Convective Precursors and Predictability in the Tropical Western Pacific

STEVEN C. SHERWOOD

School of Earth Sciences/School of Mathematics and Computing Sciences, Victoria University of Wellington, Wellington, New Zealand

(Manuscript received 23 September 1998, in final form 17 December 1998)

ABSTRACT

Conditions leading to convective outbreak in the Tropics are investigated by multivariate analysis of sounding and satellite data from the tropical western Pacific area. Circumstances that make the prediction problem difficult are discussed and addressed by applying linear “error-in-variables” and nonlinear statistical simulation techniques to a large dataset.

Low- to midtropospheric moisture is identified as the dominant factor regulating convective outbreak in this region. Based on the results it is argued that such moisture is particularly important in regulating spontaneous convective outbreak, but instability and near-surface wind speed probably play some role in allowing previous shallow or midtopped cumulus activity to deepen. Mesoscale-mean convective available potential energy sufficient for convection is found to exist almost 90% of the time.

Quantitative estimates of noise in the data are obtained and accounted for in reaching these conclusions. The results imply that large-scale mean fields alone may not contain enough information to determine the behavior of convection except probabilistically. Both types of statistical model predict that even under favorable mesoscale-mean conditions, convection is typically only 20%–30% likely to break out during a given 3-h period.

1. Introduction

Deep convective storm development remains one of the most challenging problems facing meteorologists. Though certain variables are known to affect the probability of outbreak, convection (especially over open oceans) still catches operational forecasters by surprise more often than would be hoped for. The representation of convective effects in numerical weather and climate models is also a vexing problem, since the details of how this is done can have a profound impact on model behavior, and theoretical treatments of the subject are still developing (see Blyth 1993).

In predicting and modeling convection it is important to distinguish between the factors necessary for convection to begin (sometimes called “activation” criteria), and those determining its subsequent behavior once it has started (Moncrieff 1992). The latter problem has received more attention, partly because it has been considered more complex and less well understood than the activation problem (Moncrieff 1981). However, this study will focus on activation.

Many observational analyses of convective systems have been performed to determine predictive factors—mainly in midlatitude, continental situations. Several factors have been identified that seem necessary or at least favorable for the development of organized convection, including buoyant instability, water vapor above the trade wind inversion, upper-level divergence, low-level convergence or upward motion at some level, and upper-level vorticity. Which of these factors are important depends on the scale of the disturbance; synoptic-scale conditions that are required for cyclogenesis, for example, may not be necessary for development of the more ubiquitous mesoscale convective system (MCS), the conditions for which in turn are likely more stringent than those for individual cumuli. Furthermore, important factors may differ between the Tropics and midlatitudes. Predictability has been examined extensively at midlatitudes and found to be modest at best (Pepper and Lamb 1989, and references therein). Despite reasonably good qualitative understanding of the factors necessary for the growth of individual cumulus clouds, studies attempting to quantify the relative importance of the various influences on organized, tropical mesoscale development directly from observations have encountered difficulty (Mapes and Houze 1992; Alexander and Young 1992).

The empirical studies have been thwarted by two serious obstacles, which are particularly severe in the Tropics. First, convection appears to be very sensitive to small changes in the background atmospheric state, which must therefore be measured with corresponding accuracy. Unfortunately, current observations (inclu-
We progress beyond previous efforts in two respects. First, the problem is investigated more thoroughly than previously possible in the Tropics, using a novel data-gathering technique and attacking the difficulties mentioned above by means of (a) a relatively large dataset and (b) inclusion of measurement error explicitly in the analysis. Second, the prediction problem is clearly posed, and we go beyond the attribution problem to estimate the degree to which tropical convection is actually determined at all by mean field behavior. In other words, when observational problems are taken into account, how much unpredictability is left in the system? This question is highly relevant to attempts to simplify tropical dynamics in models and theory. It is distinct from the simple problem of measuring predictability in practice and requires more work.

Accordingly, a key element of this study is the explicit treatment of measurement error. Here we acknowledge the fact, often unremarked, that appropriate predictors are areal mean quantities, though most relevant measurements are not. Many predictors are poorly observed or highly variable in space. It is impossible to quantify the randomness or determinacy of convection from a large-scale perspective without explicitly modeling this noise. Noise is estimated here using data from the Central Equatorial Pacific Experiment (CEPEX) and other field programs. The error estimates are used to perform so-called error-in-variables analyses, including nonlinear, conjunctive approaches with noise. It is argued based on the results that large-scale thermodynamic fields do not, by themselves, determine the subsequent behavior of convection within forecast periods much less than perhaps half a day.

2. Variables and data
a. Candidate causes

Following is a list of variables that may be expected to influence the probability of convective outbreak, or the degree of growth.

- **Deep (conditional) instability.** This is the most fundamental variable of importance to cumulus cell development. It is sensitive to the lapse rate, but also to the humidity of rising parcels as noted by Normand (1938). Here, “deep” instability refers to the buoyancy of a near-surface air parcel, integrated as high as the parcel could go if ascending undiluted from its level of free convection. Deep instability is often identified as an important precursor for organized deep convection in the Tropics (e.g., Johnson and Kriete 1982), has well-known prediction skill in midlatitudes (e.g., Peppler and Lamb 1989), and is clearly correlated with convection in global climatologies. More recent studies of particular regions within the Tropics, however, have not found strong relationships between convective available potential energy (CAPE), a measure of deep instability,
and convective development (Mapes and Houze 1992; Alexander and Young 1992).

- **Shallow instability.** It is widely believed that shallow inversions are an effective inhibitor of convection even in the deep Tropics, regardless of the presence of deep instability (e.g., Mapes and Houze 1992), although this is sometimes disputed (Emanuel et al. 1994).

- **Free moisture.** It was recognized early on (Stommel 1947) that entrainment of dry air would dilute the properties of cumulus updrafts, and that the resulting buoyancy loss would depend on the amounts of vapor present in the free troposphere above the inversion. Vapor at these levels has indeed been identified as an important factor for convective development in observational studies of the Tropics (Malkus 1954; Brown and Zhang 1997), in midlatitude statistical studies (Peppler and Lamb 1989; Burpee and Lahiff 1984; Zawadzki and Ro 1978), and in numerical and conceptual models (Nicholls et al. 1988; Raymond and Torres 1998).

- **Vertical velocity/low-level convergence.** This has been observed in advance of tropical convective development (Gray and Jacobson 1977; Frank 1978), has been emphasized in previous explanations of convective location and organization (McBride and Gray 1980), and was famously incorporated by Kuo (1965) and others into convective schemes for global models. However, any direct influence of slow, large-scale, vertical displacements on rapid, cumulus-scale, eddy development is very difficult to justify mechanistically; it has been forcefully argued that the correlation between the two is almost certainly mediated by one or more of the thermodynamic restraints on convection listed above, each of which can easily be removed by modest lifting (Emanuel et al. 1994; Raymond 1995). In this study, we must adopt their argument as a working assumption, since no adequate vertical velocity data are available to be included in the analysis.

- **Shear.** Vertical shear in the large-scale wind field clearly influences the type of organization exhibited by convective systems once they form (Barnes and Sieckman 1984; Nicholls et al. 1988; Rotunno et al. 1988). Whether shear is important for activation of convection is less clear, but some models suggest that it is (Rotunno et al. 1988; Moncrieff 1992). Shear in the upper levels is thought to inhibit convection, according to these studies.

- **Surface wind speed.** Wind speed near the surface has been implicated by Raymond (1995) as an indirect cause of convection over tropical oceans through its impact on surface energy fluxes and low-level parcel \( \theta_e \). Note however that if sufficiently accurate measurements of boundary-layer \( \theta_e \), shallow, and/or deep instability can be obtained, these should render the wind speed predictor irrelevant, at least regarding this mechanism (more explanation on this important point below). Another way that wind speed could promote convection is by enabling parcels to overcome weak inversions, since the turbulent kinetic energy available to parcels increases with mean air velocity in the surface layer.

### b. Observations

All variables in the list above, except vertical velocity, are obtainable from standard radiosondes—albeit with the usual sampling and accuracy problems.

This study employs a set of 6501 soundings from Sherwood and Wahrlich (1999). The soundings, which passed rigorous quality control procedures, were launched from nine operational upper-air stations in the tropical western Pacific region during all of 1995 and the first half of 1996. Each was collocated with hourly, full-resolution Geostationary Meteorological Satellite (GMS) IR1 (infrared) imagery collected from 120 km × 120 km regions near the sounding sites, starting 6 h before through 6 h after the sounding launch time. The collection region for each sounding was advected with the lower- to midtropospheric winds measured at the site, passing directly over the site at launch time, to ensure that any temporal changes observed in the convective state occurred in the same air mass that was observed by the sonde (rather than at the same geographic location). In this manner, the characteristics of preconvective air masses can be compared with those of air masses not about to experience convection, or with those having undergone recent convection. Only those soundings possessing all data, including 12-h satellite histories, were used (3300 in total).

Sherwood and Wahrlich (1998) classified the soundings on the basis of histograms of IR1 brightness temperature \( T_b \) in the collection region as a function of time relative to launch time. The fractions of pixels falling below three \( T_b \) thresholds (208, 235, 267 K) were collected at each time \(-5.5, -4.5, \ldots, +4.5, +5.5\) h, and these cloud-fraction time series were denoted \( c_{208}(t), c_{235}(t), c_{267}(t) \), respectively. Here \( c_{208} \) represented the amount of active deep convection, and peaks in \( c_{208} \) were followed by subsequent peaks in coverage at the warmer thresholds, indicating spreading anvils.

In this study, only those soundings assigned to stage 0 (nonconvective) or 1 (preconvective) by Sherwood and Wahrlich (1999) are used, with the variable \( \Omega \in \{0, 1\} \) denoting which group or “class” the sounding belongs to. Some soundings that did precede convective onset by three hours or less but fell between convective events and were assigned to the postconvective stage in that study were added here as members of the \( \Omega = 1 \) class. With the addition of these, 6.2% of the soundings belong to the \( \Omega = 1 \) class.

#### c. A predictor set

To avoid overfitting, any statistical model must be based on a manageable number of quantities, or a “predictor set.” A set of 10 (Table 1) has been chosen to represent the quantities listed earlier in the section.
Table 1. Predictor set and error amplitudes. The error amplitudes are rms deviations between point values from a radiosonde and true mesoscale-mean values, estimated as described in section 4.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Error Amplitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAPE</td>
<td>Convective available potential energy</td>
<td>294 J kg⁻¹</td>
</tr>
<tr>
<td>CIN</td>
<td>Convective inhibitive energy</td>
<td>(50 J kg⁻¹) *</td>
</tr>
<tr>
<td>RHlow</td>
<td>RH, 900–800 hPa mean</td>
<td>8.1%</td>
</tr>
<tr>
<td>RHmed</td>
<td>RH, 700–500 hPa mean</td>
<td>7.3%</td>
</tr>
<tr>
<td>RHhigh</td>
<td>RH, 400–300 hPa mean</td>
<td>9.0%</td>
</tr>
<tr>
<td>Tmed</td>
<td>Temperature, 700–500 hPa mean</td>
<td>0.45 K</td>
</tr>
<tr>
<td>Thigh</td>
<td>Temperature, 400–300 hPa mean</td>
<td>0.36 K</td>
</tr>
<tr>
<td>WSPD</td>
<td>Speed of mean wind, 1000–900 hPa</td>
<td>1.25 m s⁻¹</td>
</tr>
<tr>
<td>WSHRlow</td>
<td>Wind shear, 700–850 hPa</td>
<td>1.93 m s⁻¹</td>
</tr>
<tr>
<td>WSHRhigh</td>
<td>Wind shear, 400–700 hPa</td>
<td>2.95 m s⁻¹</td>
</tr>
</tbody>
</table>

* The noise of the transformed-CIN variable is 0.91, equivalent to a factor of 2.8 in CIN.

The first predictor, CAPE, is a standard measure of deep instability, here calculated as the energy released by a parcel taken from a starting level $p_0$ and lifted to its level of neutral buoyancy, averaged as a function of $p_0$ from 1000 to 900 hPa. Convective inhibition energy (CIN), a measure of shallow stability, is the energy required to reach the level of free convection, similarly averaged. In this study the transformation $\ln|\text{CIN}| (\text{J kg}^{-1}) + 10$, having an approximately Gaussian distribution, is employed in place of CIN in the linear analyses. CIN is very difficult to measure, as it is small and contained in a shallow layer, and has the poorest signal-to-noise ratio (see Table 1 and section 4). Relative humidity (RH) is also included in three layers, and temperature ($T$) in two layers. Two measures of wind shear are included: the magnitude of the vector difference between the 700- and 850-hPa winds ($\text{WSHR}_{\text{low}}$), and that between the 700- and 400-hPa levels ($\text{WSHR}_{\text{high}}$), respectively. Finally, wind speed averaged between 900 and 1000 hPa (WSPD) is included as well. Table 1 also lists the estimates of the noise in each variable, which are discussed in section 4.

3. Overview: Conditional independence and the linear graphical chain model

Sherwood and Wahrlich (1999) composited soundings according to their placement in the convective life cycle, an analysis technique that illustrated the changing atmospheric mean state accompanying convective development and decay. In the present study, the focus will be placed on distinguishing the preconvective and nonconvective sounding classes. First, these two converse views—the development of the mean-field environment during and after convection, and the influence of the environment on the development of convection—are unified here in the form of a graphical chain model.

a. Description

Generally, a graphical model (or simply graph) is a model of the connections between variables in a system. Variables are represented by nodes in the graph. Two nodes in the graph are connected by an edge unless (and only unless) the corresponding variables are conditionally independent, given the rest. Conditional independence means that the variables are completely independent once values of the others in the system are known, and implies that they do not interact directly—though they may still appear correlated in the data through mutual interaction with one or more other, intermediary variables. For example, the contentions made earlier concerning vertical velocity (cf. Emanuel et al. 1994) amount to the belief that convective development $\Omega$ is conditionally independent of vertical velocity $w$, though $\Omega$ may be correlated with $w$, the relationship would (in this belief) be mediated by thermodynamic variables that are correlated with both $\Omega$ and $w$. A good introduction to graphical models and their causal interpretation may be found in Pearl (1988), along with further discussion and a mathematical statement of the conditional independence concept.

In a graphical model in which nothing is known about the variables a priori, edges imply causal links but cannot generally indicate the direction of influence. If, however, the variables can be divided into a few ordered groups or “blocks,” where it is known a priori that variables can influence subsequent blocks but not previous ones, then a chain graph can be constructed in which links spanning between blocks become arrows indicating cause and effect (though the existence of a significant link does not prove that the relationship is causal, unless one is certain that all causes have been included in the analysis; potential worries associated with “hidden variables” are discussed in section 5). Simple temporal ordering of the observations is the usual way of establishing the blocks, since effects cannot precede their causes. The (temporal) sequence of blocks forms a chain. Once blocks are arranged, variables in each block are in essence multiply regressed onto those in the previous blocks. Here, we will consider a linear version of this type of model.

The technique of fitting a linear, graphical chain model to observed data will not be described here in detail; the reader is referred to Whittaker (1990), whose technique is employed here. The exercise involves a test at each edge to determine whether the conditional independence hypothesis can be rejected at a preselected confidence level (here, 95%). Edges that are not found to be significant are removed, and the linear model is refitted. In the final model, all remaining edges are significant. Each such edge may be quantified according to the partial correlation between the connected variables given the rest, which is the linear measure of conditional independence.

The graph here is composed of three blocks. Block 1 will contain variables characterizing convective activity prior to a sounding; block 2, the sounding observables; block 3, a single variable $\Omega$ indicating whether or not convection ensued during the 3 h after the
the tendency of the atmosphere to damp horizontal changes in virtual rather than actual temperature could also be contributing. Horizontal winds are not strongly related to the local thermodynamic variables. In block 1, we see that $c_{235}^{(12)}$ is conditionally independent of $c_{208}^{(12)}$ given $c_{235}^{(30)}$ despite the fairly strong correlation between them (not shown), also an intuitively appealing result. In general, the relationships shown within blocks are consistent with what we would expect.

The influences of block-1 variables on block 2 shown in Fig. 1 also correspond with previous results (Sherwood and Wahrlich 1999, and references therein). There is a strong moistening and warming influence at upper levels (heavy arrows pointing from block 1 to $T_{\text{high}}$ and RH$_{\text{high}}$), a weaker, cooling/moistening influence at middle levels, and a modest increase of WSPD. The direct moistening effect is significant at the 99% level only at upper levels, though moistening is shown at the mid- and low levels too. The moistening at low levels is peculiar since RH$_{\text{low}}$ actually falls slightly during convection (Zipser 1977; Sherwood and Wahrlich 1999).

These results can be reconciled by noting that the block-1 variables are also connected to RH$_{\text{low}}$ by an indirect sequence of two links, mediated by $T_{\text{mid}}$, and that each of these links is stronger than the direct one. The signs of the two indirect links (convection lowering $T_{\text{mid}}$, and this lowering RH$_{\text{low}}$) are fully consistent with Zipser’s (1977) interpretation of the 800-hPa drying during convection as an adiabatic effect caused by midlevel cooling and sinking outside of convective-scale drafts.

b. Results

The model fitted to these variables is displayed in Fig. 1. The partial correlation (or interaction strength) of each edge is proportional to the thickness of the line.

1) CONVective EFFECTS ON SOUNDING VARIABLES

One thing clear from the graph is that correlations among the sounding variables, and those among the satellite variables in blocks 1 and 3, are stronger than correlations between sounding and satellite variables. The difference is not surprising, since sampling variability will be correlated among the sounding variables, and satellite observations may be expected to correlate highly over time lags short compared to the synoptic decorrelation time of several days. Strong links are shown between CAPE and CIN (which is expected since both depend on low-level parcel buoyancy), between RH at the different levels, between $T$ at the different levels, and between $T$ and CAPE (which is expected since $T$ is used in calculating CAPE). The strongest link of all is a negative one between RH and $T$ at midlevels, indicating that (all other things being equal) warm and dry anomalies there tend to coincide; this is probably the signature of adiabatic vertical displacements, though

2) SOUNDING VARIABLE EFFECTS ON CONVECTION

Bolstered by the success of the model in capturing previous results, we turn to the influence of the predictors on subsequent convection, $\Omega$. Two are significant at the 99% confidence level: RH$_{\text{low}}$ and RH$_{\text{mid}}$; CIN and $T_{\text{high}}$, anomalies act to inhibit and enhance convection, respectively, but these effects are weaker. The CIN effect is expected, but the temperature connection (discussed below) is not what one would at first expect.

The absence of any link between CAPE and $\Omega$ is notable. We anticipate such absence from studies such as Mapes and Houze (1992), who found that deep instability did not seem to play a role in initiating convection. The result here does not mean that CAPE is unnecessary for convection; later (section 7) it is concluded that there is enough CAPE around in the tropical western Pacific area so that other variables are left as the limiting factors. From Fig. 1, the dominant such factor appears to be the RH up to at least 500 hPa. Caution must be observed, however, due to the thickness of the arrows pointing from blocks 1 to 3 in Fig. 1. This implies that the sounding variables in toto are not able to account for the convective behavior any better than persistence is able to do. The attribution problem therefore needs further discussion.

![Fig. 1](image-url)

Fig. 1. The partial correlation (or interaction strength) of each edge is proportional to the corresponding partial correlation coefficient. Solid and dashed links indicate positive and negative coefficients, respectively, which are significant at 99% confidence; dot-dashed and dotted links indicate positive and negative coefficients, respectively, which are significant only at 95% confidence.
4. Probability, predictability, and predictor errors

Our goal is to discover (or approximate as well as possible) the function

$$\Pr(\Omega = 1 | x) = f(x),$$

(1)

where $\Pr(\Omega = 1 | x)$ is the probability that $\Omega = 1$ given the state vector $x$ describing the local atmospheric state. Both $\Omega$ and $x$ must be defined over a finite area (not a point), as discussed in the introduction. Here $f$ is a nonlinear function or operator. In terms of this equation, the two difficulties mentioned in the introduction are that (i) the product of the observational uncertainty in $x$ and the matrices of cross derivatives of $f$ is not negligible, and (ii) the inverse problem (computation of $f$ from observations of $\Omega$ and $x$, e.g., using the normal equations of linear theory) is poorly conditioned due to high correlation of different components of $x$.

A sufficiently rich state vector $x$ would be adequate to determine $\Omega$ absolutely. In this case, the range of $f$ would be zero or one only. As the observation vector $x$ is reduced in scope (i.e., if certain relevant variables are not included), the predictability decreases, with $f$ taking on fractional values between zero and one. If no useful information is left in $x$, then $f$ will be a small constant. In our case, $x$ includes the large-scale means of thermodynamic and wind fields (except vertical velocity) typically available to forecasters and prognosed by numerical models. It would be interesting to know how well convection in the real atmosphere is actually determined by these quantities. One index of the degree of determinacy is the mean value of $f$ in cases where convection actually breaks out:

$$\langle f \rangle_{t_1} = \frac{\int_{\mathfrak{R}_x} f^2 d[\Pr(x)]}{\int_{\mathfrak{R}_x} f d[\Pr(x)]},$$

(2)

where the range of integration is the entire predictor-space $\mathfrak{R}_x$, though in practice, the integrals of $d[\Pr(x)]$ over $\mathfrak{R}_x$ are simply approximated by sums over the observed data. This function can range from the uninformed probability $\Pr(\Omega = 1)$ if $f(x)$ is constant, to unity if $x$ is a perfect predictor. Estimates of $\langle f \rangle_{t_1}$ are given in section 8.

a. Predictor error modeling

Missing predictors in $x$ is not the only factor degrading the predictability of $\Omega$. Determination of $f$ itself is complicated by the fact that we do not observe the vector of areal-mean variables $x$, but rather $x'$, the sum of $x$, and a noise component $e$. Part of $e$ is instrument error, but most results from the fact that the sounding is a point measurement that captures unwanted variability on all scales smaller than the one being investigated.

Noise has two effects. First, the problem of predicting $\Omega$ will appear to be more stochastic than it really is, if we treat $x'$ as though it were $x$. Second, noise in any predictor leads to systematic underestimation of its influence on the target variable in any regression calculation, unless that noise is explicitly included in the data model on which the regression is based. To take noise into account, we must estimate the statistics of $e$ and $x$ separately. It is assumed here that they are generated by independent processes.

Accordingly, a so-called error-in-variables model is now introduced. Given a data covariance matrix $M$ and a sample error covariance matrix $\Gamma$, the minimum-variance estimator of the signal covariance matrix $S$ is (Ful- ler 1987)

$$S = M - \left( g - p + 4 + \frac{2N/N_e}{N - 1} \right) \Gamma$$

$$g = \min \left( \lambda - \frac{1}{N - 1} , 1 \right),$$

(3)

where $\lambda$ minimizes $\det(\mathbf{M} - \lambda \mathbf{\Gamma})/\det(\mathbf{M})$, $p$ is the number of predictors, and $N, N_e$ are the numbers of independent samples used to estimate $\mathbf{M}$ and $\mathbf{\Gamma}$, respectively.

Unfortunately, estimating the full error covariance matrix will put a serious strain on the available observations that can be used to quantify noise (see below), so an alternative signal estimate,

$$S' = r^T M r, \quad r_i = \sqrt{\frac{M_{ii} - I_{ii}}{M_{ii}}},$$

(4)

is also considered. Here, the error correlation structure is assumed to be equal to that of the signal, and only the error amplitudes,

$$\sigma_i = \sqrt{\Gamma_{ii}},$$

(5)

must be determined from the noise observations themselves. Noise amplitudes can be estimated more robustly from limited data than can noise correlation, and they can be estimated from platforms other than soundings. It will turn out that, with available observations, the error correlation structure does not differ from that of the data at the 95% confidence level, which makes $S'$ and $S$ equally plausible estimates.

b. Observational estimates of predictor error

Here $\Gamma$ is obtained from a set of soundings launched as part of CEPEX, in which dense sampling of a small area occurred. Approximately 80 radiosondes and dropsondes were deployed from ship and aircraft along the 2°S lat line, from 170°E to 170°W long, most over a 10-day period (Kley et al. 1997). Synoptic variability during this sampling period was light, with a steady zonal gradient of cloud cover and water vapor during most of the period.
Consider a set of three soundings launched reasonably close together ($\Delta t \leq 200$ km, $\delta t \leq 6$ h), denoted $x_1$, $x_0$, and $x_{-2}$. The middle observation $x_0$ and the mean $(x_1' + x_2')/2$ may each be regarded as an independent, unbiased estimate of the true, large-scale mean $x$ at the time of the middle observation, as long as the signal change with time is linear. If the noise process is Gaussian with variance $\text{var}(e)$, these estimates will have variances about $x$ of $\text{var}(e)$ and $\text{var}(e)/2$, respectively, so the difference between these estimates will be distributed as a Gaussian process with variance 1.5 $\text{var}(e)$. Taking two-thirds of the sample covariance matrix of these differences therefore yields $\Gamma$ at approximately the length and timescales of interest. This has been done using 55 such triads from the CEPEX data. The error due to small-scale variability may be overestimated, since the ($\Delta t$, $\delta t$) window used was somewhat larger than the ~120-km region size used to gather satellite data, but on the other hand instrumental errors may be underestimated if the accuracy of operational soundings is inferior to that of the CEPEX sondes, which is probably true to some degree. The noise amplitudes $\sigma$ that were obtained by this method appear in Table 1.

Though the estimate of the full error-covariance matrix $\Gamma$ must be used with caution, many of the amplitudes $\sigma$ can be checked against other observations. For example, the noise levels in CAPE (Table 1), and in near-surface temperature and humidity (not shown), are smaller than the standard deviations of those variables in the main dataset by a factor of about 4. This ratio is broadly consistent with the results of comparisons between aircraft and buoy observations in CEPEX and the Tropical Ocean Global Atmosphere Comprehensive Ocean–Atmosphere Research Experiment (TOGA COARE), colocated within distances of order 100 km, in which the two platforms differed by much less than the overall variability observed (Fig. 4 of Serra et al. 1987). The noise levels found here in 1000-hPa humidity (0.65 g kg$^{-1}$) and WSPD (1.25 m s$^{-1}$) also agree with standard deviations of $\sim$0.2–0.3 g kg$^{-1}$ and $\sim$1 m s$^{-1}$ found during 80-km flight legs from COARE (Figs. 5 and 16 of Hagan et al. 1997), if random instrumental errors of $\sim$3% RH and 0.5 m s$^{-1}$ in the soundings are added in quadrature to the the reported natural variabilities.1 Mid- to upper-tropospheric humidity variability within the 40-km scale has been reported from COARE microwave measurements at roughly 3%–7% RH (Deeter and Evans 1997), which is consistent with the 9% value found here since sounding instrumental error probably grows to at least 5% RH at higher levels. That study also found that most variability was concentrated within 10-km and lower scales, which means that looseness here about the scale cutoff (40–200 km) should not affect the $\sigma$ estimates much. Thus the noise estimates used here, obtained from CEPEX soundings and representing both instrument error and submesoscale variability, seem consistent with variability estimates from other sources.

5. Further issues and experiment design

Before progressing to the use of (3)–(4) in statistical models, a few additional issues should be considered. These affect attribution, and are addressed by appropriate model design strategies.

a. Shear and nonlinearity

Previous results have found that a wind difference (between approximately 850 and 700 hPa) of about 20 m s$^{-1}$ is ideal for sustenance of simulated squall lines once they have begun (Rotunno et al. 1988). This means that convection may be aided by moderate shear values, but inhibited by too much shear, a situation that would be fitted poorly by a linear model.

There is a slight indication of this in the data (Fig. 2), though weak. Shear values in this light-wind region rarely approach 20 m s$^{-1}$, and the most helpful value of shear appears to be closer to 3–5 m s$^{-1}$, where some convective cases occur even in dry environments. In view of this, the shear-difference variable

$$\text{SHDIF} = 10\text{abs} \left( \frac{\text{WSHR}_{\text{low}} - 5 \text{ m s}^{-1}}{\text{WSHR}_{\text{low}} + 5 \text{ m s}^{-1}} \right)$$

(which is smallest for shear values near 5 m s$^{-1}$) is used henceforth as a predictor instead of the shear magnitude.
itself. SHDIF turns out to evince some predictive power over \( \Omega \) in the multivariate calculations shown below, while \( \text{WSHR}_{\text{low}} \) does not and is not considered further.

**b. Hidden causes, controlling, and gradual growth**

1) **DISCUSSION**

Next we address a subtle problem of statistical attribution. If not all of the influences on the target variable are included in \( \mathbf{x} \), and if some “hidden variables” (those which influence \( \Omega \) but are absent from \( \mathbf{x} \)) are correlated with components of \( \mathbf{x} \), then those components will be identified incorrectly as influences on \( \Omega \). For example, the strong influence of block-1 predictors on \( \Omega \) shown in Fig. 1 is probably not entirely because clouds beget more clouds, but rather because some convection-inducing circumstances (which are imperfectly captured by the soundings) persist long enough to contribute to both the previous and subsequent cloud coverage. Spurious attribution of this hidden influence to sounding variables is minimized in this study by including the earlier satellite counts in the regression, a strategy known as “controlling” the regression.

A related worry stems from the definition of \( \Omega \). A convective onset has been defined as the first time 208 K cloud coverage exceeds 10% (Sherwood and Wahrlich 1999). However, in some cases convection may be in an early building stage when the sounding is released, without having reached this threshold. Then, precursors that appear statistically relevant may in fact be effects of the early stages. In particular, the tendency toward high moisture in the low- to midtroposphere in \( \Omega = 1 \) cases may be a symptom of cumulus congestus moistening the troposphere before building to the deep-convective stage [although Sherwood and Wahrlich (1999) showed that this moisture is present for at least 4 h prior to outbreak, longer than the time usually taken for convection to develop]. It is important to establish whether the midlevel moisture is actually instrumental in causing the convection to continue building to the deep stage, or whether it is just an inert by-product of the earlier stages in the life cycle.

2) **MODEL DESIGN**

Two approaches to these problems were adopted. First, in addition to \( \Omega \), a more sensitive index \( \Omega_1 \) was considered where convective onsets were defined as the first time when any brightness temperatures (i.e., a single pixel) below 208 K appeared. Second, the linear analysis was performed separately on two partitions of the dataset: a “clear” partition, in which \( c_{235} < 5\% \), and a “cloudy” partition in which \( 90\% > c_{235} > 10\% \). These partitions contain 80% and 14% of the data, respectively (86% and 9% with \( \Omega_1 \)). Regression of \( \Omega_1 \) in the clear data should be immune from the effects described above, and should indicate which factors allow new convection to break out. Predictors that are related to \( \Omega \) in the cloudy data should indicate factors that are important in sustaining convection, allowing cumulus congestus or midtopped convection to grow deeper. In the regressions below, \( c_{235} \) and \( c_{208} \) were included as control variables.

6. **Results of further linear analysis**

Figure 3 shows the results from the total dataset for both convective indexes and all noise models, while Figs. 4 and 5 show results for the clear and cloudy portions separately. Results are shown as partial correlations (equal to the regression coefficients multiplied by the ratio of the standard deviation of the predictor to that of \( \Omega \)). Calculations with region sizes other than 120 km (not shown) look similar. Confidence intervals were estimated from a set of 600 “bootstrap samples” (Efron and Tibshirani 1993; Sherwood and Wahrlich 1999). The samples are generated by resampling the available data with replacement (including the noise samples, for the two models that use them), repeating the entire model-fitting procedure for each bootstrap sample, and plotting the 16th and 84th percentile values, equivalent to 68% confidence (one standard error). This confidence value was used as a threshold for keeping nonzero coefficients in each fit; the value was reduced from that in section 3 to improve the stability of the final results.

These further analyses still identify RH above the mixed layer as the most evident overall precursor to convective outbreak, even when the relatively higher noisiness of some of the other variables is taken into account in the measurement-error models. This conclusion is made stronger by the results of regressing \( \Omega_1 \) in clear cases (Fig. 4), those most immune from the interpretation problems discussed in section 5. If CAPE and CIN are abandoned as predictors, and replaced by the additional variables \( T_{\text{low}} \), RH\(_{\text{ML}} \), and \( T_{\text{ML}} \), forming a complete set through the troposphere (where the last two are means over the “mixed layer” 1000–900 hPa), the vertical distribution of thermodynamic influence may be investigated more directly. The result (shown in Fig. 6) shows a striking difference between clear and cloudy situations in the vertical trend of humidity influence, which reaches all the way to upper levels for clear conditions but is highly weighted toward lower levels in cloudy conditions. The model fit improves slightly with this predictor set. Both predictor sets indicate that moisture throughout most of the troposphere—not just in or near the boundary layer— aids convective development, all other things being equal.

The role of stability is harder to identify. In clear-sky cases, CAPE and CIN appear to play no significant role in regulating convection. In Fig. 5 they sometimes attain significance, but coefficients fluctuate in the experiments in a way that suggests at first that no conclusions should be drawn. However, the importance of low-level
parcel buoyancy in the cloudy cases appears clearly in Fig. 6. Much of the scatter in Fig. 5 is due to the high, negative correlations between CAPE and CIN, and between CAPE and $T_{\text{high}}$. Note that the difference between the weights attached to CAPE and CIN is about the same in all six panels (about 0.1). This indicates that low-level parcel buoyancy is in fact a significant promoter of convection, though the data were insufficient to determine unequivocally whether it was deep or shallow instability that really mattered.

Fig. 3. Standardized coefficients from regression of $\Omega$ onto the set of 10 sounding predictors. The 68% confidence intervals are shown as error bars. Top row, convective onset defined at 10% coverage of $c_{200}$; bottom row, onset defined at one pixel. Left column, standard model; middle column, model with constrained noise estimate $S'$; right column, model with minimum-variance noise estimate $S$.

Fig. 4. As in Fig. 3 but only for data with $c_{200} < 5\%$. 
In general, no wind variables showed influence on convection consistently through the experiments. The data suggest tentatively that stronger surface winds help activate convection, and (when predictor noise is taken into account) that shear values near 5 m s$^{-1}$ may aid convection. WSHR$_{\text{high}}$ tends to show inhibitive effects, as predicted from numerical models, but these effects are not really statistically significant in this study.

$T_{\text{high}}$ shows a tendency to aid convection in Figs. 3–4. Previous studies have found that large-scale, upper-tropospheric warm anomalies accompany increases in tropical convective activity. Mapes and Houze (1992) noted further that the hydrostatic pressure changes associated with these anomalies were maximal at the surface, where the warm temperatures appeared hypsometrically as lows (as opposed to highs in the upper troposphere, which might also have been expected). This finding suggested that the anomalies were associated with low-level convergence. In the analysis here, it would be hoped that the effects of low-level convergence would be observed directly (through CIN, etc.), and that $T_{\text{high}}$ would be found to be conditionally independent of $\Omega$. This does appear to occur in Fig. 6 to within experimental uncertainties, though we cannot rule out the possibility that positive $T_{\text{high}}$ anomalies are associated with convective promotion in some other way.

The overall results do not differ greatly from those found in midlatitude continental regression studies, despite the differing mean conditions (Peppler and Lamb 1989).
7. A nonlinear model

Though behavior of a nonlinear function \( f \) can be captured reasonably reliably by linear models as long as \( f \) is not too complicated, linear models may not be optimal here, as convection is well recognized to be a nonlinear process. For example, most popular convective and cloud cover schemes for numerical models employ switches at critical values of stability and humidity, respectively. Also, we have seen that the role of shear (though difficult to detect in the data), appears to be nonlinear.

The most common assumption about convection is that it requires deep instability and some other factor (often, low-level mass or moisture convergence)—a simple conjunctive (Boolean AND) relationship. Thus, it was decided to test conjunctive hypotheses of the form, “CAPE must exceed \( \xi \) and CIN must be less than \( \xi \) and . . . .” In formulating the hypotheses here, it was assumed a priori that CAPE, RH, and WSPD must be greater than a critical value for convection to be possible, while SHDIF and CIN must be less than a critical value. WSHR high and the temperatures were not included, but an upper-level lapse variable \( \Gamma_{\text{high}} = T_{\text{high}} - T_{\text{mid}} \) was included, which (based on the linear results) is assumed to allow convection above a critical value. In these hypotheses, convection was assumed to occur with probability \( P_i \), if all conditions are met, and with zero probability otherwise.

The task then amounts to finding the threshold \( \xi \) for each variable \( i \), and the conditional probability \( P_i \). This was done by finding the parameter values that maximize the likelihood of the data given a noise model. A detailed description of the method is given in the appendix.

Results

The threshold parameters found from this procedure are listed in Table 2. Variables not listed were found to be irrelevant, that is, a threshold was chosen that was large or small enough to include all points. Humidity variables again assume greatest importance, in a manner consistent with results from the linear analysis.

CIN was not found to be relevant for clear cases, and the threshold for cloudy cases was very weak, only excluding 4% of these cases. The threshold CIN value of 91 J kg\(^{-1}\) may seem unphysically large, since CIN values much less than this should be sufficient to inhibit updrafts with reasonable kinetic energies (a 2 m s\(^{-1}\) draft has only 4 J kg\(^{-1}\) of kinetic energy). It is necessary to remember, however, that the threshold is for the mean value of CIN over a mesoscale area. Localized values within this area may vary substantially and could reach zero somewhere even when the mean value is rather high. It is also possible that the variable is so noisy that the noise treatment here is simply inadequate to establish the value properly.

The most interesting result from this exercise is the low CAPE value required for convection to occur. The CAPE thresholds appearing in the table were exceeded 86% of the time in the dataset overall. This result is evident from the joint distribution of CAPE and RH\(_{\text{high}}\) in the clear dataset as shown in Fig. 7. Region I is the region hypothesized to allow convection (see appendix). We see that CAPE below 300 J kg\(^{-1}\) significantly reduces the likelihood of convection, since \( \Omega = 1 \) cases are relatively rare. But above \( \sim 300 \) J kg\(^{-1}\), the marginal distributions of CAPE among \( \Omega = 0 \) and \( \Omega = 1 \) cases are indistinguishable—in other words, CAPE growth beyond 300 J kg\(^{-1}\) has no further effect on the probability of convective outbreak. A nonlinear model is therefore more able to identify the role of this variable than is a linear one. This discrimination does not appear with CIN. The conjunctive modeling effort is, like the linear analysis, a very rough stab at a difficult problem, and the results should not be overinterpreted. The exact value of the CAPE threshold is admittedly uncertain, though clearly not much greater than 300 J kg\(^{-1}\) given the noise level established from the CEPEX data.

This result is significant since, despite widespread agreement that a stability threshold for convection exists, the exact value of this threshold depends on the thermodynamic path followed by convective parcels and

<table>
<thead>
<tr>
<th>Variable</th>
<th>Clear</th>
<th>Cloudy</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAPE</td>
<td>&gt;319</td>
<td>&gt;35 J kg(^{-1})</td>
</tr>
<tr>
<td>CIN</td>
<td></td>
<td>&lt;91 J kg(^{-1})</td>
</tr>
<tr>
<td>RH(_{\text{low}})</td>
<td>&gt;77</td>
<td>&gt;70%</td>
</tr>
<tr>
<td>RH(_{\text{mid}})</td>
<td>&gt;57</td>
<td>&gt;44%</td>
</tr>
<tr>
<td>RH(_{\text{high}})</td>
<td>&gt;24</td>
<td>—</td>
</tr>
<tr>
<td>RH(_{\text{high}})</td>
<td></td>
<td>—</td>
</tr>
<tr>
<td>WSHR</td>
<td></td>
<td>&gt;3.3 m s(^{-1})</td>
</tr>
<tr>
<td>( P_i )</td>
<td>0.15</td>
<td>0.42</td>
</tr>
</tbody>
</table>

Fig. 7. Plot of the clear dataset, CAPE vs RH\(_{\text{low}}\): \( \Omega = 0 \) (dots) and \( \Omega = 1 \) (crosses). The three labeled regions are discussed in the appendix. Observation error two-way standard deviation is shown as crosshatch.
is therefore not known from basic principles (Xu and Emanuel 1989). The uncertainty is significant, since CAPE can vary by ~2000 J kg\(^{-1}\) under different assumptions (e.g., Williams and Renno 1993). Thus it has been easy to compare the stability of two soundings, but difficult to ascertain whether or how often the atmosphere is truly unstable. The present analysis shows that the tropical west Pacific is in fact nearly always sufficiently unstable to support convection, other factors permitting.

8. Predictability and determinacy

So far we have concentrated on the attribution of convective outbreak to mean-field conditions. Another key problem is the overall degree to which the mean-field state determines whether convection will occur—in other words, how random convection is, or how much it depends on unobserved factors. These factors might include finer structure in the vertical than what has been included in \(x\); horizontal variability within the meso-scale; or other fields, like vertical velocity. Recall that we quantified the ability of mean fields to determine convection by the quantity \(\langle f \rangle_{\Omega} \) [see (2)]. For the conjunctive models, \(\langle f \rangle_{\Omega} = P_{i}\). The \(P_{i}\) values in Table 2 suggest that, for 3-h prediction, unobserved factors are still of dominant importance in determining whether convection will occur.

To obtain an equivalent estimate of \(\langle f \rangle_{\Omega}\) from the linear models, linear discriminant analysis (LDA) may be used to estimate \(f\). LDA involves fitting a multivariate Gaussian to the distribution of \(x\) within each class, centered on the sample mean for that class. Each class is fit with a Gaussian having the same covariance matrix. Bayes’ rule then gives the “posterior” probability that a sample with state vector \(x\) belongs to class \(j\) out of \(p\) classes (for the general case):

\[
Pr(\Omega = j|x) = \frac{Pr(x|\Omega = j)Pr(\Omega = j)}{\sum_{i=0}^{n} Pr(x|\Omega = i)Pr(\Omega = i)}.
\]

It turns out that the resulting posterior probability depends only on \(p-1\) linear combinations of the predictors called the discriminant functions, \(z_i\). In our two-class case there is only one discriminant function, \(z = \beta \cdot x\), which is identical (within a constant multiple) to the linear regression function obtained when \(\Omega\) is treated as the target variable. Many textbooks treat this subject (e.g., see Manly 1994).

Figure 8 shows the posterior probability of convective outbreak, plotted against \(z\) (top panel) and against the rank of \(z\) (bottom panel). To test the validity of the LDA model, the observations were divided into 20 equally populated bins according to their values of \(z\), and the sample proportion having \(\Omega = 1\) was computed in each bin, plotted as a square symbol. This comparison shows that the model’s probability assignments are consistent with the actual distributions. Also in the upper panel, the observational error in \(z\) \(\left(\beta^{-1} \Gamma \beta\right)^{1/2}\) is shown by a two-sided error bar. This uncertainty is significantly smaller than the spread in the probability function, indicating again that observational error is not capable of explaining very much of the lack of predictability. In these calculations, as with the conjunctive model calculations, satellite variables have not been included as control variables, so that the sounding variables are left on their own to explain \(\Omega\).

The lower panel of Fig. 8 shows the same curve plotted (dashed) against the rank of \(z\); in other words, the observations are equally spaced on the horizontal axis in order of increasing \(z\). The solid curve shows the results when separate models are fitted to the clear and cloudy data; (dot–dashed, dot–dot–dashed) estimated curves with separate fits and two noise estimates. Dotted line shows a perfect discriminator.
nal estimates $S$ and $S'$ for each class. This graph confirms that, at least in a linear model, noise is capable of explaining little of the observed stochasticity, regardless of noise covariance assumptions. Estimated probabilities do grow significantly in the most favorable environments (right side of the plot), and the noise is capable of explaining many of the convective instances among the least-favorable half of the dataset. However, the fundamental fact that much of the area under the curves lies well to the left of the 93.8 percentile line (the dotted line in the figure, which shows what the curve would look like for a perfect predictor) is not really changed by taking noise into account.

Computation of $\langle f \rangle_0 = P_c$ from (2) gives 0.28 for the best noise model, compared to the overall value of 0.24 from the conjunctive models (obtained by weighted mean of the clear and cloudy $P_c$ values). The predictability of the real $f$ might be better than those of our crude fits, but it should also be remembered that the performance of the LDA analysis on the data used to fit the model itself will always be at least a slight overestimate of its true performance. These predictability results are also specific to the 3-h time period and 100-km length scale. As longer forecast periods or larger regions are considered, the predictability should improve, though it is not clear how fast. Future work could address this question.

9. Conclusions

This study has tried to extricate the factors most responsible for regulating convective outbreak over the western Pacific region, by analyzing data. Many difficulties exist in following this path—difficulties that have been discussed here in some depth, but have not been overcome in a completely convincing way. Many variables give hints that they are important in differing circumstances, but in many cases, the uncertainties in the results and their interpretation make it dangerous to try to draw firm conclusions. Observation noise was found to have significant effects on attribution.

Nonetheless, there is strong evidence that moisture levels above the boundary layer—which are often not taken into account by models in deciding whether convection will occur—do in fact play a large role in regulating convective development, a conclusion also reached by Brown and Zhang (1997). Others have found similar results outside the Tropics (Burpee and Lahiff 1984; Peppler and Lamb 1989). Moisture appears particularly important in the early stages of convective growth (i.e., in regions mostly free of middle or high clouds at the time of the sounding), but stability and/or wind-induced moisture supply appear to become important in allowing shallow or midtopped cumulus activity to develop into deep convection. The latter conclusions could be made firmer by the consideration of a much larger dataset. These results overall support the tropical convective mechanisms argued by Raymond and Torres (1998). Due to the complexity of convective behavior, the results cannot necessarily be extended to other regions.

Conjunctive test analysis suggests that a minimum mesoscale-mean CAPE of 100–300 J kg$^{-1}$ is required for convection (calculated using a 1000-hPa parcel this threshold would be closer to 400–700 J kg$^{-1}$). This threshold is just a bit lower than the typical floor value attained after convective stabilization (Sherwood and Wahrlich 1999). There is no indication that further increases in CAPE have any impact on the probability of convective activation, once this threshold is exceeded. The requisite values are attained 86% of the time in the western Pacific region (according to observed estimates of CAPE noise and variability) so CAPE is not an important factor there in an overall sense, though such a condition would certainly be a more important constraint on convection elsewhere on the globe. This conclusion is important in light of previous uncertainty in the thermodynamic path typically followed by convective parcels, and concomitant uncertainty in required CAPE (Emanuel et al. 1994).

Previous studies have found relationships between the atmospheric state and convection to be weak in the Tropics (Zawadzki and Ro 1978; Mapes and Houze 1992; Alexander and Young 1992) but have not attempted to distinguish observational shortcomings from true stochasticity. In the dataset examined here, convection was 6.2% likely to break out during 3 h after a sounding was released (where only soundings not taken during convection were considered). Both linear and conjunctive models indicate that, taking observation noise into account in interpreting the data, the best that could be hoped for with perfect observations of the mean-field state is a typical assigned probability of around 20%–30% under favorable conditions. This result depends on the type of model used, and also relies on accurate estimates of noise amplitude, made here from a limited sample of sounding data, which may not represent all the problems inherent in operational data and launch sites. Nonetheless, the “predictability gap” found between the noise amount and the apparent stochasticity is large, strongly suggesting that observation noise and small-scale variability—though significant from the standpoint of attribution—are not adequate to explain the apparent stochasticity.

This result has implications for the representation of organized cumulus activity in models. Such representations are often based on the concept of rapid response (timescales of $\sim$1–2 h) to changing large-scale conditions. The results of Sherwood and Wahrlich (1999) confirmed that adjustment times as short as 2–3 h may be appropriate, once convection begins. However, the results here indicate that convective activation might be delayed by many hours after the attainment of appropriate large-scale conditions. To predict convective behavior more accurately on hour-to-hour timescales, additional variables would apparently be needed. It is very
likely that horizontal variability on small scales is the most important missing information. In squall lines, for instance, convective lines are often observed to form along narrow gust fronts, whose widths are tiny compared to 100 km. It is possible that prognosis of subgrid-scale variability or wave activity in coarse atmospheric models would be helpful to their simulation of convective effects.

Acknowledgments. The author wishes to thank Ralph Wahrlich for his work on the satellite data, and Eric Jones and Tony Plate for many helpful discussions, and the reviewers for useful comments. Sounding data were prepared and archived by the Data Support Section of the National Center for Atmospheric Research.

APPENDIX

Max-Likelihood Method for Conjunctive Hypothesis Fitting

It would be desirable to have some method—a nonlinear analog to multiple linear regression—to find nonlinear behavior or "feature interaction" objectively. Many such methods exist, but they work best with large amounts of noise-free data. One that was tried as part of this study was the c4.5 decision tree system (Quinlan 1993). Unfortunately, this algorithm was unable to find portions of predictor space in which convection consistently broke out—except small fringe regions, encompassing few cases, and inspiring little confidence. This was either because the algorithm could not directly model noise in the predictors, or because the problem is just too stochastic. A different approach was clearly needed.

The basic problem with nonlinear statistical models is the enormous number of possible feature interactions that they must consider. A better way to proceed with limited data is to restrict consideration to a small number of simple hypotheses that make sense based on existing, physically based knowledge. Another advantage of such an approach is that explicit modeling of the observational error (which is absent from c4.5 and all other nonlinear methods in common use) becomes tractable.

The method described here tests a hypothesis that each variable must be less than or greater than a particular threshold value for convection to be possible. If all conditions are met, convection is assumed to occur with probability \( P \), otherwise it cannot occur. The set of threshold parameters \( \xi \) is adjusted until the likelihood of the data given the model is maximized, a so-called maximum-likelihood approach. In principle, we could test more general hypotheses such as, "(condition 1 or condition 2) and condition 3 . . . ," but a thorough likelihood search of such hypotheses did not seem warranted by the size of the dataset.

To evaluate the likelihood of the data given a particular \( \xi \), several steps are required. First, the distribution of the set of observations \( x \) (the true state) must be estimated. This was done by subjecting the observed dataset to an affine transformation, shrinking it toward its mean by multiplication by a constant fraction in each dimension until its new variance had shrunk by an amount equal to the noise variance \( \sigma^2 \) in that dimension. This produces a sample distribution of \( x \), which is as similar as possible to that of the observations \( x' \), but has univariate statistics consistent with those of \( e \) and \( x' \). We assume here, as before, that the noise \( e \) has zero mean and is uncorrelated with \( x \).

Next, the rescaled dataset was sampled randomly with replacement 100 000 times to generate a synthetic "true" dataset. To each point in this set was added a noise vector, also sampled randomly with replacement from the set of noise samples, yielding a synthetic "observed" dataset. Additionally, 1% of the synthetic observations were completely randomized, to represent the effects of non-Gaussian noise, which occasionally happen due to catastrophic sensor failure, reporting errors, or other possible problems. This prevents models from being rejected outright just because of, say, a single outlier in the data.

Given a proposed \( \xi \), each synthetic sample can be assigned a value of \( \Omega \). Those whose true value \( x \) does not meet the convective conditions were assigned \( \Omega = 0 \). Samples meeting the convective conditions were flagged and counted, and their number was divided by the expected number of \( \Omega = 1 \) points to give the estimate \( P \), consistent with the conditions and the distribution of \( x \). Finally, a fraction \( P \) of the flagged points (chosen at random) were assigned \( \Omega = 1 \), the rest \( \Omega = 0 \).

The likelihood of the data was then estimated by dividing the state space \( \Omega \) into three regions, indicated graphically in Fig. 7: I, the interior region where all convective conditions are met; II, an intermediate region lying within 1.5 \( \sigma \) of the first region in any dimension; and III, the rest. Each region has a synthetic (i.e., expected) and actual population of observations. The expected population of \( \Omega = 1 \) cases in region III is very small, since extreme values of noise are required for such samples to be found there. The likelihood of the actual set of observations \( X' \) given the model is

\[
Pr(X' | \xi) = \prod_j Pr(X'_j | \xi) \tag{A1}
\]

\[
= \prod_{q=1}^3 \left[ \frac{n_q(0)}{n_q(0) + n_q(1)} \right]^{N_q(0)} \left[ \frac{n_q(1)}{n_q(0) + n_q(1)} \right]^{N_q(1)}, \tag{A2}
\]

where \( j \) is an index running over all the observations; \( q \) is an index running over the three regions; \( n_q(0) \) and \( n_q(1) \) are the simulated populations of observations in region \( q \) having \( \Omega = 0 \) and \( \Omega = 1 \), respectively; and \( N_q(0) \) and \( N_q(1) \) are the corresponding actual populations. A true calculation of the likelihood function from (A1) would require integrating the conditional likelihood over the entire synthetic dataset for each observed.
point, a computationally unfeasible task, but (A2) is a useful approximation that gave credible results when tested on single-variable cases that could be examined by eye.

The maximum of this likelihood function was found using Powell’s method as implemented by Press et al. (1992).

REFERENCES


