Detection of the ITCZ in the east Pacific using
Markov random fields on instantaneous satellite data

CAROLINE BAIN *, JORGE DEPAZ, JASON KRAMER,
GUDRUN MAGNUSDOTTIR, PADHRAIC SMYTH,
HAL STERN AND CHIA-CHI WANG†

University of California, Irvine, USA

*Corresponding author address: Caroline Bain c/o Gudrun Magnusdottir, Department of Earth System
Science, University of California, Irvine, CA 92697.

E-mail: gudrun@uci.edu

†now at: Department of Atmospheric Sciences, Chinese Culture University, Taipei, Taiwan
ABSTRACT

A Markov random field (MRF) statistical model is introduced, developed and validated for detecting the east Pacific Intertropical Convergence Zone (ITCZ) in instantaneous satellite data during the summer half year. The MRF statistical model uses satellite data at a fixed point as well as its neighboring points (in time and space) to inform the decision as to whether the given point is classified as ITCZ or non-ITCZ. Two different labels of ITCZ occurrence are produced. IR-only labels result from running the model with 3-hourly infrared data available for the 29 yr period, 1980–2008. Data-all labels result from running the model with additional satellite data (visible and total precipitable water) that are available from 1995. IR-only labels result in less area of ITCZ than Data-all labels, especially where the ITCZ is shallower. Yet, qualitatively, the results for the two sets of labels are similar.

The seasonal distribution of the ITCZ through boreal summertime is presented, showing typical location and extent. The ITCZ is mostly confined to the eastern Pacific in May, and becomes more zonally distributed towards September and October each year. Northward shifts in the location of the ITCZ occur in line with the seasonal cycle. The ITCZ is highly variable on inter annual time scales and highly correlated with ENSO variability. When we removed the ENSO signal from labels, inter annual variability remained high. The resulting IR-only labels, representing the longer time series, showed no evidence of a trend in location nor evidence of a trend in area for the 29 year period.
1. Introduction

The Intertropical Convergence Zone (ITCZ) is one of the most recognizable aspects of the global circulation as seen from space. It represents the rising branch of the Hadley circulation, forming a zonally elongated band of convection and cloudiness at low latitudes, where the northeasterly and southeasterly trade winds converge. Previous observational studies of the global ITCZ (e.g. Mitchell and Wallace 1992; Waliser and Gautier 1993) focused on the annual cycle in different regions. They found very distinct longitudinal variations in the ITCZ such that in the Indo-western Pacific region the summer ITCZ is broad in latitude and ill-defined due to the extensive warm pool in the ocean and monsoonal circulations. However, in the east Pacific the mean summer ITCZ is narrow and long, generally located at the southern boundary of the eastern Pacific warm pool, north of the strongest meridional gradient of sea surface temperature (Raymond et al. 2003). The summer half year (May through October) is the period during which the east Pacific ITCZ is particularly strong when present. During the winter half year the ITCZ remains in the northern hemisphere, but its signature is considerably weaker and gets mixed in with signatures of extratropical frontal systems due to cold air outbreaks (Wang and Magnusdottir 2006, hereafter WM06). The focus of this paper is to introduce a new method to detect the ITCZ using high temporal and spatial resolution satellite products that have recently been archived. We apply the method in the Pacific region east of the dateline.

The picture of the seasonal (order of 100 days) ITCZ in the east Pacific is quite different from the synoptic (order of 1–20 days) ITCZ. From day to day during the summer half year, the ITCZ is highly dynamic and changeable as seen in visible (VS) and infrared (IR) images.
from the Geostationary Operational Environmental Satellites (GOES), which depict cloud reflectivity and cloud-top temperature, respectively. The ITCZ may be quite narrow (2° latitude) and stretched over an extensive longitudinal distance (up to 70°) for hours to days, until a part of its structure (usually the easternmost part) or the entire ITCZ structure breaks down into individual disturbances that may move away or dissipate in place. The ITCZ usually reforms within a day or two of breakdown (WM06). This variability in the ITCZ presents a serious challenge to its automatic detection in data. WM06 present a detailed study of daily variability of the ITCZ in the east Pacific for five years, 1999-2003, using visual inspection of satellite images in combination with the field of 975 hPa relative vorticity from the NCEP Global Forecast System (GFS) analysis and relative vorticity calculated from QuikSCAT scatterometer surface winds. Here we extend their study by presenting a new method to automate the detection so that a greater number of active seasons may be analyzed.

Recent studies (e.g. Peters et al. 2008; Back and Bretherton 2006) discuss the importance of eddy transport of energy and water vapor for tropical circulations, where there is a well-defined narrow convergence zone with the accompanying shallow convergence layer, such as occurs in the east Pacific. According to energy budget calculations, eddy activity appears to drive the narrow convergence zone. Even though the importance of tropical eddies has been suggested it remains to define and quantify the tropical eddy activity. Furthermore it remains to examine what role extratropical eddy activity may play in driving the tropical activity. To study these synoptic scale interactions, a data set of ITCZ location and extent with synoptic temporal resolution is required. This has previously not been available.

Organized synoptic-scale variability in the east Pacific may be described in terms of ITCZ
variability and variability of westward propagating disturbances (WPDs; Magnusdottir and Wang 2008), which are taken to include isolated tropical disturbances ranging from easterly waves to tropical cyclones. ITCZs and WPDs are intricately related yet quite different in character. Both have the signature of a local maximum in lower tropospheric relative vorticity. The ITCZ has the signature of a zonally elongated vorticity strip, whereas the vorticity signature of WPDs is more axisymmetric in shape especially for the stronger cases corresponding to an isolated vortex. WPDs may lead to the breakdown of the ITCZ. WPDs may also form during ITCZ breakdown from the merging of the isolated vorticity pools that remain after the main structure has dissipated (WM06; Kieu and Zhang 2008; Davis et al. 2008). In some instances weak WPDs have been found to propagate through a well-formed ITCZ (Scharenbroich et al. 2010), deforming it in the process without breakdown. East Pacific WPDs have been found to develop as a result of tropical waves entering the region, disturbances that have propagated from the Caribbean but may have originated as far away as Africa (e.g. Avila 1991).

Due to the inherent variability of the synoptic ITCZ with its highly variable time and space evolution, it is difficult to present a composite picture of this phenomenon even though it is easily recognizable by the human eye in visible and infrared satellite images. This was partly overcome in Magnusdottir and Wang (2008) who used spectral analysis of 850 hPa relative vorticity calculated from daily averaged ERA-40 reanalysis winds spanning 1979–2001 to filter the signal of interest in a wave-number frequency diagram. Even though this method cannot separate WPDs out of the ITCZ signal, the east Pacific is dominated by ITCZ variability (as opposed to the Atlantic region). Others have used thresholding of brightness temperature in combination with high reflectivity to identify the ITCZ (e.g.
Waliser and Gautier 1993). A certain maximum temperature threshold is chosen to represent the warmest cloud top temperature within the ITCZ, such that temperatures below this threshold are considered to represent clouds that are part of the convection within the ITCZ. To filter out smaller systems an area limit is often imposed so that cloud clusters must be of some minimum area. Little attempt has been made to date to group the colder clouds into a continuous zonally elongated feature, nor to remove noise and accommodate shallower convection such as takes place in the narrow convergence zone of the east Pacific (Zhang et al. 2004).

In this paper we introduce and validate a Markov random field (MRF) model that is used to identify the ITCZ as binary labels (identifies the presence or absence of ITCZ) in the tropical east Pacific, north of the equator. The model was developed at UC Irvine for automatically identifying the ITCZ region in high-resolution satellite data without human intervention. Missing or incomplete data are easily accommodated as are shifts in the number of input variables available at each time point.

The statistical model introduced here also has the advantage that it can be adapted for applications elsewhere. It would also be appropriate for use in other regions where a narrow ITCZ tends to form such as the Atlantic and the east Pacific south of the equator, as well as in the region of the South Pacific Convergence Zone (SPCZ).

This paper is organized as follows: Section 2 describes the satellite data sets. Section 3 introduces the methodology, including the manual labelling of satellite data for use as training data, the MRF statistical model for the ITCZ, and inference of the presence/absence of ITCZ in data. Section 4 evaluates our approach to ITCZ detection. In Section 5 we present results for the east Pacific for two different outputs from the model: a longer (1980–2008)
time series of labels obtained using IR data only and a shorter (1995–2008) time series of
tables obtained using all three sources of data (IR, VS and total precipitable water, TPW).
The results focus on depicting the mean position of the ITCZ through the summer half year
as well as inter annual variability. Finally, Section 6 presents concluding remarks.

2. Satellite Data

Satellite IR, VS, and TPW images from 90 to 180°W were used for analysis. The VS and
IR data come from radiometers on geostationary satellites measuring in the visible and the
infrared-window part of the spectrum, respectively. The VS data are a measure of reflectivity
of clouds, the IR data correspond to cloud-top temperature. The TPW data give the water
vapor content of the entire column in units of equivalent thickness of liquid water (0–75
mm).

The IR and VS data sets were obtained from the National Climatic Data Center (NCDC)
of NOAA. These two data sets are from the HURSAT Basin data (Knapp 2008) and are
closely related to HURSAT B1 data (Knapp and Kossin 2007), which are derived from ISCCP
B1 data. The data were collected from different geostationary satellites. The IR channel
data were recalibrated to reduce inter-satellite differences. The HURSAT Basin data have
the same spatial and temporal resolution as HURSAT B1 data, but are assembled, gridded
and archived for each tropical ocean basin separately1.

The IR data set is available every three hours (00 UTC etc.) from 1980-2008 at a spatial
resolution of 10 km. We chose to reduce the spatial resolution to 0.5° in order to reduce the

1for more information see www.ncdc.noaa.gov/oa/rsat/hursat
number of pixels and decrease computation time. Because of the longitudinal extent of our area of interest (90-180°W), the entire area is in daylight only once per day (21 UTC) even though the data are available every three hours. Thus the VS data set is only useful once per day at the same spatial resolution as the IR data, which we similarly coarsened to a 0.5° spatial resolution. We also found that the earlier parts of the record were problematic for the VS channel. We therefore chose to only use the VS data set for the time period 1995–2008.

The TPW data set was assembled and archived by Remote Sensing Systems\(^2\). The data set is a composite of available microwave data (from the SSM/I, TMI, AMSR-E satellites). It covers 1987–2008, however the earlier part of the record is at lower temporal resolution and we chose to only use the data set for 1995–2008. The data set is at a spatial resolution of 0.25° that we coarsen to 0.5° for computational efficiency. The images are from 00, 06, 12, 18 UTC and contain data from +/- 3 hours, in practice making them +/- 9 hours resolution at each recorded time point. Furthermore, we applied linear interpolation between the time points, making TPW available at the same time as IR.

To summarize, one data set, IR, is available from 1980–2008. The other two data sets, VS and TPW, are used for the latter part of the time period or 1995–2008. We therefore chose to run our analysis of the east Pacific ITCZ twice. First using all three satellite data sets for the period 1995–2008. Secondly, we ran our analysis for the entire time period, 1980–2008, using the IR data only.

Since there are two versions of the model with different inputs, it is critical that we have consistent terminology to identify the output (i.e. estimated ITCZ labels). The term ‘All-data labels’ refers to the set of ITCZ labels obtained from the MRF model using IR, VS

\(^2\)www.remss.com
and TPW satellite data as input. The term ‘IR-only labels’ refers to the set of ITCZ labels obtained from the MRF model using only IR for input.

3. Method description

The statistical model relies on a probabilistic framework that incorporates observed satellite data, a latent Markov random field that represents the presence/absence of the ITCZ, and ‘a priori’ information about the most likely spatial location of ITCZ formation. Formally let $X_{ijt}$ serve as an indicator of the presence ($X_{ijt} = 1$) or absence ($X_{ijt} = 0$) of the ITCZ at longitude grid point $i$, latitude grid point $j$, and discretized time point $t$. $X_{ijt}$ is not directly observable and must be inferred from the satellite data and other information for all grid and time points, $i = 1, \ldots, I; j = 1, \ldots, J; t = 1, \ldots, T$. The observed satellite data at a particular grid point at time $t$ is denoted $Y_{ijt}$, a 3-dimensional vector comprised of satellite determinations of IR, VS and TPW. Not all satellite readings are available at all time points but the probability model can accommodate missing measurements. Our approach is to use the probability model to infer the probability distribution for the ITCZ indicators $X$ (for all grid points at all times) conditional on the satellite data $Y$, that is $P(X|Y)$. This posterior distribution provides estimated ITCZ fields along with a measure of uncertainty regarding the presence/absence and extent of the ITCZ. Results are obtained by sampling from the posterior distribution:

$$P(X|Y) = \frac{P(Y|X)P(X)}{\sum_X P(Y|X)P(X)} \quad (1)$$
where $P(Y|X)$ describes the distribution of the satellite data given the ITCZ indicators and $P(X)$ is a prior distribution on the ITCZ indicators that incorporates spatial information. We describe the elements of the model, our computational approach, and other implementation details in the remainder of this section.

a. Learning the distribution of satellite data, $P(Y|X)$

Application of the model requires knowledge of what values of the satellite fields are characteristic of ITCZ and non-ITCZ events. This information was obtained from training data provided by manual labelling of satellite data. Three meteorologists identified closed ITCZ regions (if any) at 2100 UTC for each day during August 2000. The labellers were provided with a graphical user interface and a tablet board with stylus to maximize precision when labelling the region desired. They were given VS, IR and TPW at times t-2, t-1, t, t+1 and t+2 where t was the time-frame being labelled. The individuals using the tablet board were told to label based on the following simple principles: (1) the ITCZ is a predominantly zonal feature. (2) Without a majority region of clouds we do not define an ITCZ. (3) We can also define an ITCZ where there are low level clouds. (4) Westward propagating disturbances including tropical cyclones are not part of the ITCZ unless they are embedded in a larger cloudy region.

The set of manual labels which performed best against the union of labels (Person 1, see table 1) was used to train the statistical model. The mean vector and covariance matrix for VS, IR, and TPW, for pixels belonging to the ITCZ ($X=1$) and for non-ITCZ pixels ($X=0$) was created using the manual labelled data. These means and covariances were then used as
the parameters for two Gaussian distributions for the satellite data $Y$, conditioned on $X=1$ and $X=0$. It is assumed that given the values of the indicators the observed satellite pixel values are independent across different pixels and times, so that

$$P(Y|X) = \prod_{ijt} P(Y_{ijt}|X_{ijt}). \quad (2)$$

b. A Markov random field model for the ITCZ, $P(X)$

The Markov random field (MRF) is a mathematical model (Kindermann and Snell 1980; Geman and Geman 1984; Li 1994; Smyth 1997) that formally specifies an a priori (that is pre-data), a probability distribution on the spatial characteristics of the ITCZ field. The known properties of the ITCZ that we would like to incorporate are typical spatial locations for the convergence zone and the informal notion that if one pixel is part of the ITCZ then this ought to increase the likelihood that neighboring pixels are also part of the ITCZ.

An MRF model is characterized by the following conditional probability relation; the probability distribution of the value of a single pixel, $X_{ijt}$, conditional on all other pixels (at this and other times) depends only on pixels that are immediate neighbors of the pixel. If we define $N_{ijt}$ as the set of neighbors of $X_{ijt}$ (neighborhood to be described in more detail below) and $X_{-ijt}$ as the set of all pixels except $X_{ijt}$, then the Markov property says that

$$P(X_{ijt}|X_{-ijt}) = P(X_{ijt}|N_{ijt}). \quad (3)$$

The neighborhood that we use here includes the immediate neighbor in the longitude ($X_{i+1,j,t}, X_{i-1,j,t}$), latitude, ($X_{i,j+1,t}, X_{i,j-1,t}$), and time ($X_{i,j,t+1}, X_{i,j,t-1}$) directions. The MRF builds in a preference for neighboring pixels to share common values with the strength
of that preference represented by one or more parameters. The MRF can also incorporate a bias or tendency for a particular pixel to be part of the ITCZ; we use this feature to accommodate historical information about the spatial regions that are likely to be part of the ITCZ. For this we specify values \( q_{ijt} = q_{ij} \), between 0 and 1, that characterize the likelihood that pixel \( ij \) is part of the ITCZ based only on its location with no other information. The conditional probability that characterizes our MRF is written as

\[
P(X_{ijt} = 1|X_{-ijt}) \propto \exp(\beta_h I(X_{i-1,j,t} = 1) + \beta_v I(X_{i,j-1,t} = 1) + \beta_t I(X_{i,j,t-1} = 1) + \beta_s \log(q_{ij}))
\]

where \( \beta_h, \beta_v, \beta_t, \beta_s \) are parameters introduced to measure the strength of the information conveyed in the zonal direction, meridional direction, neighboring times, and spatial location respectively. The probability is written as proportional to the specified term because calculating the actual probability requires computing a similar expression for the probability that \( X_{ijt} = 0 \) and then normalizing so that they sum to one.

The MRF prior distribution thus depends on a number of parameters (the \( \beta \)'s) and on the specification of spatial information about likely locations for the ITCZ (\( q_{ij} \) or 'spatial prior'). Based on some experimentation with a small sample of training data we set all \( \beta \) values equal to one. Optimization of the \( \beta \)'s, or perhaps estimating the \( \beta \)'s from training data, is a possible direction for future research.
For the spatial prior, we used a longitudinally constant strip with Gaussian variability in latitude. Figure 1a shows the study region. Figure 1b shows the two spatial priors used with the two different analyses of the ITCZ. The spatial prior takes its maximum probability, $q_{ij} = 0.25$ for all latitudes between 7.5-10°N for the All-data labels. The values decay as a Gaussian curve, with a variance of 3 degrees, from the maximum to its minimum value, $q_{ij} = 0.05$. Including TPW as an input tends to link up the convective regions, therefore higher values are required for the spatial prior used to create the IR-only labels. In this case, the spatial prior maintains the same shape as before but the maximum value is increased to $q_{ij} = 0.45$. These values were obtained through a series of sensitivity studies during model validation. The purpose of this spatial information is to assert our prior knowledge that the ITCZ is a zonal feature with stronger associations in the zonal direction than in the meridional, and is more likely in the southern region of the satellite frame.

The previous paragraphs develop a probability model for the conditional probability that a single pixel is part of the ITCZ given the status of its neighbors. A crucial theorem, the Hammersley-Clifford Theorem (see, for example, Besag 1974), guarantees that there is a single joint distribution $P(X)$ for all of the pixels (known as a Gibbs distribution) that is consistent with the conditional probabilities specified for each pixel. This joint distribution can’t be easily written in closed form because the normalizing constant (i.e., the term that ensures the distribution sums to one) involves a sum over all possible configurations for all pixels. Fortunately determining the posterior distribution $P(X|Y)$ does not require having the joint distribution.
c. Inferring the presence/absence of ITCZ, $P(X|Y)$

We take a Bayesian approach to obtain data-based inferences regarding the ITCZ. The training-data based probability model for satellite data given ITCZ status $P(Y|X)$ is combined with the MRF prior distribution on ITCZ status $P(X)$ to obtain the posterior distribution for the ITCZ status $P(X|Y)$ (Gelman et al. 2003). The most efficient way to draw these inferences is via a Markov chain Monte Carlo (MCMC) algorithm known as Gibbs sampling (Geman and Geman 1984; Gilks et al. 1993). Initial values are first generated for $X$ (each $X_{ijt}$ is set to either 0 or 1) based on the satellite data. This is done by randomly assigning $X_{ijt}$ to be 1 with probability

$$P(Y_{ijt}|X_{ijt} = 1)$$

and 0 otherwise. This is an initial approximation based on the satellite data. These initial values are then updated sequentially with each pixel updated conditional on the values of its neighbors. This is done via a conditional probability calculation that is analogous to the MRF probability given in equation (4) except that it now includes a term for the satellite data,

$$P(X_{ijt}|X_{-ijt}, Y_{ijt}) \propto P(Y_{ijt}|X_{ijt}) \times P(X_{ijt}|X_{-ijt}),$$

where the second term on the right is the right hand side of equation (4). Gibbs sampling requires cycling through all pixels to update their values repeatedly using the most recently obtained values for all neighboring pixels. Iterations through all pixels are repeated until the resulting Markov chain is found to have converged to stationary behavior. This stationary
behavior reflects the desired distribution $P(X|Y)$. We found that 200 iterations was sufficient to obtain stationary behavior. We take the next 50 simulations as representatives of the posterior distribution. These 50 simulations are used to define an estimated ITCZ by including all pixels that are judged to part of the ITCZ in 50% or more of the simulations. The simulations also provide information about the probability of ITCZ being present at each pixel.

The calculation described here is computationally intensive. Fortunately there are features of our model that allow speeding up this process. Most importantly for the neighborhood structure that we are using the data can be viewed as a 3-dimensional "chess" board with each pixel as a square and time serving as the third dimension. The Gibbs sampling update for the "black" squares depends only on the status of neighboring "white" squares and vice versa. This means that we can use a vector calculation to simultaneously update all of the "black" squares in a single step rather than use pixel-by-pixel updating. Parallelizing the computations of disparate time points is also feasible, although not explored in this paper.

d. Post Processing

Post processing is performed on the model output as a form of quality control. The post processing was designed to eliminate any noise not removed by the MRF since we are looking for a large continuous feature. By using a post processing algorithm to remove noise we reduce the need for more iterations and therefore increase the efficiency of the modeling process. Post-processing on the ITCZ labels is applied in three steps: (1) ITCZ regions are
closed by dilation and erosion\textsuperscript{3}, joining disjointed areas and smoothing edges, (2) Any holes are filled in as it is assumed that if a region is surrounded by ‘ITCZ’ then it also must be part of it, and (3) any identified ‘on’ regions that are smaller than 100 pixels (approx 5° by 5°) are removed.

4. Evaluation of ITCZ detection

Figure 2 is an example of an ITCZ label automatically identified by the statistical model using all three types of satellite data as input. The label incorporates regions of variable cloud top temperatures and excludes regions that are statistically unlikely to be within the ITCZ, as defined by the MRF model. Assessing the validity of this label presents a challenge since the ITCZ has not been definitively identified as a binary signal in instantaneous data. Here we evaluate ITCZ labels from the model against the performance of manual labellers and against thresholding techniques. Three independent manual labellers identified the ITCZ as a discrete weather feature at 2100 UTC every day for August 2000. Data from Person 1 were used to create the probability distribution $P(Y|X)$ for the satellite data (see Section 3a), as that labeller appeared closest to the “consensus” (i.e., that labeller scored the least error against the union of the other labellers).

In addition to the human labellers, we used two alternative ITCZ proxies to compare to: thresholded IR and thresholded TPW. Different values of thresholds were tested until an optimum value (which minimized errors in comparison to the union of manual labellers) was found. For IR this was a cloud-top temperature threshold of 270 K, for TPW this was

\textsuperscript{3}Documentation on erosion and dilation algorithms at www.mathworks.com/access/helpdesk/help/techdoc
a value of 53 mm. Post-processing was performed on the thresholded returns such that only
cloud systems exceeding 200 km in perimeter were used (approximately similar to Garcia
1985).

Figure 3 can be used to describe how the labels were evaluated. A base method considered
to be ground truth and a test method were defined. The test method was scored against
the base method by identifying how many pixels were wrongly identified. Referring to the
diagram,

\[
\text{False negative} = \frac{C}{A + B + C + D},
\]

\[
\text{False positive} = \frac{B}{A + B + C + D},
\]

\[
\text{Total absolute error} = \frac{B + C}{A + B + C + D}.
\]

If the base method and the test method are similar then all three scores are low. In addition
to these error measures, a measure of agreement of overlap that is not affected by the size
of the domain is

\[
\text{intersection/union} = \frac{A}{A + B + C}.
\]

This measure is 100 for perfect agreement and 0 when no pixels are identified by both
methods as ITCZ (no overlap).

Scores were calculated for each image in the evaluation data set and table 1 shows the
mean of these errors over the whole month of August 2000. The first block is a comparison
of each manual labeller against a union of the other labellers. For example Person 1 labels
are compared to the union of labels from Person 2 and 3. The second block is a comparison
between automated methods and the union of labels from Person 1, 2 and 3. The final blocks
of the table compare the various automated methods to each other. The results are obtained
by comparing labels in the region from 0 to 20°N and 170°W to 100°W. It was found that
the manual labellers were unlikely to label to the edges of the full domain as they could not
see past them, so a narrow comparison region was a more reasonable choice.

One limitation of the error rates is that they can be hard to calibrate; a low score is good
but it is not obvious how to recognize an unacceptably high score. For example, a total error
score of 100 could only be obtained if the test image were the negative image of the base
method image, which is not at all likely. To put the scores in perspective, the error scores
for ‘all domain defined as ITCZ’ and ‘all domain defined as non-ITCZ’ are also given.

The first three columns are the error scores, where a lower score indicates more skill
for the method tested. The results show that overall All-data labels are more similar to
manual labels on all measures than either thresholding method. Thresholded IR generally
underestimates the ITCZ region and thresholded TPW generally overestimates. The IR-only
labels somewhat underestimate the ITCZ region, but not by as much as thresholding IR.

The intersection/union score is shown in the final column of Table 1. For this column a
higher score indicates more skill. Again, the All-data labels score the best in comparison to
the other automated methods. The IR-only labels are an improvement on thresholded IR,
but do not score as well as thresholded TPW. The bottom section of table 1 compares the
various automatic methods with each other. The labels generated by the MRF model are
more similar to each than the thresholding techniques are to each other. The model labels
seem to capture much of the information in the thresholding approaches (they seem to agree
fairly often) and as described above do better at matching the human labellers.

Therefore, by these evaluations, the ITCZ labels generated by the statistical model per-
form better than any other automated techniques against manual labels by experts. This
justifies using the statistical model output to assist in studying the climatology of the ITCZ.

5. Results: climatology of the ITCZ

The labels of ITCZ show much inter annual, intraseasonal and synoptic scale variability, emphasising the complex dynamic nature of the weather feature. In this Section we present the composite picture of the ITCZ from All-data labels (available 1995-2008), to provide the basic climatology and inter-annual variability of ITCZ. We use IT-only labels (available 1980-2008), to investigate long term climatic trends.

Figure 4 shows the mean location of the ITCZ obtained from the sets of ITCZ labels for the same time period (1995-2008). The plot shows that qualitatively the labels show similar locations, but they are quantitatively different. The ITCZ is diagnosed less when only IR is used as input. The correlation between the two fields is 0.989. The All-data labels produce a field that is on average 1.3 times the field obtained from the IR-only labels in the region 0-15°N (the agreement is higher in the northern portion where there was no ITCZ present). The IR-only labels give a more conservative estimate of ITCZ because VS is not included, which may discount low cloud. In addition TPW, which is a smooth field linking up convective regions is not included. Despite this, Fig. 4 demonstrates that the mean location is reproduced for both sets of labels, giving confidence that statements made about general characteristics from averaged fields are consistent across the label data sets generated by different approaches.
a. Seasonal evolution of the ITCZ

The time evolution of mean ITCZ area through the season from the All-data labels for 1995-2008 is shown in Fig. 5. The time series was smoothed using a running mean with a 7.5 day window. The standard deviation was approximately \(8 \times 10^5\) km\(^2\). The domain was split into three longitudinal boxes (shown in Fig. 1a) to highlight the differences in ITCZ development across the Pacific region.

The plot shows the dominance of cloud in the east Pacific for most of the season. The area of the ITCZ in the 90-120\(^\circ\)W region (bold curve) has two peaks, in late June and late August and declines in September and October. The ITCZ area in the 120-150\(^\circ\)W region (dashed curve) peaks in July and August and decreases after this. The area of ITCZ in the 150-180\(^\circ\)W region (thin curve) increases steadily from May to October. The plot indicates that the ITCZ is most present in the eastern Pacific at the beginning of the season and becomes more zonally distributed as the summer months progress.

Part (b) shows the time evolution of mean area of cloud associated with tropical cyclones for the whole domain, found using the algorithm described in Appendix A. Tropical cyclone cloud peaks in August and September, coinciding with a recovery in the decline of ITCZ amount in the east Pacific (the double-peak effect). Therefore it is likely that some of the ITCZ defined in the east Pacific at this time is attributable to tropical cyclone activity, which is mostly confined to these longitudes. Another reason for a recovery in the decline in the ITCZ in late August may be the increase in the number of WPDs associated with African Easterly Waves, which peak annually in August and September (Thorncroft and Hodges 2001).
This image was reproduced using the IR-only labels as well as the thresholded IR and thresholded TPW data sets (not shown). This confirmed that the features observed were consistent between data sets. In spite of some quantitative differences, the plots were qualitatively similar: the double-peak in the eastern Pacific cloud area was always detected as was the westward spread of ITCZ activity towards the end of the season. The main differences between different labels were in the central region (120-150°W) where clouds are lower and warmer and are therefore less likely to be detected in labels that only rely on IR data.

Figure 6 shows monthly anomalies from the annual mean of ITCZ location (shown in Fig. 4a) overlaid on the monthly mean sea surface temperature (SST). The median anomaly is used rather than the mean anomaly to reduce the influence of the El Nino year in 1997, when the ITCZ was located particularly far south towards the end of the season.

The plot demonstrates the zonal extension of the ITCZ by the end of the season. Positive anomalies can be seen in the eastern Pacific in June and July; these become negative in September and October. Positive anomalies were found in the western region in September and October demonstrating the westward redistribution of the ITCZ. The anomalies also show the northwards migration of the ITCZ, with positive anomalies to the south of the domain in May and October and positive anomalies to the north in mid-summer. This seasonal shift in location is consistent with the northward migration of warm SSTs.

Raymond et al. (2006) stated that ITCZ formation in the east Pacific was most likely to occur in the region of maximum SST gradient on the southern boundary of the SST maximum. On inspection of the SST data, this location often coincides with the location of the ITCZ in the region from the Central American coastline to 120°W. The relationship is the reverse in the central Pacific (approximately 140°W to edge of domain bounding box at
180°W), where we typically find the ITCZ on the maximum SST gradient north of the SST maximum. In between these regions there is a transition zone from 120° W to 140° W where the ITCZ coincides with the region of highest SSTs. It should be noted that this transition zone overlaps with the region, 130° W to 160° W, where the clouds are shallower and ITCZ is least detected if total ITCZ occurrence for each longitude is calculated.

Figure 7 shows the median latitudinal distribution of the ITCZ for May to October from All-data labels. Again, the median was used to reduce the influence from the extreme 1997 El Nino year. The eastern and western longitudinal regions are shown to highlight the difference between the eastern and central Pacific. A tilt in the ITCZ from south-west to north-east is evident from figures. This tilt has been found in other observational studies (e.g. WM06; Magnusdottir and Wang 2008) and in idealized studies (Wang and Magnusdottir 2005; Ferreira and Schubert 1997), where it arises from advection by the forced vorticity field. It is also shown that the east Pacific ITCZ covers a larger region at the beginning of the season, but the envelopes of standard deviation around both curves demonstrates the considerable variability in the area of ITCZ in May. There is a northwards shift from May to September of approximately 3 degrees latitude for both east and west domains, but the southwards tilt to the west remains throughout the season. The redistribution towards the end of the season is also clear in this plot.

b. Inter annual variability

The All-data labels were used to examine inter annual variability in location, shown in Fig. 8. The plots are anomalies from the mean location of ITCZ (Fig. 4). The figure
shows considerable variability in the location of the ITCZ from year to year and especially
demonstrates the strong influence of El Nino. Particularly noticeable are the El Nino events
in 1997 and 2002 when the ITCZ region increased both in size and frequency of occurrence.
The El Nino in 1997 was particularly notable for the large average area of ITCZ – this was
also a strong El Nino event (McPhaden 1999). The La Nina events were in 1995 (weak),
1998, 2000 and 2008. In general, the ITCZ during La Nina was smaller than average in area,
although that was not the case in 2000.

Two types of El Nino and La Nina events have been noted (Kao and Yu 2009): those
which originate in the east Pacific and those which originate in the central Pacific. The 1997
El Nino was an event that originated in the East Pacific. The 2002 El Nino was a central
Pacific event. Conversely, the La Nina event of 1998-2000 was identified as a central Pacific
event, in comparison to the La Nina events in 1995 and 2000. This is somewhat reflected in
the location anomalies in Fig. 8 which suggest a reduction in the ITCZ in the central Pacific
region.

The IR-only labels were used to investigate climatic trends in the ITCZ from 1980 to 2008.
The IR satellite record had some data gaps in the early 1980s and although the statistical
model is able to handle gaps in data, diagnosis is less reliable if data is not available for
several sequential time steps. Therefore IR images where greater than 20% of the image
was missing were removed unless the images on each side in time were more complete. This
led to less information going into the data series in 1980-1991 where on average 2 days per
month were missing or incomplete in the satellite record.

Figure 8 was reproduced using IR-only labels to examine location trends from 1980-2008.
As mentioned at the beginning of the Section, the area of the ITCZ was reduced in the 120-
150W region due to lower cloud not being identified in the IR. Again, there were positive anomalies in location in El Nino years and mostly negative in La Nina years. However, the figure is qualitatively similar to Fig. 8. Linear regression was performed to remove the influence of ENSO (in a similar manner to Vimont et al. 2001) on the mean annual ITCZ location so that climatic trends could be studied. The years where the multivariate ENSO index (Wolter and Timlin 1998) was greater than 0.5 were selected and the anomalies from mean location combined and weighted by ENSO index to produce a composite picture of typical El Nino year ITCZ location anomaly. This typical composite could then be removed from the El Nino years (weighted by the strength of ENSO). The same was performed for La Nina years where ENSO index was less than -0.5. By this means, the influence of ENSO was removed or reduced, and any climatic shifts in the location of the ITCZ not due to ENSO should be revealed. Several analyses were done on the resulting data and no significant or consistent trends in location were found. However, there is much variability from year to year even after removal of the ENSO signature.

In addition to this analysis, histograms of latitudinal distribution each year were investigated and again, no climatic shifts in the position of the ITCZ could be detected. This suggests that any trends in ITCZ location over the 29 years (that are not due to ENSO), are quite small or non-existent.

Figure 9a shows the mean area of ITCZ from the IR-only labels for May to October each year, alongside the ENSO index. It should be noted that the time axis (x-axis) is not continuous. Only the months of May through October are depicted each year. The plot shows high correlation between the ENSO index and the area covered by ITCZ. The correlation coefficient is 0.67, confirming the influence of ENSO in determining ITCZ area.
To determine the long term trends in the area of ITCZ independent of ENSO, the data were normalized and the normalized ENSO index was removed. The resulting time series is shown in Fig. 9b. The plot shows that there is no overall trend in ITCZ area over the 29 years but there is much variability in size from one year to the next.

6. Concluding remarks

Our method has identified the ITCZ as a distinct weather feature in the east Pacific. The application of an MRF statistical model to the multi-year satellite data sets has revealed the seasonal evolution of the ITCZ in the east Pacific and its inter annual variability. The location and area of the ITCZ varies significantly on inter annual timescales and is highly correlated with the ENSO index. During El Nino years there were significant shifts in ITCZ location in tandem with shifts in ocean warm-pool regions and the average area of the ITCZ was greater than in ENSO-neutral years. Inspection of the 29 year ITCZ data set using IR-only labels, showed no consistent trend in the area covered by ITCZ or shifts in ITCZ location when the influence of ENSO was removed.

The ITCZ climatology of Waliser and Gautier (1993) which used cold cloud thresholding to define the convergence zone, described the ITCZ in the Pacific in a general sense. The mean latitude location was given as 8° N. This is in general agreement with our results, but moreover our results show the seasonal migration of the ITCZ and significant inter annual variability. The latitudinal location is controlled by the SST. Our comparisons between ITCZ location and SSTs showed that in the far eastern Pacific the ITCZ is located to the south of the maximum SSTs, whereas in the central Pacific (150-180°W), the ITCZ is located to
the north of the maximum SSTs.

Our results show that there is a westward shift in ITCZ location through the summer half of the year. The ITCZ is more concentrated in the east Pacific in May to July, and more distributed to the central Pacific in September and October. The ITCZ area in the east Pacific was noted to peak in late June and late August. The secondary maximum could be due to an increase in the number of tropical cyclones and other WPDs in August.

The application of statistical techniques in identifying the ITCZ has been shown to be successful. The labels automatically generated from the statistical model present a time-saving alternative to manual labelling. The model allows vast quantities of data to be analysed for the presence of meteorological features at relatively little computational cost. The model uses fast, portable code which can be run on any desktop: to identify the ITCZ for 6 months of 3-hourly data takes approximately 20 minutes on a parallelized computer using 8 processors.

The use of the statistical model is ideal for this particular weather feature due to the ITCZ’s persistent nature (rendering the use of the time component in the Markov random field highly appropriate), and the need to discard unrelated cloud features in the satellite images. Validation against manual labellers found that the statistical model was better equipped to identify coherent structures than thresholding techniques. The primary advantage of the MRF approach over other techniques is its accuracy in determining the ITCZ envelope. The model is able to take expert opinion, in the form of manual labels, and systematize the evaluation of the phenomenon over a long time period.

With the success of the method there is the possibility of applying the technique to the detection of the ITCZ in other regions of the world. The model is thus far trained on
manual labels of the ITCZ in the Pacific region but has the potential to be retrained for other global locations if manual labelling is carried out. Post processing can be tailored to the needs of any user. Another future research possibility is to apply this MRF approach to tracking non-ITCZ features such as weather systems in the extra-tropical storm tracks. It also has the potential to be used in other fields such as oceanography to track ocean eddies or phytoplankton blooms.

Future work on east Pacific ITCZ variability will take advantage of the high temporal sampling of the new ITCZ labels. We will focus on analysis of synoptic and diurnal variations of the ITCZ, the role of eddies within the tropics and interaction with extra-tropical Rossby wave breaking.

Acknowledgments.

The authors would like to acknowledge Ken Knapp at NOAA’s National Climatic Data Center for providing all IR and VS data from the HURSAT-Basin archive. We would also like to thank Deborah Smith at Remote Sensing Systems for providing the TPW data from combined radar satellite instruments. This research was supported by NSF Grant ATM0530926.
REFERENCES


Identification of tropical cyclones

Tropical cyclones produce large cloud signatures that can be embedded or isolated from the main cloudy region of the ITCZ. It is therefore of interest to know what parts of the ITCZ labels may be associated with tropical cyclones.

The horizontal extent of tropical cyclones is most often distinguished by regions of sustained wind speed over 17 m s\(^{-1}\), e.g. as specified by the National Hurricane Center\(^4\). However, higher wind speeds do not necessarily equate to the full extent of the cloud cover associated with the cyclone, and wind speeds are often only above tropical storm speed within a few degrees of the eye centre. Therefore a second algorithm using the same IR satellite data as our statistical model was applied to identify the clouds that belonged to tropical cyclones. Best track data from the HURDAT data set\(^1\) was used to establish the location of the eye of the storm. The data was linearly interpolated from the 6 hourly data set to every 3 hours. High clouds (< 255 K) overlapping a 9 pixel neighborhood near the eye were selected, along with any intersecting low cloud (< 270 K), but not non-intersecting high cloud. In this way if a cyclone was embedded in the ITCZ, the cloud belonging to the cyclone (only) was distinguished. A maximum radius was defined as 400 km, but this was rarely reached by the algorithm.

This method was seen as a balance between using more complex methods of pattern

\(^4\)www.nhc.noaa.gov
recognition or more simple methods such as defining a generic circle around all cyclone eyes as tropical cyclone cloud. The algorithm was able to adapt when cyclone-associated cloud was unusually positioned, or the cyclone was embedded within a larger cloudy region. The inferred cyclone data is used in Section 5a, and may be used for future investigations on interactions between cyclones and the ITCZ.
Table showing the comparisons between a union of the manual labels defined as ITCZ and each method of detection for August 2000. False negative is the number of ITCZ pixels not detected by each method divided by the number of ITCZ pixels. False positive is the number of pixels incorrectly detected as ITCZ by each method divided by the number of ITCZ pixels in the union. All results given as percentages. See text for more details.
Table 1. Table showing the comparisons between a union of the manual labels defined as ITCZ and each method of detection for August 2000. False negative is the number of ITCZ pixels not detected by each method divided by the number of ITCZ pixels. False positive is the number of pixels incorrectly detected as ITCZ by each method divided by the number of ITCZ pixels in the union. All results given as percentages. See text for more details.

<table>
<thead>
<tr>
<th>Method tested</th>
<th>Base method tested against</th>
<th>False negative</th>
<th>False positive</th>
<th>Total absolute error</th>
<th>intersection/union</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person P1</td>
<td>Union P2, P3</td>
<td>5.9</td>
<td>2.3</td>
<td>8.2</td>
<td>57.9</td>
</tr>
<tr>
<td>Person P2</td>
<td>Union P1, P3</td>
<td>7.2</td>
<td>2.8</td>
<td>10.1</td>
<td>51.4</td>
</tr>
<tr>
<td>Person P3</td>
<td>Union P1, P2</td>
<td>7.7</td>
<td>2.2</td>
<td>9.9</td>
<td>49.3</td>
</tr>
<tr>
<td>All-data labels</td>
<td>Union P1, P2, P3</td>
<td>7.5</td>
<td>4.4</td>
<td>11.9</td>
<td>49.2</td>
</tr>
<tr>
<td>IR-only labels</td>
<td></td>
<td>11.3</td>
<td>2.6</td>
<td>13.9</td>
<td>37.9</td>
</tr>
<tr>
<td>Threshold IR</td>
<td></td>
<td>12.3</td>
<td>3.7</td>
<td>16.0</td>
<td>32.6</td>
</tr>
<tr>
<td>Threshold TPW</td>
<td></td>
<td>6.7</td>
<td>9.8</td>
<td>16.4</td>
<td>43.4</td>
</tr>
<tr>
<td>If all domain 'ITCZ'</td>
<td></td>
<td>0.0</td>
<td>78.8</td>
<td>78.8</td>
<td>21.2</td>
</tr>
<tr>
<td>If all domain 'non-ITCZ'</td>
<td></td>
<td>21.2</td>
<td>0.0</td>
<td>21.2</td>
<td>0.0</td>
</tr>
<tr>
<td>All-data labels</td>
<td>Threshold IR</td>
<td>1.9</td>
<td>7.3</td>
<td>9.2</td>
<td>53.5</td>
</tr>
<tr>
<td>IR-only labels</td>
<td></td>
<td>2.2</td>
<td>2.0</td>
<td>4.2</td>
<td>67.5</td>
</tr>
<tr>
<td>Threshold TPW</td>
<td></td>
<td>2.7</td>
<td>14.3</td>
<td>17.0</td>
<td>35.4</td>
</tr>
<tr>
<td>If all domain ITCZ</td>
<td></td>
<td>0.0</td>
<td>87.3</td>
<td>87.3</td>
<td>12.7</td>
</tr>
<tr>
<td>If all domain non-ITCZ</td>
<td></td>
<td>12.7</td>
<td>0.0</td>
<td>12.7</td>
<td>0.0</td>
</tr>
<tr>
<td>All-data labels</td>
<td>Threshold TPW</td>
<td>8.2</td>
<td>1.9</td>
<td>10.0</td>
<td>57.3</td>
</tr>
<tr>
<td>IR-only labels</td>
<td></td>
<td>13.9</td>
<td>2.0</td>
<td>15.9</td>
<td>36.6</td>
</tr>
<tr>
<td>Threshold IR</td>
<td></td>
<td>14.3</td>
<td>2.7</td>
<td>17.0</td>
<td>35.4</td>
</tr>
<tr>
<td>If all domain ITCZ</td>
<td></td>
<td>0.0</td>
<td>75.6</td>
<td>75.6</td>
<td>24.4</td>
</tr>
<tr>
<td>If all domain non-ITCZ</td>
<td></td>
<td>24.4</td>
<td>0</td>
<td>24.4</td>
<td>0</td>
</tr>
</tbody>
</table>
List of Figures

1 (a) Map of the study region (black rectangle) split into 3 longitudinal boxes in gray (used in Section 5). (b) Spatial prior for the model for All-data labels (solid) and IR-only labels (dashed) as a function of latitude. Each spatial prior is zonally uniform across the domain. 37

2 An example of an ITCZ label automatically identified by the statistical model using IR, VS and TPW satellite data for 19 August 2000 at 2100 UTC. Shading represents the IR image with the temperature scale provided to the right. The bold contour demarks the identified ITCZ region. 38

3 Schematic showing the areas that are used to assess the agreement between ITCZ labels found using different methods. The outline of the ITCZ is shown by the solid curve for the base method and by the dashed curve for the test method. 39

4 Mean number of days when ITCZ is present (May-October) from 1995-2008 with (a) All-data labels, (b) IR-only labels. 40

5 Time evolution of ITCZ All-data labels (1995-2008) through the season. (a) Mean area of ITCZ for three longitudinal boxes: 90-120°W (bold), 120-150°W (dashed), 150-180°W (thin). (b) Mean area of cloud associated with tropical cyclones. 41
Monthly anomalies of ITCZ location (contoured) overlaid on monthly mean SST for the period 1995-2008. The contours represent median number of days in the month when ITCZ is present in comparison to the mean (Fig. 4a). Black curves are positive anomalies, gray curves are negative anomalies. The contour interval is 2 days. SST is in Celsius as shown in the color bar.

Median latitude of ITCZ distribution in two longitudinal box regions: 90-120°W (bold curve) and 150-180°W (dashed curve). Standard deviation is shown in the shading.

Location anomalies from 1995 to 2008 using All-data labels. Black lines are positive anomalies, gray lines are negative anomalies. Each line represents a 10 day anomaly in mean occurrence over the season May-October.

(a) Mean area of ITCZ labels in black (monthly means May to October) shown alongside ENSO index for the same months. (b) Normalized time series of area of ITCZ with the signal from ENSO removed.
Fig. 1. (a) Map of the study region (black rectangle) split into 3 longitudinal boxes in gray (used in Section 5). (b) Spatial prior for the model for All-data labels (solid) and IR-only labels (dashed) as a function of latitude. Each spatial prior is zonally uniform across the domain.
Fig. 2. An example of an ITCZ label automatically identified by the statistical model using IR, VS and TPW satellite data for 19 August 2000 at 2100 UTC. Shading represents the IR image with the temperature scale provided to the right. The bold contour marks the identified ITCZ region.
Fig. 3. Schematic showing the areas that are used to assess the agreement between ITCZ labels found using different methods. The outline of the ITCZ is shown by the solid curve for the base method and by the dashed curve for the test method.
Fig. 4. Mean number of days when ITCZ is present (May-October) from 1995-2008 with (a) All-data labels, (b) IR-only labels.
Fig. 5. Time evolution of ITCZ All-data labels (1995-2008) through the season. (a) Mean area of ITCZ for three longitudinal boxes: 90-120°W (bold), 120-150°W (dashed), 150-180°W (thin). (b) Mean area of cloud associated with tropical cyclones.
Fig. 6. Monthly anomalies of ITCZ location (contoured) overlaid on monthly mean SST for the period 1995-2008. The contours represent median number of days in the month when ITCZ is present in comparison to the mean (Fig. 4a). Black curves are positive anomalies, gray curves are negative anomalies. The contour interval is 2 days. SST is in Celsius as shown in the color bar.
Fig. 7. Median latitude of ITCZ distribution in two longitudinal box regions: 90-120°W (bold curve) and 150-180°W (dashed curve). Standard deviation is shown in the shading.
Fig. 8. Location anomalies from 1995 to 2008 using All-data labels. Black lines are positive anomalies, gray lines are negative anomalies. Each line represents a 10 day anomaly in mean occurrence over the season May-October.
Fig. 9. (a) Mean area of ITCZ labels in black (monthly means May to October) shown alongside ENSO index for the same months. (b) Normalized time series of area of ITCZ with the signal from ENSO removed.