INFORMATION THEORY AND DYNAMICAL SYSTEM PREDICTABILITY

RICHARD KLEEMAN

Abstract. Predicting the future state of a turbulent dynamical system such as the atmosphere has been recognized for several decades to be an essentially statistical undertaking: Uncertainties from a variety of sources are magnified by dynamical mechanisms and given sufficient time, compromise any prediction. In the last decade or so this process of uncertainty evolution has been studied using a variety of tools from information theory. These provide both a conceptually general view of the problem as well as a way of probing its nonlinearity. Here we review these advances from both a theoretical and practical perspective. Connections with other theoretical areas such as statistical mechanics are emphasized however the importance of obtaining practical results for prediction also guides the development presented.

1. Introduction

Prediction within dynamical systems originated within the modern era in the study of the solar system. The regularity of such a system on time scales of centuries meant that very precise predictions of phenomena such as eclipses are possible at such lead times. On longer times scales of order million or more years, chaotic behavior due to the multi-body gravitational interaction becomes evident (see e.g. [1]) and effects such as planetary collisions or system ejection can occur. Naturally at such leads where chaotic behavior dominates predictions become far less precise. Of course, predictions over such a timescale are only theoretical and not subject to observational verification.

In the last century prediction within a greater range of practical dynamical systems has been attempted. Perhaps the best known of these have been turbulent fluids such as the atmosphere and ocean as well as earthquake prediction for which the system can be considered even more non-linear. It is fair to say that considerable progress has been made in the first area; the second area has potential currently limited by a lack of observations while the third area more limited success has been achieved (see e.g. [2]).

Predictions made using a dynamical model typically suffer primarily from two deficiencies[1]: Firstly the model used may have certain inadequacies as a representation of reality and secondly initial conditions for a prediction may not be known exactly. Such problems are known as *model errors* and *initial condition errors* respectively.

Progress can be made in improving predictions either by improving physical depictions within models or by improving the observational network and thereby reducing errors in the initial conditions. There is considerable evidence for the

[1]Uncertainty in boundary conditions can sometimes be of importance as well. We omit a discussion of this effect in the present review.

atmosphere however that no practical observational network will ever eliminate significant prediction errors due to initial condition errors. Even if one was able to define such conditions to the round-off accuracy of the computing device deployed, at some practical prediction time even these minute errors would grow sufficiently large as to overwhelm the forecast made. Such behavior is of course characteristic of dynamical systems classified loosely as chaotic or turbulent.

In general model errors are almost by definition quite hard to study in a systematic fashion since they are caused by quite diverse factors which are not very well understood. In fact if they *were* better understood they would be removed by improving the dynamical model using this knowledge. Thus the issue of model error tends primarily to be an engineering rather than a theoretical study. We therefore focus our attention here on initial condition errors but it should be borne in mind that model errors can often cause substantial problems for forecasters. It is interesting that the general public often confuses the two issues and attributes errors due to initial condition uncertainty to the inadequacy of meteorologists in reducing model errors. Of course meteorologists are not always averse to using the reverse strategy in response.

Inevitable deficiencies in observing networks imply uncertainty about the initial conditions used in predictions which can therefore be considered to be random variables. The study of the evolution of such variables thus in a general sense can be considered to be the study of predictability. Naturally functionals connected with uncertainty defined on such variables play an important role in such studies. In the context of atmospheric prediction, most attention has focused on the first two moments of the associated probability functions since often such functions are quasi-normal. Study of the problems from the viewpoint of entropic functionals is the natural generalization of this which allows for a more general treatment and this has received considerable attention in the last decade and is the subject of this review.

The dynamical systems of practical interest are generally of rather high dimensionality and this has posed particular problems for predictability studies. It means that study of the evolution of the multivariate probability distribution is generally impossible and studies are often confined to Monte Carlo (ensemble) predictions. The size of such ensembles are usually far less than the system dimensionality which has led to a wide variety of reduced state space techniques for identifying important error growth directions. Related reduced state space methods can also be used to truncate the dynamical system to a stochastically forced low order system. Such reductions have proved very useful in illuminating dominant dynamical processes and may also be amenable to complete analytical solution.

With regard to predictability studies, the functionals which have been studied most are differential entropy, relative entropy and mutual information. The dynamical evolutions of the first two are of particular interest with the second being defined with respect to the time evolving probability distribution and the equilibrium (climatological) probability distribution of the system. In general these two distributions converge with time and when they are statistically indistinguishable in some sense (to be made precise below) all predictability has been lost. Indeed an illuminating way to view predictability is as a measure of the disequilibrium of a statistical system. All this is discussed further in sections 2, 3 and 5 below.

The review is structured as follows: The next section is a brief review of the mostly well known properties of the relevant entropic functionals to be discussed. Section 3 discusses general entropic evolution equations in dynamical systems for which the generic Chapman Kolmogorov equation is relevant. Several of these results are not new but all of the material is not widely known outside statistical physics. Section 4 discusses the concept of information (uncertainty) flow within a dynamical system and its applications. Section 5 discusses various approaches to the study of predictability using the previous tools and outlines a number of new results obtained in the last decade. Finally Section 6 draws some general conclusions on work to date and potential future directions.

The approach taken in this review is slanted toward the authors and co-workers work in this area but other perspectives are also outlined and referenced. In general the nature of this work is rather eclectic given the attempt to balance the theoretical with the practical. It is hoped there will be materials here of interest to a rather broad audience. The intention is to connect information theoretic approaches to predictability with the broad area of statistical physics. A different approach to this subject which emphasizes more geophysical applications may be found in the excellent review of [3]. We have also chosen in the interests of brevity to omit the treatment of data assimilation and data filters from an information theoretic perspective but remind the reader that there is significant effort in that area as well (see, for example, [4] and [5]).

2. Relevant information theoretic functionals and their properties

Results quoted without proof may be found in [6]. Consider two vector random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ with associated countable outcome alphabet A and associated probability functions $p(x)$ and $q(x)$ with $x \in A$. The entropy or uncertainty is defined by

$$H(\boldsymbol{X}) \equiv \sum_{x \in A} p(x) \log \left(\frac{1}{p(x)} \right)$$

which is obviously non-negative. The relative entropy between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is defined by

$$D(p||q) \equiv \sum_{x \in A} p(x) \log \left(\frac{p(x)}{q(x)} \right)$$

This is also non-negative and only vanishes when $p = q$ so can be considered a "distance" function between functions although it is neither symmetric nor satisfies the triangle identity of Euclidean distance[2].

If we concatenate $\boldsymbol{X}$ and $\boldsymbol{Y}$ then we can define a joint probability function $p(x, y)$ with associated marginal distributions $p(x)$ and $p(y)$. The mutual information then measures the "distance" between this joint distribution and one for which $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent i.e. $p(x, y) = p(x)p(y)$. Thus we have

$$I(\boldsymbol{X}; \boldsymbol{Y}) \equiv D(p(x, y)||p(x)p(y)) = \sum_{x,y \in A} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right)$$

[2]In the limit that the two probability functions are "close" then indeed it does satisfy these identities to arbitrary accuracy. This is an indication that the relative entropy induces a metric structure in the sense of differential geometry on the space of probability densities. Further discussion can be found in [7].

Given its definition, the mutual information represents the degree of dependency between different random variables. Finally we can use the joint distribution $p(x, y)$ to define the conditional entropy $H(\boldsymbol{X}|\boldsymbol{Y})$ as the expected uncertainty in $\boldsymbol{X}$ given that $\boldsymbol{Y}$ is known precisely.

The entropic functionals defined above can be generalized to so called *differential* entropic functionals defined on random variables with continuous alphabets. It is interesting however to attempt to interpret them as limits of their countable analogs. Thus for example the differential entropy $h(\boldsymbol{X})$ is defined as:

$$h(\boldsymbol{X}) \equiv -\int_S p(x)\log(p(x))dx \tag{2.1}$$

where S is the continuous outcome set for $\boldsymbol{X}$. If we convert this to a Riemann sum we obtain

$$h(\boldsymbol{X}) \sim -\sum_{i\in\Lambda} p(x_i^*)\log(p(x_i^*))\Delta = -\sum_{i\in\Lambda} P_i \log P_i + \log\Delta = H(\tilde{\boldsymbol{X}}) + \log\Delta \tag{2.2}$$

where Δ is the (assumed constant) volume element for the Riemann sum partitioning Λ chosen. Clearly as this approaches zero, the second term approaches $-\infty$ and the differential entropy is finite only because $H(\tilde{\boldsymbol{X}})$ diverges to $+\infty$. This latter divergence occurs because the larger the size of the index/alphabet set Λ the larger the entropy since there is increased choice in outcomes. One can overcome this rather awkward limiting process by restricting attention to entropy differences of different random variables in which case the $\log\Delta$ term cancels.

By contrast the differential relative entropy is a straightforward limit of the ordinary relative entropy:

$$\begin{array}{ll} D(p||q) \equiv \int_S p(x)\log(p(x)/q(x))dx & \sim \sum_{i\in\Lambda} p(x_i^*)\log(p(x_i^*)/q(x_i^*))\Delta \\ & = -\sum_{i\in\Lambda} P_i \log(P_i/Q_i) = D(P||Q) \end{array} \tag{2.3}$$

Note the cancellation of Δ in the third expression here.

This cancellation effect is also important to the transformational properties of the differential functionals. In particular suppose we have the following general non-linear change of variables:

$$\mathbf{y} = \mathbf{r}(\mathbf{x})$$

The transformed probability density $p'(\mathbf{y})$ is well known to be given by

$$p'(\mathbf{y}) = p(\mathbf{r^{-1}}(\mathbf{y}))\left|\det\left\{\frac{\partial \mathbf{r^{-1}}(\mathbf{y})}{\partial \mathbf{y}}\right\}\right| \tag{2.4}$$

and the change of variables formula for integration has it that

$$d\mathbf{x} = \mathbf{dy}\left|\det\left\{\frac{\partial \mathbf{r^{-1}}(\mathbf{y})}{\partial \mathbf{y}}\right\}\right| \equiv d\mathbf{y}\,|\det J| \tag{2.5}$$

where J is the so-called Jacobian of the transformation. So we have

$$\begin{array}{ll} D(p'||q') = \int_{S'} p'(\mathbf{y})\log(p'(\mathbf{y})/q'(\mathbf{y}))d\mathbf{y} & = \int_S p(\mathbf{x})\log((p(\mathbf{x})\,|\det J|)\,/\,(q(\mathbf{x})\,|\det J|))d\mathbf{x} \\ & = D(p||q) \end{array}$$

providing the determinant of the transformation does not vanish which is a condition for the non-singularity of the transformation. Notice that this proof does not work for the differential entropy because of the lack of cancellation. The *difference* of the differential entropy of two random variables will be invariant under *affine*

transformations because then $\det J$ is constant and the integral of the probability density is also constant (unity). For an affine transformation of the form

$$\boldsymbol{x}' = A\boldsymbol{x} + \boldsymbol{c}$$

where the matrix A and vector $\boldsymbol{c}$ are constant one can easily establish in using the similar arguments to above that

$$h(A\boldsymbol{X}) = h(\mathbf{X}) + \log|\det A| \tag{2.6}$$

Notice that this equation also implies that new distributions that are given by

$$p'(\boldsymbol{x}) = p(\boldsymbol{x}')$$

have the same differential entropy providing that the affine transformation is volume preserving i.e. $\det A = \pm 1$. Note however that in general $D(p'||p)$ will be non-zero and positive unless the probability distribution has a symmetry under the transformation.

In addition to the above general transformational invariance and straightforward limiting property, the relative entropy also satisfies an intuitive fine graining relationship. Thus if one subdivides a particular partitioning Λ into a finer partitioning Λ' then one can easily establish (using the log sum inequality) that the relative entropy defined on the new partition is at least as large as the original coarse grained functional. Thus if the limit to the continuous functional is taken is this fine graining fashion then the relative entropy converges monotonically to the continuous limit. This has the intuitive interpretation that as finer and finer scales are considered, that "differences" in probability functions become larger as finer structure is used to make the assessment of this "difference".

3. Time evolution of entropic functionals

We begin our discussion within the context of a general stochastic process. A well known result from information theory called the generalized second law of thermodynamics states that under certain conditions the relative entropy is a non-increasing function of time. We formulate this in terms of causality.

Definition 1. Suppose we have two stochastic processes $\boldsymbol{X}(t)$ and $\boldsymbol{Y}(t)$ with $t \in R$ denoting time and associated probability functions p and q the same outcome sets $\{\boldsymbol{x}(t)\}$. We say that the two processes are *causally similar* if

$$p\left(\boldsymbol{x}(t)|\boldsymbol{x}(t-a)\right) = q\left(\boldsymbol{x}(t)|\boldsymbol{x}(t-a)\right) \quad \forall \boldsymbol{x} \quad \forall t \in R \quad \textit{and} \quad \forall a > 0 \tag{3.1}$$

This condition expresses intuitively the notion that the physical system giving rise to the two processes is identical *and* that precise knowledge of the set of values for outcomes at a particular time are sufficient to determine the future evolution of probabilities. This is intuitively the case for a closed physical system however not the case for an open system since in this second case other apparently external variables may influence the dynamical evolution. Note that the condition of probabilistic causality (3.1) introduces an arrow in time (via the restriction $a > 0$). Also note that a time-homogeneous Markov process satisfies this condition.

It is now easy to show that the relative entropy of two causally similar processes is non-increasing (the proof can be found in [6]). In some systems referred to as reversible (see below) this functional will actually be conserved while in others termed irreversible, it will exhibit a strict decline. Note that this result applies to *all* causally similar processes but often one is taken to be the equilibrium process and in

irreversible processes the relative entropy then measures the degree of equilibration of the system.

Consider now a continuous time continuous state Markov process governed by a Fokker Planck equation (FPE):

$$\partial_t p = -\sum_{i=1}^{N} \partial_i \left[A_i(\mathbf{x},t)p\right] + \frac{1}{2}\sum_{i,j=1}^{N} \partial_i\partial_j \left\{ \left[\mathbf{B}(\mathbf{x},t)\mathbf{B}^t(\mathbf{x},t)\right]_{ij} p \right\} \tag{3.2}$$

$$C_{ij} \equiv \left[\mathbf{B}(\mathbf{x},t)\mathbf{B}^t(\mathbf{x},t)\right]_{ij} \qquad non-negative\ definite \tag{3.3}$$

We can derive a series of interesting results regarding the time evolution of both differential entropy and relative entropy. Detailed proofs may be found in the Appendix. The first result is well known in statistical physics and dynamical systems studies (see, for example, [8] and [9]).

Theorem 2. *Suppose we have a stochastic process obeying equation (3.2) with* $\mathbf{B} = 0$

and associated probability function f *then the ordinary (differential) entropy satisfies the evolution equation*

$$h_t = \int f \nabla \cdot \mathbf{A} dx = \langle \nabla \cdot \mathbf{A} \rangle_f \tag{3.4}$$

Notice the importance of $\nabla \cdot \mathbf{A}$ to the entropy evolution. This also measures the rate at which an infinitesimal volume element expands or contracts in the dynamical system. When it vanishes the system is sometimes said to satisfy a Liouville condition. Hamiltonian systems which includes many inviscid (frictionless) fluids satisfy such a condition. We shall use equation (3.4) in a central way in the next section when we consider the concept of information flow.

A particular instructive case when $\nabla \cdot \mathbf{A} < 0$ occurs in the case of dissipative systems. A damped linear system has this quantity as a negative constant. In such a case the entropy declines since all trajectories end in the same fixed point. In the case that the system has some stochastic forcing and so the diffusion term in the Fokker Planck equation does not vanish then the stochastic forcing generally acts to increase entropy as the following extension shows:

Theorem 3. *Suppose we have a general stochastic process with probability function* p *governed by a Fokker Planck equation which is restricted by the stochastic forcing condition*

$$\partial_i C_{ij} = 0 \tag{3.5}$$

where we are assuming the summation convention. Then the evolution of the differential entropy is given by the equation

$$h_t = \langle \nabla \bullet \mathbf{A} \rangle_p + \frac{1}{2} \left\langle C_{ij} \partial_i \left(\log p\right) \partial_j \left(\log p\right) \right\rangle_p \tag{3.6}$$

The second term here clearly results in no decline in entropy given that C_{ij} *is non-negative definite.*

In the case of a stochastically forced dissipative system it is possible for the system to reach an equilibrium where the entropy creation due to the stochastic forcing is balanced by its destruction via the dissipative deterministic dynamics. This balance is an example of a fluctuation dissipation result. A particular case

of this theorem with C_{ij} constant was stated in [8] and the two terms on the right hand side of (3.6) were referred to as entropy production and entropy flux respectively. Also within this reference is a discussion of the relationship of this type of entropy evolution equation to others proposed in statistical physics using Lyupanov exponents (e.g. [10]).

In contrast to the evolution of differential entropy, the relative entropy is conserved in all systems even dissipative ones providing $\mathbf{B} = 0$:

Theorem 4. *Suppose we have two stochastic processes obeying equation (3.2) which have the additional condition that* $\mathbf{B}(\mathbf{x}, t) = 0$ *then the relative entropy of the two processes (if defined) is time invariant.*

In many dynamical systems with $\mathbf{B} = 0$ if one calculates the relative entropy with respect to a particular finite partitioning of state space rather than in the limit of infinitesimal partitioning then the conservation property no longer holds and in many interesting cases it declines with time instead and the system equilibrates. This reflects the fact that as time increases the difference in the distributions tends to occur on the unresolved scales which are not measured by the second relative entropy calculation. Such a coarse graining effect is of course the origin of irreversibility and often the "unresolved" scales are modeled via a stochastic forcing i.e. we set $\mathbf{B} \neq 0$. In such a case we get a strict decline with time of relative entropy. The following result was due originally to [11], the proof here follows [12]

Theorem 5. *Suppose we have two distinct*[3] *stochastic processes obeying (3.2) with* $C = \mathbf{B}(\mathbf{x}, t)\mathbf{B}^{\mathbf{t}}(\mathbf{x}, \mathbf{t})$ *positive definite almost everywhere and with associated probability densities* f *and* g *then the relative entropy strictly declines and satisfies the evolution equation*

$$D(f||g)_t = -\left\langle C_{ij}\partial_i \left(\log(f/g)\right) \partial_j \left(\log(f/g)\right)\right\rangle_f \tag{3.7}$$

This final theorem shows the central role of stochastic forcing in causing relative entropy to decline and hence to the modeling of irreversibility. In the context of statistical physics, the non-positive term on the right hand side of (3.7) with g set to the equilibrium distribution (invariant measure) forms a part of discussions on non-equilibrium entropy production (see [8]).

In stochastic modeling of dynamical systems it is common to separate the state space into fast and slow components and model the former with noise terms and dissipation of the slow modes. Presumably in this case if such a model works well as a coarse grained model for the total unforced and undissipated system then the last two theorems imply that there is a "leakage" of relative entropy from the slow to the fast components of the system.

It is an empirical fact that in many physical systems of interest, we find that if a slow subspace is chosen which represents a large fraction of the variability of the system then the subspace relative entropy will always show a monotonic decline which is suggestive that stochastic models of this system may work well. In the more general dynamical system context however it is an interesting question under what circumstances the fine scales can cause an increase in the relative entropy with respect to the coarse scales. After all the coarse grained system is not closed and so the monotonicity result quoted at the beginning of this section need not necessarily apply.

[3] In other words differing on a set of measure greater than zero.

It is possible to extend the Fokker Planck equation to include discontinuous jump processes and then this equation becomes the more general (differential) Chapman-Kolmogorov equation. The additional terms are often referred to (on their own) as the master equation. It is then possible by similar arguments to those given above to conclude that the jump processes result in an additional strict monotonic decline in relative entropy. The interested reader is referred to Chapter 3 of [12] for a sketch proof and more information and references.

There is also a well known connection between these results and the classical kinetic theory of Boltzmann which applies to dilute gases. In the latter case the probability distribution of molecules in the absence of collisions is controlled by an equation identical to the Fokker Planck equation with $\mathbf{B} = 0$. The molecules themselves are controlled by a Hamiltonian formulation which means that the probability equation can be shown to also satisfy $\nabla \bullet \mathbf{A} = 0$. Thus both the entropy and relative entropy are conserved. Collisions between molecules are modeled probabilistically using a hypothesis known as the Stosszahlansatz or molecular chaos hypothesis. This amounts to the insertion of master equation terms and so the resulting Boltzmann equation can be viewed as a Chapman Kolmogorov equation. The particular form of the master equation ensures that both relative entropy and differential entropy satisfy monotonic declines and increases respectively. This result is known as Boltzmann's H-theorem and is the traditional origin of irreversibility in statistical mechanics. More detail can be found in standard treatments of statistical mechanics.

4. Information Flow

4.1. **Theory.** If we partition a closed dynamical system then an interesting question arises regarding the evolution of uncertainty within the subsets chosen. How does it depend on the evolution of the other partitions? In general as we saw in the last section, entropy is not conserved globally unless the Liouville property holds, so the flow of uncertainty within a system does not usually follow the standard continuity equation satisfied by quantities such as energy, charge or mass. This concept of uncertainty flow has some important practical applications since it is clear that reduction of uncertainty in one partition member at a particular time may depend importantly on the reduction of uncertainty in another partition member at an earlier time. Clearly then understanding the flow of uncertainty is potentially important to optimizing predictions.

This issue was first addressed by [13] who studied the propagation of perturbations in simple non-linear dynamical systems using a "moving frame" or co-moving Lyupanov exponent defined by

$$\lambda(v; x_1, x_2) \equiv \lim_{t\to\infty} \frac{1}{t} \ln\left[\frac{\Delta(v, x_1, x_2, t)}{\Delta(v, x_1, x_2, 0)}\right]$$

where perturbation amplitudes Δ are defined using an L_2 norm on a moving interval:

$$\Delta(v, x_1, x_2, t) \equiv \left[\int_{x_1+vt}^{x_2+vt} |\delta\psi(x,t)|^2 \, dx\right]^{\frac{1}{2}}$$

with respect to small deviations in the system variables $\psi(x, t)$. Maximization of this with respect to the moving frame velocity v showed the preferred velocity of growing perturbations. Since regular Lyupanov exponents are often related to

entropy production it was natural to try to find an information theoretic counterpart for the co-moving exponents. This turned out empirically and in the systems studied, to be the time lagged mutual information (TLMI) of random variables $I(\boldsymbol{X}(t_1);\boldsymbol{Y}(t_2))$ where the random variables are located at different spatial locations. Associated with the TLMI is the natural velocity scale

$$v' \equiv \frac{d(\boldsymbol{X},\boldsymbol{Y})}{|t_2 - t_1|}$$

where d is the spatial distance between the random variables. The TLMI turned out to be maximized when this velocity matched that which maximized the co-moving Lyupanov exponent.

In addition to this match of physical propagation scales, mutual information has an appealing interpretation as the reduction in uncertainty of $\boldsymbol{X}(t_1)$ due to perfect knowledge of $\boldsymbol{Y}(t_2)$ i.e. roughly speaking, the contribution of uncertainty in the former due to uncertainty in the latter. This follows from the identity

$$I(\boldsymbol{X}(t_1);\boldsymbol{Y}(t_2)) = h(\boldsymbol{X}(t_1)) - h(\boldsymbol{X}(t_1)|\boldsymbol{Y}(t_2)) \tag{4.1}$$

This measure of information flow was further verified as physically plausible in more complex and realistic dynamical systems by [14]. It was however shown to give misleading results in certain pathological situations by [15]. In particular when both $\boldsymbol{X}$ and $\boldsymbol{Y}$ are subjected to a synchronized source of uncertainty then unphysical transfers are possibly indicated by the lagged mutual information. This is somewhat analogous to over interpreting a correlation as causative. Note that the mutual information reduces to a simple function of correlation in the case that the distributions are Gaussian.

[16] suggested a new information theoretic measure of flow which overcame the problem of lack of causality identified in the earlier study. The situation was considered where each spatial location was a Markov process of order q with time being discretized with an interval of Δt. Thus the probability function at any particular spatial point depends only on the values at this particular location for the previous q times. In such a situation there can, by construction, be no information flow between spatial points. He then tested the deviation from this null hypothesis using a conditional relative entropy functional. For an order $q = 1$ Markov process this transfer entropy is defined as

$$T(\boldsymbol{Y} \to \boldsymbol{X}, t) \equiv \int\int\int p(x(t+\Delta t), x(t), y(t)) \log \frac{p(x(t+\Delta t)|x(t), y(t))}{p(x(t+\Delta t)|x(t))} dx(t+\Delta t) dx(t) dy(t)$$

Conceptually this represents the expected change in the probability function at a particular spatial location due to perfect knowledge of a random variable at another location and earlier time beyond which that would result from the perfect knowledge of the first variable at the earlier time. This rather long winded description is perhaps better expressed by writing the transfer entropy (TE) in terms of conditional entropies:

$$T(\boldsymbol{Y} \to \boldsymbol{X}, t) = h(\boldsymbol{X}(t+\Delta t)|\boldsymbol{X}(t)) - h(\boldsymbol{X}(t+\Delta t)|\boldsymbol{X}(t), \boldsymbol{Y}(t)) \tag{4.2}$$

These results can be easily generalized to the case $q > 1$ with the additional penalty of further complexity of form.

An important aspect of the above two measures is their practical computability. The TLMI is a bivariate functional while the TE is a functional of order $q + 1$.

Computation of entropic functionals in a practical context generally requires Monte Carlo methods (in the predictability context, the so-called ensemble prediction) and often the choice of a coarse graining or "binning" with respect to random variable outcomes. The coarse graining is required in conjunction with Monte Carlo samples to estimate required probability densities. This as usual creates problems in defining functionals of order greater than perhaps five or so since there are then for most practical cases, insufficient ensemble members to reliably sample the required dimensions of the problem. This issue is commonly referred to as the "curse of dimensionality". The functionals mentioned above can sometimes avoid this problem because of their low order.

A different approach to this problem was taken in [17] who took as their starting point the basic entropy evolution equation (3.4) in the absence of stochastic forcing. To simplify things they considered a two random variable system. In this case it is possible to calculate the entropy evolution equation for one of the random variables alone by integrating out the other variable in the (Liouville) evolution equation:

$$\frac{dH_1}{dt} = -\iint p(x_1, x_2)\left[\frac{A_1}{p(x_1)}\frac{\partial p(x_1)}{\partial x_1}\right] dx_1 dx_2$$

where A_i are the deterministic time tendency components (see equations (3.2) and (3.4)). Now if the random variable X_1 was not influenced by the random variable X_2 then we would expect the entropy of this first variable to evolve according to the one dimensional version of (3.4) i.e.

$$\frac{\partial H_1^*}{\partial t} = \left\langle \frac{\partial A_1}{\partial x_1} \right\rangle_p .$$

The difference between the actual X_1 entropy H_1 and the idealized isolated entropy H_1^* can be interpreted as the flow of uncertainty from X_2 to X_1 or expressed differently as the "information flow"

$$T_{2\to 1} \equiv H_1 - H_1^*.$$

In a number of simple dynamical systems this information flow was computed and was found to behave qualitatively but not quantitatively like the transfer entropy of Schreiber discussed above. In particular it was observed that in general flow is not symmetric in both cases i.e.

$$T_{2\to 1} \neq T_{1\to 2}$$

One would, of course, expect symmetry in the case of the flow of a conserved quantity which is an indication of the peculiarity of uncertainty/information flow.

The approach above was generalized to the case of a finite number of random variables in the two papers ([18]) and ([19]). Much work remains to be done in exploring the implications of this approach in dynamical systems of practical interest. A different generalization of the formalism was proposed in ([20]) who considered the information flow between two finite dimensional subspaces of random variables. Denoting the two outcome vectors by x and y, these authors considered the (Ito) stochastic system

$$\begin{aligned} dx &= (F_1(x) + F_{12}(x, y))dt \\ dy &= F_2(x, y)dt + BdW \end{aligned} \tag{4.3}$$

where B is a constant matrix governing the additive stochastic forcing (see (3.2)). The application envisaged by this system is where the x are slow coarse grained variables while the y are fast fine grained variables. The information flow $T_{y\to x}$ then plays the role of the second entropy production term of equation (3.6) since it represents uncertainty generation in the coarse grained variables due to the interaction with fine grained variables. Associated with this system is evidently a full Fokker Planck rather than the simpler Liouville equation however the general idea of ([17]) of flow between y and x still goes through in a similar manner. In the important case that the distribution for x is exponential

$$p_1(x,t) = \exp(-\phi(x,t))$$

it is easily derived that the information flow has the simple form

$$T_{y\to x} = -\langle \nabla_x \cdot F_{12}\rangle_p + \langle F_{12}\cdot \nabla_x \phi\rangle_p$$

and in the case that ϕ and F_{12} are polynomials in x this reduces to a matrix equation in moments of the appropriate exponential distribution family. Such distributions are, of course, widely seen in equilibrium statistical mechanics (Gibbs ensembles) or in quasi-equilibrium formulations of non-equilibrium statistical mechanics ([21]).

In the case that the F_1, F_{12} and F_2 satisfy certain natural divergence free conditions common in fluid systems, the authors are able to derive an interesting relation between the relative entropy rate of change and the information and energy flows within the system:

$$D(p_1||p_{1eq})_t = -T_{y\to x} + \frac{1}{\sigma^2}E_{y\to x}$$

where we are assuming B in (4.3) is diagonal with the white noise forcing components having constant variance σ^2 and p_{1eq} is the coarse grained equilibrium distribution. This equation is to be compared with (3.7) where the direct stochastic representation of the fine grained variables causes the relative entropy to decline.

4.2. **Applications.** One important practical application of the preceding formal development occurs in the problem of forecasting within dynamical systems. In high dimensional systems, errors in the initial conditions inevitably occur because the observing system typically is only able to partially resolve state space. More specifically for geophysical prediction, the high dimensionality is caused by the requirement of adequately resolving the spatial domain of interest and any observing network used to define initial conditions is typically not comprehensive with regard to model resolution.

These errors propagate and magnify as the prediction time increases (see next section). One approach to improving predictions is therefore obviously to reduce initial condition errors. Unfortunately however the very large improvements in observing platforms required to achieve this for all initial conditions are very expensive. Frequently however predictions in specific areas are regarded as having a high utility (for example, a storm forecast over a large city) and the initial conditions errors affecting such specific predictions are believed to be confined to very particular geographical locations. Given such a scenario, a targeted observation platform improvement strategy may be a very effective and inexpensive method of improving important predictions.

For the above strategy to work however the sensitivity of prediction errors to initial condition errors must be well known. In atmospheric prediction, common methods of identifying this sensitivity are via linearization of the atmospheric dynamical system (see eg. [22]) or by the use of a Kalman filter (see eg. [23]). Such methods however make rather restrictive assumptions namely that linear dynamics are appropriate or that prediction distributions are Gaussian. These are questionable particularly for long range, low skill predictions since the dynamical systems involved are highly non-linear and the typical states therefore turbulent.

The measures of uncertainty/information flow discussed above make none of these restrictive assumptions so are obvious candidates for improving this sensitivity identification. This can be seen directly in the simplest measure, the TLMI, by consideration of equation (4.1) with t_1 set to the prediction time and t_2 to the initial time. The TLMI then tells us how much the prediction random variable $X(t_1)$ would have its uncertainty reduced if we could reduce the uncertainty of the initial condition random variable $Y(t_2)$ to zero. This is precisely the sensitivity we require for targeted initial condition improvement. The fact that the TLMI is a bivariate functional also means that Monte Carlo techniques are sufficient for its calculation.

The feasibility of using this functional in a practical setting was investigated in the atmospheric case by [24]. The author used a fairly realistic global model of the atmosphere. Initial conditions for the model were drawn as sample members from a simple multivariate Gaussian distribution intended to represent errors due to observing network sparseness. Each sample/ensemble member was integrated for around a week using the dynamical model and the TLMI calculated for a prediction at a specific location and dynamical variable of interest. In general the functional tended (in mid-latitudes at least) to peak strongly in small regions to the west of the prediction location. This geographically confined sensitivity lasted for the duration of the ensemble prediction. Two examples at 6 days are shown in Figure 4.1a and 4.1b. Calculated here are the TLMI with prediction and initial condition variables both the same and respectively temperature and zonal wind. Both random variables are at a near surface location as there is little vertical propagation of uncertainty. The sensitivity is well confined geographically and it is worth observing that by 6 days an ensemble prediction for the regions displayed typically exhibits marked non-linear effects. The TE of order 1 between the prediction time and the initial conditions was also calculated and found to give similar results except in the vicinity of the location of the particular prediction variable under consideration. Consideration of the right hand side of equation (4.2) shows this result to not be surprising.

The application of TE to non-linear time series such as those useful in biological situations has been considered by [15] who studied the interaction of heart and breathing rates. He found that unlike the TLMI, the TE exhibited an asymmetry with the transfer entropy from heart to breathing being generally greater than that in the reverse direction. This suggested the dominance causally of the heart which accords with basic biological understanding. Their results were however somewhat sensitive to the coarse graining chosen to evaluate the various probability densities. This sensitivity was not apparent qualitatively in the predictability applications discussed earlier given a sufficiently large ensemble.

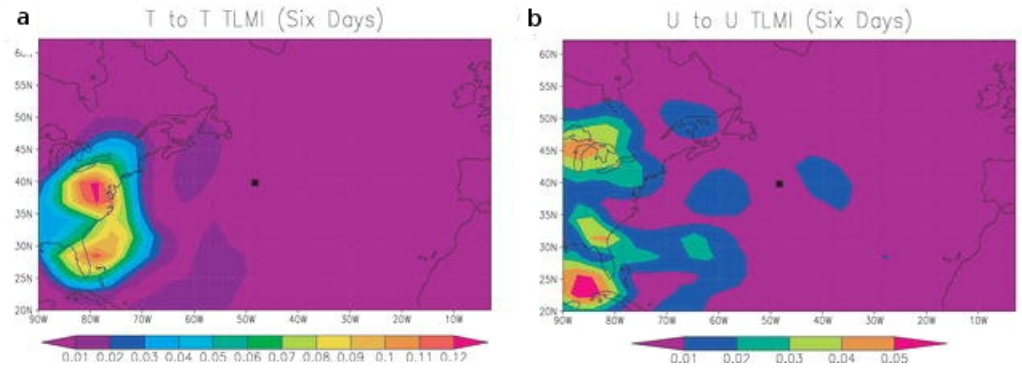

FIGURE 4.1. The TLMI regions for a six day prediction at the center of the diagram (shown as a solid circle). The left panel shows near surface temperature while the right shows zonal wind. Predictions were for the same quantities.

5. Predictability

5.1. **Introduction.** In seeking to make predictions in any dynamical system one always needs to deal with uncertainty. This is particularly so if such a system is large and complex. In general predictions within such systems are usually made by solving an initial value problem using a set of partial differential equations. Uncertainty is introduced since the initial value vector can never be precisely defined and indeed observational limitations mean that this error can be a small but appreciable fraction of the desired accuracy of any practical prediction.

As mentioned earlier, when predictions of the real world are attempted one also faces an uncertainty associated with its representation by a set of differential equations. In general, this second kind of uncertainty which is often called model error, can only be measured rather indirectly via the careful comparison of model predictions to reality and an attempted disentanglement of the two kinds of error. This is a complex undertaking and so here we restrict our attention only to errors of the first kind and deal with what is sometimes referred to perfect model predictions.

We also restrict our focus to dynamical systems associated with turbulent fluids such as the atmosphere and ocean. Here routine predictions are common and in many situations demonstrably skillful. Additionally the degradation of this skill with prediction time is particularly evident. Predictions within the solar system are generally much more skillful but the degradation of this skill is not as pronounced on practical time scales (see e.g. [1]). As we shall argue below, the degradation time scale is closely related to the statistical equilibration time of the dynamical system. For the atmosphere this is of the order of a month or so. In the ocean it can be two to three orders of magnitude longer than this. In the solar system it is several orders of magnitude longer still[4] and so cannot be observed.

Given the uncertainty in the initial value vector, a powerful perspective regarding prediction is to regard it fundamentally as a statistical rather than deterministic

[4]Lyupanov exponents of order $10^{-6} - 10^{-7}$ $(years)^{-1}$ have been noted in realistic model integrations which sets an analogous timescale for loss of predictability in the solar system (see [1]).

undertaking. One is interested not just in the prediction itself but also in the associated uncertainty. Thus from a mathematical viewpoint, prediction is best considered as the study of the temporal evolution of probability distributions associated with variables of practical interest. Such a probability distribution will be referred to here as the prediction distribution.

Statistical predictions within such systems are often not considered in isolation but are evaluated against some reference distribution associated with the known long term behavior of the system. This might be considered from a Bayesian perspective as representing a prior distribution of the system since it gives the statistical knowledge of the system prior to any attempted prediction. This reference or prior distribution will be referred to here as the equilibrium distribution since under an ergodic hypothesis, the longer term statistical behavior of the system can be considered to also be the equilibrium behavior. In Bayesian terminology the prediction distribution is referred to as the posterior since it is a revised description of the desired prediction variables once initial condition knowledge is taken into account.

Different aspects of the prediction and equilibrium distribution figure in the assessment of prediction skill. Indeed a large variety of different statistical functionals has been developed which are tailored to specific needs in assessing such predictions. It is, of course, a subjective decision as to what is and is not important in such an evaluation and naturally the choice made reflects which particular practical application is required. An excellent discussion of all these statistical issues is to be found in [25] which is written from the practical perspective.

Given that statistical prediction is essentially concerned with the evolution of probability distributions and the assessment of such predictions is made using measures of uncertainty, it is natural to attempt to use information theoretic functionals in an attempt to develop greater theoretical understanding. In some sense such an approach seeks to avoid the multitude of statistical measures discussed above and instead focus on a limited set of measures known to have universal importance in the study of dynamical systems from a statistical viewpoint. The particular functionals chosen to do this often reflect the individual perspective of the theoreticians and it is important to carefully understand their significance and differences.

5.2. **Proposed functionals.** To facilitate the discussion we fix notation. Suppose we make a statistical prediction in a particular dynamical system. Associated with every such prediction is a time dependent random variable $\boldsymbol{Y}(t)$ with $t \geq 0$ and this variable has a probability density $p(\boldsymbol{y}, t)$ where the state vector for the dynamical system is $\boldsymbol{y}$. Naturally there is also a large collection of such statistical predictions possible each with their own particular initial condition random variable. We denote the random variable associated with the total set of predictions at a particular time by $\boldsymbol{X}(t)$ with associated density $q(\boldsymbol{y}, x, t)$ where x is some labeling of the different predictions. For consistency we have

$$p_x(\boldsymbol{y}, t) = q(\boldsymbol{y}, x, t). \tag{5.1}$$

Note that the particular manner in which a collection of statistical predictions are assembled remains to be specified. In addition another important task is to specify what the initial condition distributions are exactly. Clearly that issue is connected with an assessment of the observational network used to define initial conditions.

To the knowledge of this reviewer, the first authors to discuss information theoretic measures of predictability in the atmospheric or oceanic context were [26] who considered a particular mutual information functional of the random variables $\boldsymbol{X}(0)$ and $\boldsymbol{X}(t)$. More specifically, they considered a set of statistical predictions with deterministic initial conditions which had the equilibrium distribution. This is a reasonable assumption because it ensures a representative sample of initial conditions. The framework was considered in the specific context of a stochastic climate model introduced earlier by the authors. In a stochastic model deterministic initial conditions are meant to reflect a situation where errors in the unresolved fine grained variables of the system are the most important source of initial condition errors. In the model considered, the equilibrium distribution is time invariant which implies that this is also the distribution of $\boldsymbol{X}(t)$ for all t. As we saw in the previous section we have the relation

$$I(\boldsymbol{X}(t);\boldsymbol{X}(0)) = H(\boldsymbol{X}(t)) - H(\boldsymbol{X}(t)|\boldsymbol{X}(0)). \tag{5.2}$$

so the functional proposed measures the expected reduction in the uncertainty of a random variable given perfect knowledge of it at the initial time. This represents therefore a reasonable assessment of the expected significance of initial conditions to any prediction. As $t \to \infty$ the second conditional entropy approaches the first unconditioned entropy since knowledge of the initial conditions does not affect uncertainty of $\boldsymbol{X}(t)$. Thus as the system equilibrates this particular functional approaches zero. Note that it represents an average over a representative set of initial conditions so does not refer to any one statistical prediction. We shall refer to it in this review as the generic utility[5].

A few years later a somewhat different approach was suggested by [27] who rather than considering a representative set of statistical predictions, considered instead just one particular statistical prediction. They then defined a predictability measure which was the difference of the differential entropy of the prediction and equilibrium distributions:

$$R(t) = H(\boldsymbol{Y}(\infty)) - H(\boldsymbol{Y}(t)) \tag{5.3}$$

which they call the predictive information[6]. This has the straightforward interpretation as representing the reduction in uncertainty of a prediction variable $\boldsymbol{Y}(t)$ compared to the equilibrium or prior variable $\boldsymbol{Y}(\infty)$. This appears to be similar to equation (5.2) conceptually since if the former author's framework is considered, then $H(\boldsymbol{Y}(\infty)) = H(\boldsymbol{X}(t))$ but remember that (5.3) refers to just one particular statistical prediction. In the special case that both the prediction and equilibrium distributions are Gaussian one can derive the equation[7]

$$R(t) = -\frac{1}{2}\log\left(\det\left(C(t)C^{-1}(\infty)\right)\right) \tag{5.4}$$

[5]The authors referred to it as the transinformation. We adopt the different terminology in order to contrast it with other proposed functionals (see below).

[6]We shall refer to this functional henceforth with this terminology

[7]We are using the conventional definition of entropy here whereas [27] use one which differs by a factor of m, the state-space dimension. Shannon's axioms only define entropy up to an unfixed multiplicative factor. The result shown here differs therefore from those of the original paper by this factor. This is done to facilitate clarity within the context of a review.

where the matrix $C(t)$ is the covariance matrix for the Gaussian distribution at time t. The matrix $C(t)C^{-1}(t=\infty)$ is easily seen to be non-negative definite assuming reasonable behavior by the equilibrium covariance. The eigenvectors of this matrix ordered by their non-negative eigenvalues (from greatest to least) then contribute in order to the total predictive information. Thus these vectors are referred to as predictability patterns with the first few such patterns often dominating the total predictive information in practical applications. In a loose sense then these are the predictability analogs of the principal components whose eigenvalues explain variance contributions rather than predictive information. Note that the principal components are the eigenvectors of $C(\infty)$.

Another approach to the single statistical prediction case was put forward by the present author in [28] (see also [29]). They proposed that rather than considering the difference of uncertainty of the prior and posterior distributions that the total discrepancy of these distributions be measured using the relative entropy functional. The motivation for this comes from Bayesian learning theory (see [30] Chapter 2). Here a prior is replaced after further observational evidence by a posterior and the discrepancy between the two, as measured by the relative entropy, is taken as the utility of the learning experiment. Information theory shows that the relative entropy is the coding error made in assuming that the prior holds when instead the posterior describes the random variable of interest. Given this background, the relative entropy can be regarded as a natural measure of the utility of a perfect prediction. We shall refer to this functional as the prediction utility.

These two functionals differ conceptually since one uses the reduction in uncertainty as a measure while the other measures the overall distribution change between the prior and posterior. This difference can be seen abstractly as follows: Consider a translation and a rotation of a distribution in state-space then as we saw in section 2, the difference of the entropy of the original and transformed distributions is zero but the relative entropy of the two distributions is generally positive. These issues can be seen explicitly by calculating the relative entropy for the Gaussian case and comparing it to (5.4):

(5.5)
$$\begin{aligned} D(p(t)||p(\infty)) &= -\tfrac{1}{2}\log\left(\det\left(C(t)C^{-1}(\infty)\right)\right) + \tfrac{1}{2}tr\left(\left[C(t)-C(\infty)\right]C^{-1}(\infty)\right) \\ &+ \qquad\qquad \tfrac{1}{2}\overline{\mathbf{y}}(t)^t C^{-1}(\infty)\overline{\mathbf{y}}(t) \end{aligned}$$

where we are assuming for clarity that the mean vector $\overline{\mathbf{y}}(\infty)=0$. The first term on the right here is identical to (5.4) so the difference in the measures comes from the second and third terms. The third non-negative definite term arises from the distributions having different means which is analogous to the translation effect discussed above. The second term is analogous to the rotation effect discussed above: Suppose hypothetically that the equilibrium Gaussian distribution is in two dimensions with a diagonal covariance matrix but with variances in the two directions unequal. Consider hypothetically that the prediction distribution is a rotation of this distribution through $\pi/2$. Evidently the relative entropy of the two distributions will not occur through the first or third term in (5.5) since the mean vector is unchanged as is the entropy. Thus the second term must be the only contributor to the positive relative entropy.

A very simple concrete example illustrates the differences discussed. Suppose that temperature at a location is being predicted. Further suppose that (realistically) the historical and prediction distributions for temperature are approximately

normal. Next suppose that the historical mean temperature is 20°C with a standard deviation of $5°C$. Finally suppose that the statistical prediction for a week in advance is $30°C$ with a standard deviation of $5°C$. It is clear that the first measure discussed above would be zero whereas the second would be 2.0 and due only to the third term in (5.5). Clearly no reduction in uncertainty has occurred as a result of the prediction but the utility is evident. Obviously this example is somewhat unrealistic in that one would usually expect the prediction standard deviation to be less than $5°C$ but nevertheless the utility of a shift in mean is clear and is intuitively of a different type to that of reduced uncertainty. Prediction utility of this kind in a practical context is partially measured by the widely used anomaly correlation skill score (see [31]).

The transformational properties for the three functionals above were discussed in section 2: The relative entropy is invariant under a general non-degenerate non-linear transformation of state space. The mutual information, since it is the relative entropy of two particular distributions, can also be shown to have this property. The difference of differential entropies is invariant under the smaller class of affine transformations[8].

There is an interesting connection between the three proposed functionals that was first observed by ([3]): Suppose we take the general stochastic framework assumed above (not just the specific model but any stochastic model) and calculate the second and third functionals and form their expectation with respect to the set of statistical predictions chosen. It is easy to show then that they both result in the first functional. The proof is found in the appendix. Thus for the particular case of a stochastic system with deterministic initial conditions, the predictability functionals averaged over a representative set of predictions both give the same answer as the functional originally suggested by [26].

Finally it is conceptually interesting to view predictability as a measure of the statistical disequilibrium of a dynamical system. Asymptotically as the prediction distribution approaches the equilibrium distribution (often called the climatology) all measures of predictability proposed above approach zero. The initial conditions become less and less important and the prior distribution is approached meaning that the act of prediction has resulted in no new knowledge regarding the random variables of interest in the system. As we saw in section 3, in a variety of important stochastic dynamical systems, this disequilibrium is measured conveniently by the relative entropy functional since it is non-negative and exhibits a strict monotonic decline with time. These properties thus allow it to be used to discuss the general convergence properties of stochastic systems as discussed for example in [12] section 3.7. As we also saw in section 3 if the system is fully resolved i.e. all fine grained variables are retained, then the relative entropy is conserved. Thus it is fundamentally the coarse graining of the system that is responsible for the irreversible equilibration and hence loss of predictability. In practical fluid dynamical systems this occurs in two different ways: Firstly it is obvious that all spatial scales cannot be resolved in any model and the effects of the fine scales on the resolved scales need to be included in some way. This is often done using dissipative and stochastic formulations. Secondly in the sense of probability densities, even the retained scales can usually never be completely resolved. The corresponding Fokker

[8]This does not cover all invariant transformations for this case since there are non-affine transformations with $\det J$ constant.

Planck equation is usually not exactly solvable in non-linear systems and numerical simulations face severe practical issues in high dimensions (the well known curse of dimensionality). All of this means that in nearly all realistic cases, Monte Carlo methods are the only practical choice and these mean an effective coarse graining since there is a limit to the scale at which the probability density is observable using an ensemble. We discuss these issues further below.

5.3. **Applications.** A number of dynamical models of increasing complexity have been studied using the functionals discussed in the previous subsection. Some of these have direct physical relevance but others are included to illustrate various dynamical effects.

In considering predictability of a dynamical system one is usually interested from a practical perspective in two things:

(1) In a generic sense how does predictability decay in time?
(2) How does predictability vary from one prediction to the next?

Phrased more abstractly, what is the generic character of the equilibration and secondly how does the predictability vary with respect to the initial condition label x in equation (5.1)?

These issues have been studied from the perspective of information theory in the following types of models.

5.3.1. *Multivariate Ornstein Uhlenbeck (OU) stochastic processes.* Models of this type have seen application as simplified representations of general climate variability (e.g. [32]); of the El Nino climate oscillation (e.g. [33], [34] and [35]); of mid-latitude decadal climate variability (e.g. [36]) and of mid-latitude atmospheric turbulence (e.g. [37]). In the first three cases there is a very natural (and large) timescale separation between internal atmospheric turbulent motions and variability that is usually thought of as climatic which inherits a slow timescale from the ocean. In the fourth case the separation is associated with the various parts of the atmospheric turbulence spectrum. In general the first kind of separation is more natural and of a greater magnitude. Such strong separations are of course the physical justification for the various stochastic models. Models of this type have been used to carry out skillful real time predictions most particularly in the El Nino case but also in the mid-latitude atmospheric case.

A general time dependent OU stochastic process can be shown to have the property that Gaussian initial condition random variables remain Gaussian as time evolves. This follows from the fact that Gaussian random variables subjected to a linear transformation remain Gaussian and the general solution for these processes (see equation (4.4.83) from [12]) is a linear combination of the initial condition random variables and Gaussian Wiener random variables. In the time independent case (the usual climate model) calculation of the covariance matrix (equation (4.4.45) from [12]) shows that it depends only on the initial condition covariance matrix and the prediction time. This implies that different initial conditions with identical covariance matrices have identical covariance matrices at other times also. Consideration of equation (5.4) and (5.5) then shows that such a set of statistical predictions have identical predictive information but may have varying prediction utility depending on the initial condition mean vector. Additionally if the initial conditions have zero covariance i.e. are deterministic and are sampled in a representative fashion from the equilibrium distribution then the results of 5.2 show that

this initial condition invariant predictive information is just the generic utility of the set of predictions. In what follows we shall refer to the third term from (5.5) as the Signal since it depends on the first moment of the prediction distribution while the first two terms will be referred to as the Dispersion since they depend on the covariance of the same distribution. Notice that if the signal is averaged over a representative set of deterministic initial conditions then it simply becomes minus the second term because of the discussion at the end of the last subsection. Thus minus the second term is a measure of the average contribution of the signal to prediction utility.

A simple example, considered previously by the present author as a minimal model of El Nino (see [28] and [35]), serves to illustrate the kinds of behavior encountered. Here a stochastically forced damped oscillator is modeled with two variables. In Ito form

$$\begin{aligned} d\boldsymbol{x} &= -A\boldsymbol{x}dt + \boldsymbol{u}dW \\ A &\equiv \begin{pmatrix} 0 & 1 \\ \beta & \gamma \end{pmatrix} \quad \boldsymbol{u} = \begin{pmatrix} 0 \\ F \end{pmatrix} \\ \gamma &\equiv \frac{2}{\tau} \quad \beta \equiv \left(\frac{4\pi^2}{T^2} + \frac{1}{\tau^2} \right) \end{aligned}$$

where τ is the oscillation decay time and T is the period. Asymptotically the covariance matrix of this system can be shown to be diagonal so the basis we are using is a principal component basis. Figure 5.1 shows the prediction utility of a representative sample of predictions with deterministic initial conditions with time being the vertical coordinate and sample member the horizontal coordinate. The relaxation process is evident for all predictions and one can show analytically that it is controlled by τ. Additionally it is clear that variations in prediction utility are a significant fraction of overall utility and that the character of this variation tends to persist strongly throughout any particular prediction. In other words predictions tend to have high/low utility at all times. We refer to this effect as predictability durability. Recall that the variation in utility is driven entirely in this example by the signal. Calculation of this term shows it to be

$$Signal = \left(\frac{\overline{x}_1}{\sigma_1^2} \right)^2 + \left(\frac{\overline{x}_2}{\sigma_2^2} \right)^2$$

where the equilibrium variances are σ_i^2. Thus the signal is simply the rescaled L_2 norm squared of the prediction means (called anomalies in the climate context). In this model it appears that it takes a significant amount of time for the stochastic forcing to erode this signal which is the cause of the durability.

The simple model above was generalized in the first reference above to the time dependent case on the basis that one would expect the instability of the system to follow certain cyclic features such as the seasons. The basic qualitative results shown above however remain unaltered and in particular the variation of prediction utility was still controlled mainly by the signal even when rather large variations in the decay time τ throughout the time cycle were assumed.

5.3.2. *Low order chaotic systems.* Such systems were originally proposed as very simple analogs for atmospheric turbulence (e.g [38] and [39]). They have been extensively studied from the viewpoint of atmospheric predictability and were the original motivation for the now extensive mathematical field of chaotic dynamics.

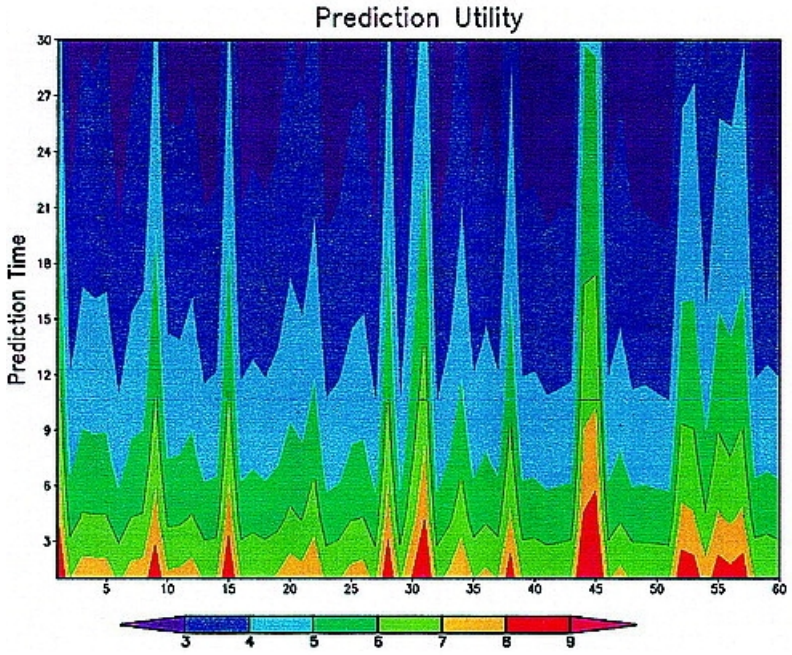

FIGURE 5.1. Prediction utility (relative entropy) for a series of representative initial conditions and at different prediction times.

Here, unlike the previous case, there is no great separation of timescales and the growth of small perturbations occurs through non-linear interaction of the degrees of freedom rather than from random forcing from fast neglected modes. These systems are often characterized by an invariant measure that is fractal in character. This implies that the dimension of the space used to calculate probability densities is non-integral which means some care needs to be taken in defining appropriate entropic functionals. As noted earlier the non-linearity of the system but also the fractal equilibrium strange attractor mean that generally the best practical approach is Monte Carlo sampling with an appropriate coarse graining of state space. One can for example define the entropy of the equilibrium strange attractor as

$$\begin{aligned} E &= \lim_{M \dashrightarrow \infty} \lim_{r \dashrightarrow 0} E(M,r) \\ E(M,r) &\equiv -\frac{1}{M} \sum_{i=1}^{M} \ln\left(\frac{N_i(r)}{M r^d}\right) \\ &= -\frac{1}{M} \sum_{i=1}^{M} \ln\left(\frac{N_i(r)}{M}\right) + d \ln r \equiv S(M,r) + d \ln r \end{aligned} \tag{5.6}$$

where $N_i(r)$ is the number of attractor sample members within a radius r of the i'th sample member; d is the so-called fractional information dimension of the attractor and M is the number of sample points. Careful comparison with (2.1) and (2.2) shows this agrees with the usual definition up to a constant which multiplies

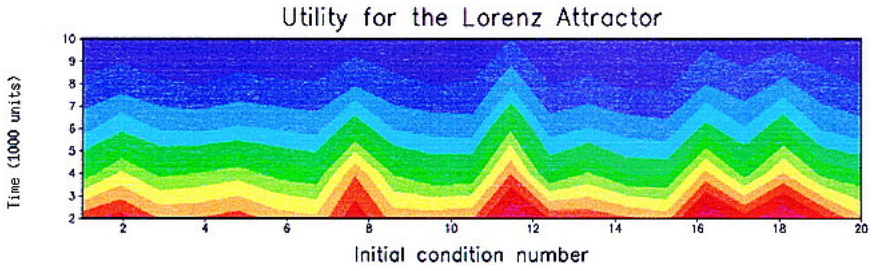

FIGURE 5.2. Same as Figure 5.1 but for the Lorenz model. Time is expressed in number of time steps (each .001).

r^d to give the volume element in a d dimensional space. The information dimension can be estimated by considering a sequence of decreasing r and calculating $S(M,r)$ (see [40] section 7.9). If one considers the calculation of the relative entropy of samples drawn from particular subregions of the attractor then a simpler expression is relevant which does not involve the information dimension:

$$D = \lim_{M \dashrightarrow \infty} \lim_{r \dashrightarrow 0} D(M,r)$$

$$D(M,r) \equiv \frac{1}{M} \sum_{i=1}^{M} \ln \left(\frac{N_i^1(r)}{N_i^2(r)} \right) \tag{5.7}$$

where $N_i^1(r)$ is the number of the first sample members within r of the i'th first sample member while $N_i^2(r)$ is the number of second sample members within the same radius of the same i'th first sample member. Naturally only a particular coarse graining $D(M,r)$ is usually practically available since M is finite.

The present author (see [28]) considered the behavior of coarse grained prediction utility (5.7) in the classical Lorenz chaotic system ([38]). Here we also compare that with the corresponding coarse grained predictive information derived from (5.6). A representative set of initial conditions with means drawn from equilibrium strange attractor were considered. For simplicity a (homogeneous) Gaussian distribution of two dimensions tangential to the attractor was considered with a standard deviation approximately three orders of magnitude smaller than the attractor dimensions. Thus the initial condition uncertainty was far less then the equilibrium or climatological uncertainty. The evolution of the relative entropy is shown in Figure 5.2 for a representative set of initial conditions (format is the same as Figure 5.1). We observe again a characteristic relaxation toward equilibrium. The temporal decline in relative entropy is almost always monotonic although as noted in the previous subsection, this is not compulsory for the coarse grained measure used.

In stark contrast to the stochastic models considered previously, variations in the prediction utility from one set of initial conditions to another is typically strongly related to variations in predictive information. This is shown for a large sample of initial conditions in Figure 5.3a which shows the relationship at around halfway through the relaxation displayed in Figure 5.2. It is worth noting that the prediction distributions although initially chosen to be Gaussian rapidly lose this property and indeed by the time of Figure 5.3a it can be shown that although the variations in the predictive information correlates well with the prediction utility, variations in the Gaussian predictive information of (5.4) show very little correlation. This suggests

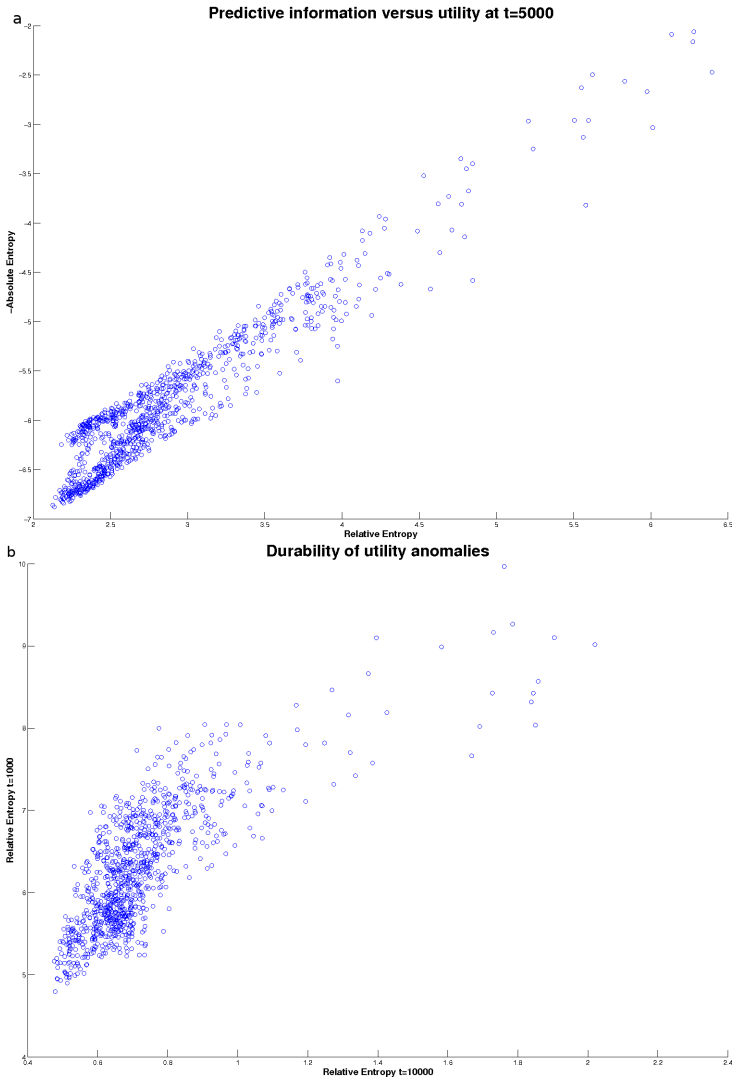

FIGURE 5.3. Top panel shows the relationship between predictive information and prediction utility at 5000 time units (see Figure 5.2). Bottom panel shows the relationship between utility at early (t=1000) and late (t=10,000) predictions.

that it is the higher order cumulant contributions to the entropy which are driving variations in utility. As in the stochastic case, the predictions show considerable durability. This is illustrated in Figure 5.3b which shows the correlation of the utility of early and late predictions.

5.3.3. *Idealized models of turbulence.* These systems attempt to simulate the realistic spectrum of turbulence with between 10^2 through 10^4 degrees of freedom and in an environment which is often homogeneous. They represent a bridge between the models of the previous subsections and realistic models of both atmosphere and

ocean. The nature of the turbulence exhibited means that there is a large range of timescales present and there are large non-linear transfers of energy between modes often in the form of a scale cascade. Thus both the effects of the previous two types of models are typically present.

Calculation of information theoretic functionals for models of this dimension involves essential difficulties. As we saw in the simple chaotic model above, the most straightforward method involves a Monte Carlo method known as a prediction ensemble and some choice of coarse graining to estimate probability densities via a partitioning of state space. When we confront models with significantly higher dimensions however this problem becomes very acute since the required number of partitions increases with the power of the dimension. In order to obtain a statistically reliable estimate of density for each partition we require that the ensemble size be at least the same order as the number of partitions chosen. It is fairly clear that only approximations to the functionals will be possible in such a situation.

Two approaches have been taken to this approximation problem. The first assumes that the ensemble is sufficiently large so that low order statistical moments can be reliably calculated and then implicitly discards sample information regarding higher order moments as not known. One then uses the moment information as a hard constraint on the hypothetical distribution and derives that distribution using a maximum entropy hypothesis by deciding that no knowledge is available regarding the higher order moments. In the case that only second order or less moments are available this implies, of course, that the hypothetical distribution is Gaussian. Typically however some higher order moments can also be reliably estimated from prediction ensembles so a more general methodology is desirable. Under the appropriate circumstances, it is possible to solve for more general distributions using a convex optimization methodology. This was originally suggested by [41] and was developed extensively with the present application in mind by [42], [43], [44], [45] and [46].

In the latter references moments up to order four were typically retained and so generalized skewness and kurtosis statistics were assumed to be reliably estimated from prediction ensembles. An interesting feature of these maximum entropy methods is that as more moments are retained the relative entropy increases. Thus the Gaussian relative entropy is bounded above by that derived from the more general distributions exhibiting kurtosis and skewness effects. This is analogous with the coarse to fine graining hierarchy effect noted in the final paragraph of section 2. One feature of many models of turbulence makes this moment maximum entropy approximation approach attractive: It is frequently observed that the low order marginal distributions are quasi Gaussian and so can usually be very well approximated by retaining only the first four moments. Such a situation contrasts strongly with that noted above in simple chaotic systems where highly non-Gaussian behavior is usual.

Another approach to the approximation issue (developed by the present author in [47]) is to accept that only marginal distributions up to a certain low order are definable from a prediction ensemble. One then calculates the average relative entropy with respect to all possible marginal distributions and calls the result the marginal relative entropy. As in the maximum entropy case, the marginal relative entropy of a certain order defined in this way strictly bounds from above the marginal relative entropies of lower order. This again is consistent with greater

retention of information as higher order marginal distributions are used to discriminate between distributions. Note also that at the top of this chain sits the full relative entropy.

An interesting aspect of these approximations is the role of the prediction ensemble in defining either the moment constraints or the marginal distributions. Some reflection on the matter shows that in fact these objects are subject to sample error and this becomes larger as higher order approximations are considered. This suggests that in the first case the constraints assumed in deriving maximum entropy distributions should not be imposed as hard constraints but instead be weak constraints as they should reflect the presence of the sample error. This rather subtle issue has received little attention to date in the literature in the context of maximum entropy (see however [44]). See [47] for an information theoretic analysis for the simpler marginal distribution estimation case.

Four different turbulence models of varying physical complexity and dimension have been analyzed in detail using the above approximation methods:

(1) A truncated version of Burgers equation detailed in [48]. This system is a one dimensional inviscid turbulence model with a set of conserved variables which enable a conventional Gibbs Gaussian equilibrium distribution. Predictability issues relevant to the present discussion can be found in [49] and [43].
(2) The one dimensional Lorenz 1996 model of mid-latitude atmospheric turbulence detailed in [50]. This model exhibits a variety of different behaviors depending on the parameters chosen. For some settings strongly regular wave-like behavior is observed while for others a more irregular pattern occurs with some resemblance to atmospheric turbulence. The most unstable linear modes of the system tend to have most weight at a particular wavenumber which is consistent with the observed atmosphere. Predictability issues from the present perspective can be found in [43].
(3) Two dimensional barotropic quasi-geostrophic turbulence detailed in, for example, [51]. Barotropic models refer to systems with no vertical degrees of freedom. Quasi-geostrophic models are rotating fluids which filter out fast waves (both gravity and sound) in order to focus on low frequency variability. The barotropic versions aim at simulating very low frequency variability and exclude by design variability associated with mid-latitude storms/eddies which have a typically shorter timescale. Predictability issues are discussed in [45] in the context of a global model with an inhomogeneous background state.
(4) Baroclinic quasi-geostrophic turbulence as discussed in, for example, [52]. This system is similar to the last except simple vertical structure is incorporated which allows the fluid to draw energy from a vertical mean shear (baroclinic instability). These systems have therefore a representation of higher frequency mid-latitude storms/eddies. This latter variability is sometimes considered to be approximately a stochastic forcing of the low frequency barotropic variability although undoubtedly the turbulence spectrum is more complicated. Predictability issues from the present perspective were discussed in [53] for a model with an homogeneous background shear. They are discussed from a more conventional predictability viewpoint in [54].

Generally speaking for all systems the predictability time scale (as measured by the prediction utility) is related to the equilibration time scale of the turbulence being considered. Variations in prediction utility with initial conditions are typically similar to those seen in stochastic or simple chaotic systems. The origin of these variations is a rather complex subject and the answers obtained also vary significantly according to the system examined. In the first subsection above we considered the case where both the prediction and equilibrium distributions were Gaussian. This meant that an analytical formula involving the lowest two moments was possible and we found it convenient conceptually to separate terms into those dependant on the first and second moments of prediction distribution which we referred to as the signal and dispersion. It is important to emphasize however that this separation does not reflect a decomposition of the utility into the predictive information (the uncertainty reduction of the prediction) and a remainder term. Indeed as we saw above, the second term of equation (5.5) which is included in the dispersion, is not at all related to uncertainty reduction but reflects differences in the two distributions unrelated to entropy/uncertainty or their means.

When we move to the maximum entropy approximation for prediction utility and consider moments of higher order it is possible to generalize this separation of prediction moment contributions to the relative entropy into those dependent on the prediction mean and those dependent on higher moments[9]. A comprehensive and rigorous description of this generalized signal/dispersion decomposition is to be found in [43]. The contribution of these two terms was examined in this reference for the first two turbulence systems listed above. The dominant contributor to overall utility variations was found to depend on the parameters chosen for the model as well as the prediction time chosen. There is also within this reference a discussion of when moments beyond the first two are important to a maximum entropy approximation retaining the first four moments.

When the higher order geostrophic turbulence models are considered (models 3 and 4 in the list above), it is usual that a considerable fraction of the equilibrium variability occurs due only to a relatively small number of modes. Predictability studies have usually focused on the equilibration behavior of such modes.

In the study of a barotropic turbulence model [45] the focus was on the first four principal components of the equilibrium distribution. The maximum entropy framework retaining four moments was used and the variability of prediction utility in terms of the generalized signal and dispersion examined with the conclusion that dispersion was overwhelmingly responsible for variations.

A baroclinic turbulence model was studied in [53]. The equilibrium turbulence spectrum in such models is well known to be dominated energetically by the large scale vertically uniform (barotropic) modes (see [55]). The predictability study therefore restricted attention to barotropic modes and the gravest four horizontal Fourier modes. The marginal distributions for both prediction and equilibrium distributions were observed to be quite close to Gaussian so only the first two moments were retained for calculation of entropic functionals by the maximum entropy method. In addition the marginal distributions with respect to these four dimensions was also calculated directly by a rather coarse partitioning. In contrast to

[9]Strictly speaking there are three contributions. One involves the prediction means; another the higher order prediction moments alone and thirdly a cross term involving both. The authors cited in the text group this latter term with the first and call it the generalized signal.

the barotropic model above, variations in prediction utility were overwhelmingly related to variations in the (Gaussian) signal. This strong difference in predictability behavior has yet to be explained. It may be due to the importance in the latter case of the stochastic like forcing of the low frequency barotropic modes by the higher frequency baroclinic modes. These are of course absent in the barotropic turbulence case. Thus it may be that barotropic model is more like the low order chaos case while the baroclinic model is more like the simple stochastic case. This is speculation however and deeper investigation is required.

5.3.4. *Realistic global primitive equation models.* Such models are simplified versions of models used in real time to predict the daily weather. In contrast to most models of the previous subsection, the background for the turbulence is very inhomogeneous as it depends on geographical features such as topography and the jet stream. The latter is driven by the temperature gradient between the equator and pole and so peaks strongly in mid-latitudes. Compared with the quasi-geostrophic turbulence systems, more modes are generally required to explain the equilibrium variance: Of order around 100 compared with less than 10 for the homogeneous baroclinic quasi-geostrophic case. This dynamical situation has tended to limit the application of the approximation approaches discussed earlier. The only study conducted to date with such models from an information theoretic perspective was by the present author in [56]. This used marginal relative entropies on the dominant 100 equilibrium modes[10] as well as the Gaussian maximum entropy approach discussed above. Generally speaking, marginal distributions of dynamical variables tend to be close to Gaussian however this issue remains to be systematically investigated. In the present context it was found that as initial conditions were varied, there was a correlation of order 0.8 between Gaussian estimates of relative entropy and those derived from marginal relative entropies. The latter tended to have a higher correlation among themselves.

In this system, the equilibration time scale implied by the above estimates of prediction utility is a strong inverse function of the amount of mean vertical shear present in the atmosphere. When this shear is strong as it is during winter it is between 1 and 2 months while for the summer weak shear case it is around 3 months. These estimates of predictability are significantly longer than those derived from practical weather prediction where two weeks is generally accepted to be the current predictability limit. This difference may be due to model error of practical predictions and/or due to simplifications present in the model studied in [56] which excludes important random processes such as moist convection which can provide a significant mechanism to reduce predictability.

The variation in prediction utility with respect to a representative set of initial conditions was also studied and the results found were somewhat similar to those noted in the baroclinic turbulence model discussed earlier. Overall relative entropy fluctuations (either Gaussian or marginal) were most strongly related to signal rather than dispersion variations. This effect was more pronounced for short range rather than long range predictions. The dispersion was however also related to utility but at a lower level than signal. Like both the stochastic and simple chaotic models discussed earlier, prediction utility variations tended to show durability i.e.

[10]Practical prediction ensemble sizes restrict marginal relative entropies to order four or less.

high utility predictions tended to remain high throughout a prediction and vice versa.

5.3.5. *Other recent applications.* There have been interesting applications of the functionals discussed here to various climate prediction problems. This has been aided by the fact that to an even greater extent than the turbulence models discussed above, the marginal distributions observed tend to be very close to Gaussian. Indeed it is almost as if a central limit theorem operates in many complex geophysical models in such a way that many distributions of dynamical variables of interest tend to be approximately normal.

The El Nino coupled ocean/atmosphere system was considered in depth with realistic models in [57] and conclusions were broadly in line with the simple stochastic models considered earlier. In general the signal rather than dispersion played a larger role in explaining predictability variation.

The problem of global warming projection was considered from the current perspective by [58]. They used the relative entropy between the equilibrium distribution of various models and the present observed climate/equilibrium distribution as a measure of the veracity of a particular model. They then found a relationship between this veracity and the amplitude of projected global average warming over the next century. Models with higher veracity tended to project greater warming. In related work, global warming model errors were analyzed using an array of empirical information theoretic tools in [59]. This allowed the authors to identify global warming patterns of highest sensitivity.

Part of the issue of long term climate projection concerns its separation into that due to ocean initial conditions (natural decadal variability) and that due to changes in forcing due to the radiative effects of greenhouse gases. This issue was addressed using relative entropy measures of predictability in [60] where it was found that the predictability switches from the first kind to the second kind at leads of roughly 5-10 years.

The predictability of the natural decadal predictability mentioned above was analyzed in detail using relative entropy measures in [61]. A focus was on the behavior of the first few principal components and the authors used the Gaussian signal/dispersion decomposition to carefully analyze the nature of this predictability and the way it transfers between principal components as predictions evolve.

Finally the Gaussian signal/dispersion decomposition of relative entropy has been used to study the nature of the predictability of polar sea ice in a very recent study by C. Bitz and coworkers at the University of Washington (private communication to the author).

6. Conclusions and Outlook

Predictability issues in dynamical models are studied very naturally using information theoretic functionals. This follows because such variables are best considered to be random variables and entropic functionals have a very intuitive interpretation in terms of uncertainty and as measures of the differences between distributions. Adding to this attraction is the fact that these functionals exhibit invariance under general transformations of the variables describing the dynamical system and that they have deep connections with concepts from statistical physics. Indeed it is argued in this review that the utility of a prediction can be measured by the degree of statistical disequilibrium within the system. Thus the study of the

predictability of a system is equivalent to the study of the nature of its statistical equilibration.

Application of the concepts considered here has often relied on the fact that the random variables under study are quasi-Gaussian since this allows application of analytical expressions. It should be emphasized however (and we saw this in Section 4) that a major attraction of information theory is the ability to go beyond this situation and consider highly non-linear and non-Gaussian distributions. As we saw in detail in section 5 considerable progress has been made using maximum entropy methods to improve approximate calculation of functionals in high dimensional system. No doubt however more work remains in this challenging area where the curse of dimensionality always lurks.

The fundamental mechanism that drives equilibration and hence loss of predictability is the dynamical process which leads to statistical irreversibility. This, as we saw in Section 3, is intimately related to the manner in which a system is coarse grained and fundamental progress in that area will drive a deeper understanding of system predictability. Thus the expectation of the author is that further progress in the area of non-equilibrium statistical mechanics will be of great benefit to the study of the basic nature of prediction.

Appendix

The proofs of the theorems in Section 3 are presented here for technical completeness.

Theorem 2.

Proof. It follows that

$$\begin{aligned} -(f\log f)_t &= -f_t\left(\log f+1\right)\\ &= \nabla\cdot(\mathbf{A}f)(\log f+1)\\ &= f\nabla\cdot\mathbf{A}(\log f+1)+\mathbf{A}.(\nabla f)(\log f+1)\\ &= f\nabla\cdot\mathbf{A}(\log f+1)+\nabla\cdot(\mathbf{A}f\log f)-\nabla\cdot\mathbf{A}f\log f\\ &= f\nabla\cdot\mathbf{A}+\nabla\cdot(\mathbf{A}f\log f) \end{aligned}$$

The second term is in the form of a divergence so may be neglected after integration. We are then left with equation (3.4). □

Theorem 3.

Proof. The first term of equation (3.6) results from Theorem 2 above. We consider therefore only the influence of the stochastic forcing part of the FPE. Let $r\equiv -p\ln p$ then the FPE shows that

$$r_t=-\frac{1}{2}\partial_i\partial_j\left(C_{ij}p\right)\left(\log p+1\right)$$

Now it is easy to see that

$$\partial_i\partial_j(C_{ij}uw)=w\partial_i\partial_j(C_{ij}u)+2(\partial_j(C_{ij}u))(\partial_i w)+C_{ij}u\partial_i\partial_j(w) \tag{6.1}$$

where we are using the symmetry of C. Using (6.1) with $u=1$ and $w=p$ this becomes after dropping a divergence term

$$r_t=-\log p\left[p\partial_i\partial_j D_{ij}+2\left(\partial_j D_{ij}\right)\partial_i p+D_{ij}\partial_i\partial_j p\right]$$

with $D_{ij}\equiv\frac{1}{2}C_{ij}$. Now it is easy to show that

$$-\log p\partial_i p=\partial_i r+\partial_i p$$

and so

$$r_t=r\partial_i\partial_j D_{ij}+2\left(\partial_j D_{ij}\right)\left(\partial_i r+\partial_i p\right)-\log pD_{ij}\partial_i\partial_j p$$

Setting $u=1$ and $w=r$ in (6.1) we have

$$\partial_i\partial_j\left(D_{ij}r\right)=r\partial_i\partial_j D_{ij}+2\left(\partial_j D_{ij}\right)\partial_i r+D_{ij}\partial_i\partial_j r$$

which enables us to write

$$r_t=\partial_i\partial_j\left(D_{ij}r\right)-D_{ij}\partial_i\partial_j r+2\left(\partial_j D_{ij}\right)\partial_i p-\log pD_{ij}\partial_i\partial_j p$$

The first term on the right can be dropped as before since it is of the form of a divergence. The third term can be written as

$$2\left(\partial_j\left[D_{ij}\partial_i p\right]-C_{ij}\partial_i\partial_j p\right)$$

which also has a divergence piece which can be dropped. We have then

$$r_t=-D_{ij}\left[\partial_i\partial_j r+\left(2+\log p\right)\partial_i\partial_j p\right]$$

Now we have

$$-\partial_i\partial_j\left(p\log p\right)=-\partial_i\left(\log p\right)\partial_j p-\left(\log p+1\right)\partial_i\partial_j p$$

and so

$$\begin{aligned} r_t &= D_{ij}\left[\partial_i\left(\log p\right)\partial_j p - \partial_i\partial_j p\right] \\ &= pD_{ij}\partial_i\left(\log p\right)\partial_j\left(\log p\right) - D_{ij}\partial_i\partial_j p \\ &= pD_{ij}\partial_i\left(\log p\right)\partial_j\left(\log p\right) + \left(\partial_i D_{ij}\right)\partial_j p - \partial_i\left[D_{ij}\partial_j p\right] \end{aligned}$$

Dropping the final divergence term, using equation (3.5) to drop the second term and integrating over all space we obtain the desired result (3.6). □

Theorem 4.

Proof. Let the two processes have probability functions f and g then it follows that

$$\begin{gathered} \left(f\log(f/g)\right)_t = f_t\log(f/g) + f_t - g_t(f/g) = f_t(\log(f/g)+1) - g_t(f/g) \\ = -\nabla\cdot(\mathbf{A}f)(\log(f/g)+1) + \nabla\cdot(\mathbf{A}g)(f/g) \\ = \left[-f\nabla\cdot\mathbf{A} - \mathbf{A}.(\nabla f)\right]\left[\log(f/g)+1\right] + \left[g\nabla\cdot\mathbf{A} + \mathbf{A}.(\nabla g)\right](f/g) \\ = -\nabla\cdot\mathbf{A}f\log(f/g) - \mathbf{A}.(\nabla f\left[\log(f/g)+1\right] - \nabla g(f/g)) \\ = -\nabla\cdot\mathbf{A}f\log(f/g) - \mathbf{A}.\nabla(f\log(f/g)) \\ = -\nabla\cdot(\mathbf{A}f\log(f/g)) \end{gathered}$$

In this case the entire right hand side of the evolution equation is in the form of a divergence and so may be neglected after integration is carried out. □

Theorem 5.

Proof. As in the previous theorem we consider the relative entropy "density" function $r = f\log(f/g)$. Clearly the proof of this shows we need only consider the time rate of change in this function due to C since that due to $\mathbf{A}$ leads to no change in time of the global integral of r. The change in r due to C is easily calculated using equation (3.2):

$$(r_c)_t = \left(\log(f/g)+1\right)\partial_i\partial_j\left(C_{ij}f\right) - \frac{f}{g}\partial_i\partial_j\left(C_{ij}g\right) \tag{6.2}$$

where we are using the summation convention for repeated Latin indices. Writing $g = f(g/f)$ and applying (6.1) we derive that the second term of equation (6.2) is

$$(2) = -\frac{f}{g}\left[\frac{g}{f}\partial_i\partial_j(C_{ij}f) + 2\partial_i(C_{ij}f)\partial_i\left(\frac{g}{f}\right) + C_{ij}f\partial_i\partial_j\left(\frac{g}{f}\right)\right]$$

combining this with the first term we get a cancellation of the first term of (2) with part of the first term of equation (6.2) and so

$$(r_c)_t = \log(f/g)\partial_i\partial_j\left(C_{ij}f\right) - 2\left(\frac{f}{g}\right)\partial_j(C_{ij}f)\partial_i\left(\frac{g}{f}\right) - \left(\frac{f}{g}\right)C_{ij}f\partial_i\partial_j\left(\frac{g}{f}\right)$$

Now to this equation we add and subtract the terms

$$2\partial_i(\log\frac{f}{g})\partial_j(C_{ij}f) + C_{ij}f\partial_i\partial_j\left(\log\frac{f}{g}\right)$$

and use equation (6.1) to deduce that

$$(r_c)_t = \partial_i\partial_j(C_{ij}r) - \left(C_{ij}f\left[\frac{f}{g}\partial_i\partial_j\left(\frac{g}{f}\right) + \partial_i\partial_j\left(\log\frac{f}{g}\right)\right]\right)$$

where we are using the definition of r as well as canceling two terms involving $\partial_j(C_{ij}f)$. It is straightforward (albeit tedious) to simplify the expression in the square brackets and obtain finally

$$(r_c)_t = \partial_i\partial_j(C_{ij}r) - fC_{ij}\partial_i(\log\frac{f}{g})\partial_j(\log\frac{f}{g}) \tag{6.3}$$

The first term on the right is of the form of a divergence and so as usual does not contribute to the evolution of the global integral of r_C. Actually the positive definite nature of C shows that it is purely diffusive of the density r. The second term is negative almost everywhere due to the fact that C is positive definite almost everywhere and that f and g differ almost everywhere. Thus in that situation if we take the global integral of r_C we conclude that the relative entropy declines strictly with time and satisfies the evolution equation (3.7). □

Relationship of the three predictability functionals. Label the set of statistical predictions with the initial condition vector $\boldsymbol{z}$ and drop the time variable in densities with the assumption that the prediction time outcome is described by the vector $\boldsymbol{x}$. For a stochastic model, the prediction distribution with this initial condition vector is then

$$p_{\boldsymbol{z}}(\boldsymbol{x}) = q(\boldsymbol{x}|\boldsymbol{z})$$

The second functional is then

$$Pr(\boldsymbol{z}) = \int q(\boldsymbol{x}|\boldsymbol{z})\log q(\boldsymbol{x}|\boldsymbol{z})d\boldsymbol{x} + H(\boldsymbol{X}(t))$$

using the fact that all $H(\boldsymbol{X})$ are equal to the equilibrium entropy for this setup. Taking expectation values with respect to the density for $\boldsymbol{z}$ which is $q(\boldsymbol{z})$ we obtain

$$\begin{aligned}\langle Pr\rangle_q &= \int\int q(\boldsymbol{z})q(\boldsymbol{x}|\boldsymbol{z})\log q(\boldsymbol{x}|\boldsymbol{z})d\boldsymbol{x}d\boldsymbol{z} + H(\boldsymbol{X}(t)) \\ &= -H(\boldsymbol{X}(t)|\boldsymbol{X}(0)) + H(\boldsymbol{X}(t)) = I(\boldsymbol{X}(t);\boldsymbol{X}(0)\end{aligned}$$

The third functional is

$$D(\boldsymbol{z}) = \int q(\boldsymbol{x}|\boldsymbol{z})\log\frac{q(\boldsymbol{x}|\boldsymbol{z})}{q(\boldsymbol{x})}d\boldsymbol{x}$$

since the distribution $q(\boldsymbol{x})$ is always the equilibrium distribution. Again taking the same expectation value we obtain

$$\begin{aligned}\langle D\rangle_q &= \int\int q(\boldsymbol{z})q(\boldsymbol{x}|\boldsymbol{z})\log\frac{q(\boldsymbol{x}|\boldsymbol{z})}{q(\boldsymbol{x})}d\boldsymbol{x}d\boldsymbol{z} \\ &= \int\int q(\boldsymbol{x},\boldsymbol{z})\log\frac{q(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})q(\boldsymbol{x})}d\boldsymbol{x}d\boldsymbol{z} = I(\boldsymbol{X}(t);\boldsymbol{X}(0))\end{aligned}$$

References

[1] Sussman, G.; Wisdom, J. Chaotic evolution of the solar system. *Science* **1992**, *257*, 56.

[2] Knopoff, L. Earthquake prediction: the scientific challenge. *P. N. A. S.* **1996**, *93*, 3719.

[3] DelSole, T. Predictability and Information Theory. Part I: Measures of Predictability. *J. Atmos. Sci.* **2004**, *61*, 2425–2440.

[4] Eyink, G.; Kim, S. A maximum entropy method for particle filtering. *Journal of Statistical Physics* **2006**, *123*, 1071–1128.

[5] Castronovo, E.; Harlim, J.; Majda, A. Mathematical test criteria for filtering complex systems: Plentiful observations. *Journal of Computational Physics* **2008**, *227*, 3678–3714.

[6] Cover, T.; Thomas, J. *Elements of information theory*, 2nd ed.; Wiley-Interscience: New York, 2006.

[7] Amari, S.; Nagaoka, H. *Methods of Information Geometry*; Translations of Mathematical Monographs, AMS, Oxford University Press, 2000.
[8] Daems, D.; Nicolis, G. Entropy production and phase space volume contraction. *Phys. Rev. E* **1999**, *59*, 4000–4006.
[9] Garbaczewski, P. Differential entropy and dynamics of uncertainty. *J. Stat. Phys.* **2006**, *123*, 315–355.
[10] Ruelle, D. Positivity of entropy production in nonequilibrium statistical mechanics. *Journal of Statistical Physics* **1996**, *85*, 1–23.
[11] Lebowitz, J.; Bergmann, P. Irreversible gibbsian ensembles. *Annals of Physics* **1957**, *1*, 1–23.
[12] Gardiner, C.W. *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*; Vol. 13, *Springer Series in Synergetics*, Springer, 2004.
[13] Kaneko, K. Lyapunov analysis and information flow in coupled map lattices. *Physica D* **1986**, *23*, 436.
[14] Vastano, J.A.; Swinney, H.L. Information transport in spatiotemporal systems. *Phys. Rev. Lett.* **1988**, *60*, 1773.
[15] Schreiber, T. Spatiotemporal structure in coupled map lattices: Two point correlations versus mutual information. *J. Phys.* **1990**, *A23*, L393–L398.
[16] Schreiber, T. Measuring Information Transfer. *Phys. Rev. Lett.* **2000**, *85*, 461–464.
[17] Liang, X.S.; Kleeman, R. Information transfer between dynamical system components. *Phys. Rev. Lett.* **2005**, *95*, 244101.
[18] Liang, X.S.; Kleeman, R. A rigorous formalism of information transfer between dynamical system components I. Discrete maps. *Physica D.* **2007**, *231*, 1–9.
[19] Liang, X.S.; Kleeman, R. A rigorous formalism of information transfer between dynamical system components II. Continuous flow. *Physica D.* **2007**, *227*, 173–182.
[20] Majda, A.J.; Harlim, J. Information flow between subspaces of complex dynamical systems. *Proceedings National Academy of Science* **2007**, *104*, 9558–9563.
[21] Zubarev, D.; Morozov, V.; Ropke, G. Statistical mechanics of nonequilibrium processes. Vol. 1: Basic concepts, kinetic theory **1996**.
[22] Palmer, T.N.; Gelaro, R.; Barkmeijer, J.; Buizza, R. Singular vectors, metrics, and adaptive observations. *J. Atmos. Sci.* **1998**, *55*, 633–653.
[23] Bishop, C.; Etherton, B.; Majumdar, S. Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. *Mon. Weath. Rev.* **2001**, *129*, 420–436.
[24] Kleeman, R. Information flow in ensemble weather predictions. *J. Atmos Sci.* **2007**, *64*, 1005–1016.
[25] Jolliffe, I.; Stephenson, D. *Forecast verification: a practitioner's guide in atmospheric science*; Wiley, 2003.
[26] Leung, L.Y.; North, G.R. Information Theory and Climate Prediction. *J. Clim.* **1990**, *3*, 5–14.
[27] Schneider, T.; Griffies, S. A conceptual framework for predictability studies. *J. Clim.* **1999**, *12*, 3133–3155.
[28] Kleeman, R. Measuring dynamical prediction utility using relative entropy. *J. Atmos. Sci.* **2002**, *59*, 2057–2072.
[29] Roulston, M.S.; Smith, L.A. Evaluating Probabilistic Forecasts Using Information Theory. *Mon. Weath. Rev.* **2002**, *130*, 1653–1660.
[30] Bernardo, J.; Smith, A. *Bayesian Theory*; John Wiley and Sons, 1994.
[31] Kleeman, R.; Moore, A.M. A new method for determining the reliability of dynamical ENSO predictions. *Mon. Weath. Rev.* **1999**, *127*, 694–705.
[32] Hasselmann, K. Stochastic climate models. Part I. Theory. *Tellus* **1976**, *28*, 473–485.
[33] Kleeman, R.; Power, S. Limits to predictability in a coupled ocean-atmosphere model due to atmospheric noise. *Tellus A* **1994**, *46*, 529–540.
[34] Penland, C.; Sardeshmukh, P. The optimal growth of tropical sea surface temperature anomalies. *J. Clim.* **1995**, *8*, 1999–2024.
[35] Kleeman, R. Spectral analysis of multi-dimensional stochastic geophysical models with an application to decadal ENSO variability. *J. Atmos. Sci.* **2010**. To appear.
[36] Saravanan, R.; McWilliams, J. Advective ocean–atmosphere interaction: An analytical stochastic model with implications for decadal variability. *Journal of Climate* **1998**, *11*, 165–188.
[37] DelSole, T.; Farrel, B.F. A stochastically excited linear system as a model for quasigeostrophic turbulence: analytic results for one- and two-layer fluids. *J. Atmos. Sci.* **1995**, *52*, 2531–2547.

[38] Lorenz, E.N. Deterministic non-periodic flows. *J. Atmos. Sci.* **1963**, *20*, 130–141.
[39] Lorenz, E.N. Irregularity: a fundamental property of the atmosphere. *Tellus* **1984**, *36A*, 98–110.
[40] Nayfeh, A.; Balachandran, B. *Applied nonlinear dynamics*; Vol. 2, Wiley Online Library, 1995. 685 pp.
[41] Mead, L.R.; Papanicolaou, N. Maximum entropy in the problem of moments. *J. Math. Phys.* **1984**, *25*, 2404–2417.
[42] Majda, A.J.; Kleeman, R.; Cai, D. A framework of predictability through relative entropy. *Meth. Appl. Anal.* **2002**, *9*, 425–444.
[43] Abramov, R.; Majda, A. Quantifying uncertainty for non-Gaussian ensembles in complex systems. *SIAM Journal on Scientific Computing* **2005**, *26*, 411–447.
[44] Haven, K.; Majda, A.; Abramov, R. Quantifying Predictability Through Information Theory: Small Sample Estimation in a Non-Gaussian Framework. *J. Comp. Physics* **2004**. Submitted.
[45] Abramov, R.; Majda, A.; Kleeman, R. Information Theory and Predictability for Low Frequency Variability. *J. Atmos Sci* **2005**, *62*, 65–87.
[46] Abramov, R. A practical computational framework for the multidimensional moment-constrained maximum entropy principle. *Journal of Computational Physics* **2006**, *211*, 198–209.
[47] Kleeman, R. Statistical predictability in the atmosphere and other dynamical systems. *Physica D* **2007**, *230*, 65–71.
[48] Majda, A.J.; Timofeyev, I. Statistical mechanics for truncations of the Burgers-Hopf equation: a model for intrinsic stochastic behavior with scaling. *Milan Journal of Mathematics* **2002**, *70(1)*, 39–96.
[49] Kleeman, R.; Majda, A.J.; Timofeyev, I. Quantifying predictability in a model with statistical features of the atmosphere. *Proc. Nat. Acad. Sci. USA* **2002**, *99*, 15291–15296.
[50] Lorenz, E. Predictability: A problem partly solved. Proc. Seminar on Predictability, 1996, Vol. 1, pp. 1–18.
[51] Selten, F. An efficient description of the dynamics of barotropic flow. *J. Atmos. Sci.* **1995**, *52*, 915–936.
[52] Salmon, R. Baroclinic instability and geostrophic turbulence. *Geophys. Astrophys. Fluid Dyn.* **1980**, *15*, 167–211.
[53] Kleeman, R.; Majda, A.J. Predictability in a model of geostrophic turbulence. *J. Atmos Sci* **2005**, *62*, 2864–2879.
[54] Vallis, G.K. On the predictability of Quasi-Geostrophic Flow: The Effects of Beta and Baroclinicity. *J. Atmos. Sci.* **1983**, *40*, 10–27.
[55] Salmon, R. *Lectures on Geophysical Fluid Dynamics*; Oxford Univ. Press, New York, 1998.
[56] Kleeman, R. Limits, variability and general behaviour of statistical predictability of the mid-latitude atmosphere. *J. Atmos Sci* **2008**, *65*, 263–275.
[57] Tang, Y.; Kleeman, R.; Moore, A. Comparison of Information-based Measures of Forecast Uncertainty in Ensemble ENSO Prediction. *J. Clim.* **2008**, *21*, 230–247.
[58] Shukla, J.; DelSole, T.; Fennessy, M.; Kinter, J.; Paolino, D. Climate model fidelity and projections of climate change. *Geophys. Res. Lett* **2006**, *33*, L07702.
[59] Majda, A.; Gershgorin, B. Quantifying uncertainty in climate change science through empirical information theory. *Proceedings of the National Academy of Sciences* **2010**, *107*, 14958.
[60] Branstator, G.; Teng, H. Two Limits of Initial-value Decadal Predictability in a CGCM. *Journal of Climate* **2010**.
[61] Teng, H.; Branstator, G. Initial-value predictability of prominent modes of North Pacific subsurface temperature in a CGCM. *Climate Dynamics* **2010**, pp. 1–22.

Acknowledgements

The author wishes to thank the Isaac Newton Institute at Cambridge University where this review was written during a visit in August 2010. Stimulating discussions over the past few years with Tim DelSole, Greg Eyink, Andy Majda, Ilya Timofeyev and Bruce Turkington are gratefully acknowledged. Support from the National Science foundation from a variety of grants is also gratefully acknowledged.

Courant Institute of Mathematical Sciences, New York University, New York