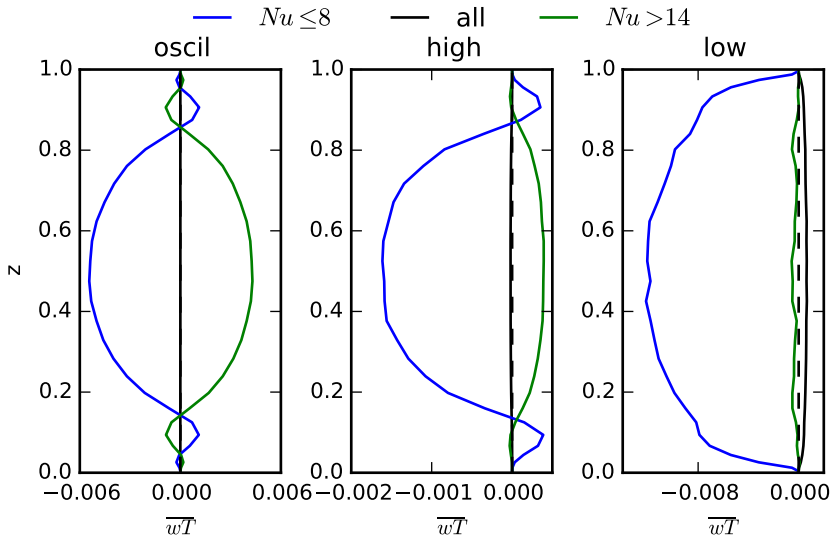


Figure 9: Convective heat flux averaged horizontally.

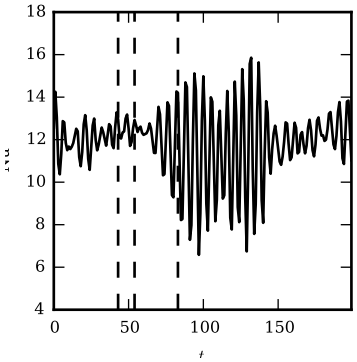

Figure 10: Portion of Nusselt number time series for the extreme heat transfer event for $t = 4950 - 5150$. The time axis has been shifted to begin at $t = 0$. The spatial reconstructions for the NLSA mode families are plotted in Figure 11 for the three time samples indicated by the vertical dashed lines.

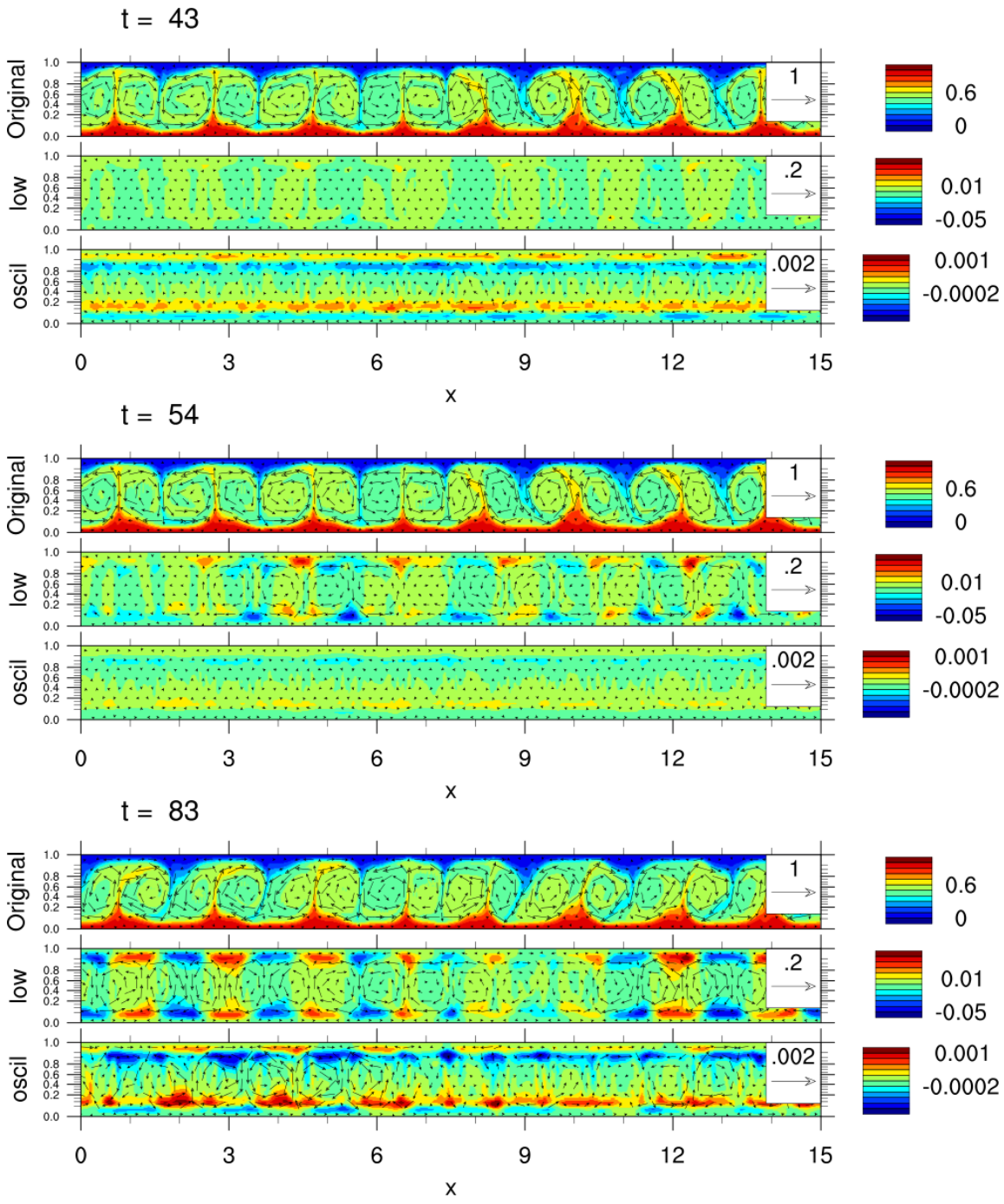


Figure 11: Snapshots of the spatially reconstructed velocity and temperature field for the time interval plotted in Figure 10. Three time samples are contained in the plot. The upper panel of each time sample plots the original data, and each subsequent row contains the spatial reconstructions for the indicated mode family. Temperature is indicated using contours and velocity with the arrows.

**Supplementary material for online publication only**

**Click here to download Supplementary material for online publication only: movie.mp4**

**LaTeX Source Files**

**Click here to download LaTeX Source Files: submission.tar.gz**