their peripheral blood for 60 days after challenge were considered to be protected against malaria.

All persons who received 2 or more doses of the vaccine developed antibodies against CSP. Antibody levels were found to peak after the second dose, then declined between second and third dose and latter attained maximum levels after the third dose. There was considerable variability in antibody response among individuals and also to different formulations of the vaccine. Responses to vaccines 2 and 3 were significantly greater than those to vaccine 1. Among those who completed the study, only one of seven subjects given vaccine 3 became infected, while 7 of 8 given vaccine 1 and 5 of 7 given vaccine 2 developed malaria. Parasitemia was found in all the six subjects, who were not immunized and served as controls, 11–13 days after sporozoite challenge. The vaccine efficacy was estimated to be 86% and relative risk of infection was computed as 0.15. Protected individuals had higher antibody titres.

In their earlier studies the investigators had demonstrated that vaccine 3 is superior for inducing strong antibody responses and antigen specific delayed hypersensitivity in primates. In mice, vaccine 3 was found to stimulate proliferative and cytolytic T cell responses as well.

A remarkable feature of the RTS, S vaccine is the conspicuous effect of certain adjuvants on the vaccine efficacy. This observation suggests that immune potentiators capable of upregulating protective cytokine production or leading to advantageous immune cell responses may be combined with anti-parasitic elements in developing more potent vaccines[11].

The crucial test for the RTS, S vaccine is yet to be done. A randomized double blind placebo-controlled field trial in areas of malarial endemicity would determine whether the hybrid vaccine has an edge over the other vaccines on trial.

1. Stoute, J. A., Slaoui, M., Heppner, G. *et al., New Engl. J. Med.*, 1997, **336**, 86–91.
2. Ramalingam, S. and Mathai, E., *Natl. Med. J. India*, 1996, **9**, 120–124.
3. Nussenzweig, V. and Nussenzweig, R. S., *Adv. Immunol.*, 1989, **45**, 283–334.
4. Deirdre, H., Davis, J., Nardin, E. *et al., Am. J. Trop. Med. Hyg.*, 1991, **45**, 539–547.
5. Brown, A. E., Singharaj, P., Webster, H. K. *et al., Vaccine*, 1994, **12**, 102–108.
6. Patarroyo, M. E., *Science*, 1990, **248**, 242.
7. Alonso, P. L., Smith, T., Armstrong Schelenberg, J. R. M. *et al., Lancet*, 1994, **334**, 1175–1181.
8. D'Alessandro, U., Leech, A., Drakeley, C. J. *et al., Lancet*, 1995, **346**, 462–467.
9. Nosten, F., Luxemberger, C., Kyle, D. E. *et al., Lancet*, 1996, **348**, 701–707.
10. Gordon, D. M., Mc Govern, T. W., Krzych, U. *et al., J. Infect. Dis.*, 1995, **171**, 1576–1585.
11. Di Rosa, F. and Matzinger, P., *J. Exp. Med.*, 1996, **183**, 2153–2163.

C. C. Kartha is in the Division of Cellular and Molecular Cardiology, Sree Chitra Tirunal Institute for Medical Sciences and Technology, Thiruvananthapuram 695 011, India.

OPINION

Experimental statistics for biology students: An experiment in motivation

V. Sitaramam

A major fallacy exists in teaching that the mere existence of a subject necessarily motivates the students to go after it, merely based on teachers' advice. Turning a blind eye to what is not immediately necessary is called short-sightedness in students and dedication to specialization and pursuit of depth among the faculty. Thus, biologists end up learning mathematics and statistics for reasons of 'subject' which necessarily leads to third rate statistics and even worse mathematics (note 1). These problems worried us at the Biotechnology programme at Pune University since we take students with both biology and quantitative backgrounds. A sustained dialogue over years with freshly admitted students in basic sciences such as zoology and biochemistry and the better-funded courses such as the National Biotechnology Masters programme wherein biochemistry is an important component, as well as with graduate students enrolled for Ph D in these sciences at the University of Pune, revealed that, with the exception of agricultural students, any background in statistics is generally non-existent. Background in mathematics is also very poor among students of biology and even in general chemistry. The practical, laboratory training that they have had is also non-existent. Mensuration, taught to engineering students routinely, is non-existent in all sciences at the bachelor level (note 2). Even the idea of a derivative requires to be painfully taught to students who are already past 20 years of age. The major problem is that they are not motivated since they see no particular point in learning it. Increasing emphasis on cookbook (qualitative) procedures, kits and readymade reagents have rendered quantitation and kinetics of any kind in subsequent years increasingly extinct (note 3).

The biotechnology programme at Pune University has been documented at length with regard to its components (see refs 1–3) and represents one of the few University programmes in India in which the progress (note 4) has been kept track of by continuously monitoring various aspects of student performance (note 5). The existential concerns related to the need to emphasize the quantitative aspects of biology (note 6) have motivated us in the teaching programme to introduce interactive experiments that convince students that there is more to science than

what meets the eye. The task is three-fold; firstly, the experiments should be posed as statements simple and commonplace enough to be understood, and yet complex enough to be nearly unanswerable; secondly, students should understand that facts do not speak for themselves and are to be made to speak; and thirdly, computers and computations, statistical and mathematical, actually make the tasks easier and not obtuse and difficult. The last is of particular importance since the opinion of students on the available books varies from 'boring' to 'forbidding rather than friendly'.

A sustained effort was made to shake up students from a rather indifferent if not negative attitude towards all things quantitative (ref. 1). We need deceptively simple experiments that actually make the students *wary* of their own observations and conclusions (note 7). This is aimed at achieving specific dividends: firstly, the message could be made clear that, measurement is a matter of serious concern, leading to increased interest in practical training; learning statistics and mathematics could be of advantage, improving the attendance and attentiveness in these subjects; and lastly, students grasp the need to ask and answer aggressively what a method or procedure means and that this question would be meaningless without verification or testing. I outline here a couple of introductory experiments on measurement, introduced at the beginning of a course in physical biochemistry and conducted over the last eight consecutive years (note 8). The experience may be summarized as a worthwhile effort (note 9), primarily based on student feedback.

Problem 1

On day one, the first year students are assembled with the following question:

'You students are admitted *via* an entrance test at the national level. By some strange machinations in New Delhi, we got wind of fact that, this year, the selection was highly arbitrary (note 10). The selection was exclusively based on body size: only the shortest of the girls and the tallest of the boys have been chosen. Prove or disprove.' They are asked to freely discuss among themselves during lunch and come back with a procedure to be implemented the same afternoon. They are forbidden to discuss this problem with the seniors in the second year, for which the seniors also fully co-operated. There are about 18–20 students per year.

What I describe here is a set of common responses and how the experiment proceeds. No two batches respond the same way, making the experiment always interesting to the teacher. It is also amazing how often the students attempt to describe the solution free of all jargon (which they are yet to learn) and how the basic idea is understood by nearly all intuitively. Some working knowledge in statistics for the instructor helps. It is important not to help the students.

The solutions

The students take the problem initially as a joke and at least 20% claim that they can settle it in an hour. They state that they wish to measure and ask for an expandable scale of the kind one finds in a clinic. Since the department has none, they settle for a tape and use the wall to mark the heights and measure these heights with a tape. Less than 10% of the students explicitly realize that one needs some device that projects perpendicularly from the wall to mark the height while avoiding error due to parallax. They are all familiar with the idea of the least count. They are emphatic that the error for measurements with a metric tape would not exceed the least count of 1.0 mm. They accept that some of them are good in measuring and some may not be and that it is worthwhile to insist that everybody measures everybody else and that we can also figure out how good the capability is. They do not speculate how to go from these measurements to the actual question. They reduce the problem to what they can handle first: Are the girls taller or shorter than the boys? They know that they should compare the mean heights. In another two hours or so, they prepare 19×19 matrix (note 11). A table is made and they obtain the mean values as well as ranges of heights. They can see that given a person x, his/her height taken by everybody has less variation, i.e. lower range than the whole class. Surprisingly this range is larger than 1.0 mm and usually goes up to 3–4 mm, i.e. 169.3–169.7 cm for the height of x; $n = 19$. The students are asked to make up their mind: was that a good measurement? The answer at that point has always been negative. The measurement would *never* match their initial notions of perfection. From this point onwards, a number of conflicting hypotheses are thrown at the group by the teacher. The first suggestion is that the tape is faulty either in the markings being non-uniform or that it is made of rubber and tends to be elastic and varies depending on the pull.

The students deny the first and swear that the second cannot be true, since they have taken a metal tape and that the room did not get appreciably hotter during the measurement. They begin to enumerate where each could have gone wrong. Some of them could be poor in handling the measurements. How can they find out? The problem is that the lowest error as detected by the range is still 2–3 times the least count. It is so regardless of whether the persons, the measurer or the measured, are tall or short. The range dramatically comes very close to the least count when the scoring of the height marked by one student was visible for the next person while marking on the wall. The students agree that it is a bad procedure since one is not really measuring and is only trying to conform to the previous measurement. They intuitively agree that the 'blind' measurements are objective and superior to such biased measurements.

At the second stage, the students compare the heights of all individuals in the class and then as sub-groups of boys and girls. Now the range is much larger from the tallest to the shortest. They did not have difficulty in understanding why the range should be very large, as much as 32 cm. The notion of biological variation among individuals is something they understand very readily. How this can be proven to exist is another matter since measurements themselves have variation! When they are asked the question whether this difference was due to error in measurement, the response was immediate, No. The reasoning, however, was slow to come. They are introduced to the terminology of intra-individual variation and inter-individual variation. It was suggested to them that it would perhaps be all right if the inter-individual variation was much larger than intra-individual variation to believe a measurement (note 12). Thus, if boys are taller than girls, this statement is reliable if and only if we provide a view that the inter-individual

variation is much larger than intra-individual variation within multiple measurements, least count representing the lowest error that could be possibly recorded.

At each stage they are egged on to answer the starting question. They feel strongly that boys being taller than girls could be universal which may have nothing to do with the selection process. In fact, all of them feel that such a statement is not possible unless we have the heights of all the students who appeared for the examination including those who have passed.

Then the students are introduced to the calculation of standard deviation and standard error of mean. There is no difficulty in understanding the meaning of mean; standard deviation takes a little more time. It is suggested to them that the range would be roughly four times the standard deviation. This leads to a discussion of frequency distribution, and the meaning of 95% confidence limit could be appreciated. They definitely like the idea that an error follows some rules in real measurements. The difference between a systematic error and random error becomes self-evident and there is no real difficulty in appreciating that even random errors follow some rules. Standard error of mean takes even a longer time to comprehend. The fact that each individual measurement as well as the mean itself could vary is understood; then standard error is also understood. It is also easier to understand that the standard deviation would be larger than the standard error of mean. (They are amused by the suggestion that one prefers to illustrate in publications data with standard error than with standard deviation!) The equations corroborated the view that mean would tend to vary less than the individual measurements. Measurements as duplicates, triplicates and so on are appreciated as a good strategy to enhance the reliability of measurement.

By this time (usually 2–3 hours each of two sessions at least were spent by then), the students agree that there is nothing called a 'perfect' measurement and also that how good a measurement is can be quantitated. Calculations by hand or with pocket calculators were painful for those who did not know how to use a spread sheet on a computer. They all agree that computers are definitely worthwhile if the data analysis is to be kept to a fraction of time for data acquisition. By then, the polarity in the class becomes very strong. The vocal invariably take the position that the question is not answerable with the data available. There are always one or two doubting Thomases. One went to the departmental office one year and obtained the marks/ranks obtained during selection in Delhi and showed that there was no difference between the means and standard deviations of performance for boys and girls, nationally and those for Pune alone, though their heights at Pune differed. In a span of eight years since these practicals were started, not even once did the students give a suggestion of a procedure. Once a student did argue that there were more boys above the mean value for themselves while the girls did not reflect this. There was massive confusion how more number of boys could be above their own mean.

Then the idea of a frequency distribution is introduced and discussed by the instructor with emphasis on the nature and meaning of skewness. Symmetric and asymmetric distributions are identified. The students immediately plot the frequency distributions from the data and look for skewness. The problem is translated into Figure 1 and then the students agree that there could be such a possibility. They also are of the firm opinion that this does not prove the point but merely affirms the possibility. At this point they are introduced to the idea of a *t*-test. They use this to check that boys and girls do differ in their mean heights; then they compute skewness and kurtosis. The students begin to sense that effective comparisons are meaningful only if the curves of frequency distributions look nearly the same. Various Pearson distributions, when formally illustrated, amplify this feeling. They all agree that it is desirable to have same number of observations in the groups for an effective comparison, considering the difficulty in getting good frequency distributions at smaller numbers. They simultaneously learnt the art of dividing observations into class intervals and about uni- and multi-modal distributions. For this, they combine the data of boys and girls and see whether the frequency distributions tell them if they are subgroups. They understand that existence of a single apparent uni-modal distribution need not rule out subpopulations, in as much as poor selection of class intervals could make every group a subpopulation! They see that there is a role for a good interpreter and that the use of statistics does not preclude the use of human mind and that it actually augments it. That is considerable progress since the best they could come up with before the experiment was something like lies, damn lies and statistics.

Usually there are some adventurous souls who try to make something like paired data out of boys and girls and fail miserably. Failure of an idea is more welcome than having no idea at all. We generally found it necessary to see to it that the students exhaust all possibilities that they could imagine. Testing has no

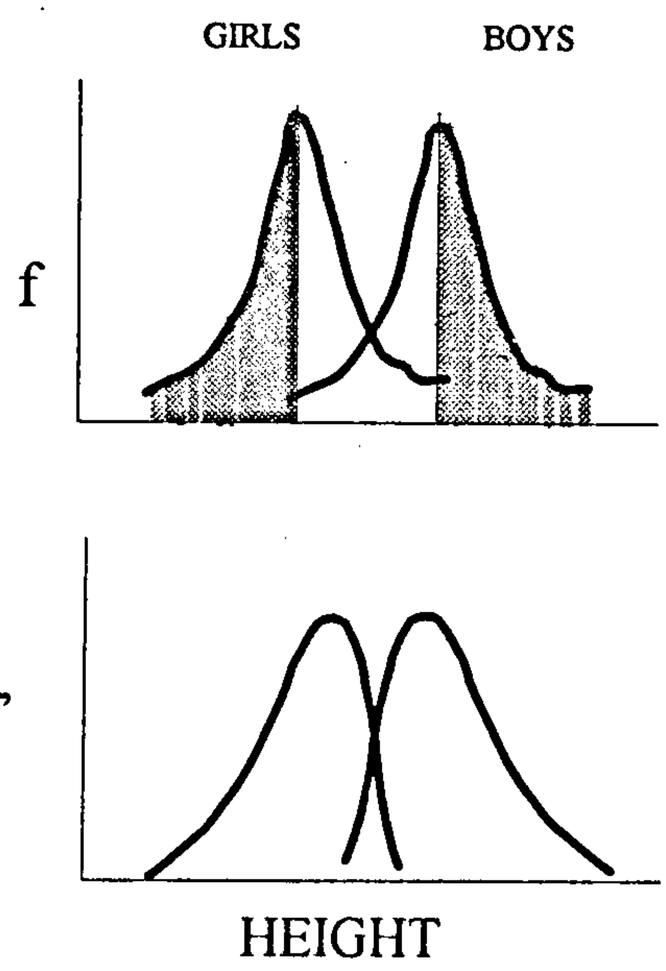

Figure 1. Problem 1 is depicted pictorially. Top panel shows that the frequency distribution of boys and girls is different. The mean/median/mode of this symmetric distribution is indicated by a vertical line for each distribution. The hypothesis implies that only the shaded lot are selected: the taller among the boys and the shorter among the girls. The bottom panel spells out the kind of distribution created by the suggested selection bias: note the asymmetry (skewness) of the distributions in the bottom panel and how it differs for the boys and the girls.

The discerning reader (to date only a referee and not even one student) would note that *f'* does not follow from *f*, given the rule as stated in the top figure by deleting the non-shaded area. The rule needs to be modified. The answer is simple but will be left unanswered here. Incidentally an important component of this kind of instruction is that answers need to be found by modelling.

meaning if alternatives are not imagined. It is necessary to play on every spontaneous idea and create a measurement around it so that the ideational process gets aroused and encouraged. By now the students heartily agree that measurement is the right thing. Subjective ideas are no good. It is possible to veer the discussions towards realistic situations wherein the importance shifts to paired data.

There were years when the data suggested that the starting premise was right and there were years when the starting hypothesis was rejected outright. The former was most instructive in understanding what a chance association is. The students actually take 20–30 sets, each of 8–10 random numbers, and do t test to see if the significance ($P < 0.05$) ever emerges due to chance alone. Since it always does, such results have a chastening effect on their thinking.

Then the second problem is posed to them.

The second problem

It is not possible to accurately transfer sugar into a beaker just with a spoon and without a balance. The initial reaction is affirmative. The test procedure is also conjured up readily. Take one spoonful, two and three and so on and see how reproducible the weighment per spoon is. When they do the experiment, they are asked to test how good it is. A few promptly convert it into 4–5 measurements of weight per spoon, though they initially plot a linear relationship between number of spoons of sugar and the weight. They get the mean and standard deviation. However, they are at a loss to assess just how good the spooning is!

They are introduced to linear regression by the method of least squares (note 13). They assess the slope and intercept as well as the coefficient of correlation for the X on Y as well as Y on X regressions. Linear regression by the method of least squares has rather rigorous assumptions which are worth knowing. Foremost of them is the point that one has to choose the variables correctly for the axes. The method assumes that the error is in assessing Y such as the height of the individual (or the absorbance of a reaction) and the age (or the concentration of the substance like glucose) which is generally, and these days, reasonably known. They are startled to see that the slope and intercept are not the same while the coefficient of correlation remains the same when they do linear regression after transposing the axes one for the other. After the exercise, they understand readily the three cardinal assumptions of linear regression, of the error being on Y, normally distributed, and constant at all values of X. This last assumption, viz. homoscedasticity, becomes not too difficult a concept after the experiment. They guess that in real measurements, error would be high at low values measured since they approach the least count. They are already familiar that actual error due to intra-individual variation would be larger than the least count! They agree that homoscedasticity is a pipe dream in difficult measurements, i.e. at the limits of sensitivity.

More striking is the realization an year later that biochemists generally teach the plotting of standard curves the wrong way. There were at least half a dozen students who argued in these eight or more years that, to assess the unknown protein or glucose, they should regress the reverse way, i.e. X on Y and not Y (the colour) on X (the glucose or protein concentration) for the standard curve. There have been some embarrassing moments due to conflicts with teachers who would usually insist on Y on X regression. A more common response from teachers (even from some of the referees of prestigious journals, in my experience) is that it does not matter if the assays are of 'quality'.

The next level becomes really interesting when there is an opportunity to discuss with the students that what is important is the variance and not the mean. An explanation of $(1 - r^2)$ being the variance unaccounted for by glucose and this should be the lowest for a good measurement brings home the all important message of quality control. There is a scramble for who gets the lowest $(1 - r^2)$ in the spoon experiment. At least one or two repeat the experiment just to prove the instructor wrong that nearly perfect fits can be obtained by measuring by eye alone. Bless them.

Concluding remarks

The distractions are the most important part of this instructional process. The students do not mind being proved wrong as long as this is for a good reason. Every group has its own leaders. It is fascinating how these self-styled leaders get cut to size by their own inflexible positions in such group interactions. This has the salutary effect of promoting intellectual anarchy or free thinking. These are not just *post facto* analyses by the author. Students take part in the subsequent months in such discussions and laugh about it. Among the major benefits from this experiment are the ease with which chromatography (wherein the width of the eluting band has interesting possibilities on the source of error) could be taught later, as also a sensitization to non-normal errors as in radio-immunoassays. Kinetics could be taught more ambitiously and rather critically (note 14). These exercises are not without drawbacks. Major among them is what the instructor has to face. After one semester of such a set of experiments, he never gets an immediate reply for any question from any student than 'it depends'! They become professorial, leaving the instructor to make all the mistakes.

References

1. Sitaramam, V., *Curr. Sci.*, 1991, **60**, 537–540.
2. Sitaramam, V., *Curr. Sci.*, 1995, **68**, 779–782.
3. Sitaramam, V., *Curr. Sci.*, 1996, **70**, 335–340.

Notes

1. I refer to the majority rather than the gifted minority since this minority does not need these views.
2. We must not consider medical students among these since the Indian medical student learns any and everything that he can only by rote. Most of the modern biology that these Western medical textbooks increasingly refer to remains as strange as voodoo in most medical colleges where science must be taught by those who have medical and not science degrees.
3. It would be disconcerting if the future generations of biochemistry students find it difficult to derive even the km.
4. As well as the lack of it.
5. As also the departmental, by national committees over the years.
6. A concern shared only with the IITs and JNU, Delhi, while we at Pune remain the only University with unrestricted admittance from either stream.
7. In fact, this is the major bane of teaching in biology wherein hypotheses are preferred

to be accepted unquestioningly rather than be understood for their internal contradictions and uncertainties. Teaching takes the same dogmatic route.

8. A subdued version of this same experiment may be carried out using high school students of 14–16 years age without a significant loss of content.
9. This is in spite of the fact that the experiment does involve a lot of effort and tends to be viewed not as biochemistry/biology experiment by fellow teachers. I am yet to come across a statistician who believes that the experimental component of biology is any of their concern in teaching.
10. A disclaimer is perhaps now in order that the problem does not refer to any actual instances and any resemblance to reality is purely a coincidence.
11. Following the well-known aphorism 'a barber is one who shaves every one except himself'.
12. It often comes as a revelation that there is no absolute measurement and even that a measurement need not be far better than what the conclusions warrant.
13. Often one needs to estimate the concentration of protein/glucose in an unknown sample. The standard way to do it is to prepare solutions with known concentration, and measure some property, which is usually the absorbance. The plot of absorbance (Y axis) vs concentration (X axis) is known as the standard curve.
14. It is interesting that the author never came across a paper that critically tested whether the point of intersection of the reciprocal plots to detect the order of the reaction was actually above or below the line ($1/v = 0$). It is the faith and not the substrate that can reach a concentration of infinity readily.
15. Throughout the text, I deliberately omitted commonplace statistical definitions and equations to emphasize the need for intuitive understanding. Nearly everything here and much more is available in any primer on statistical methodology, except the motivation to open these pages.

ACKNOWLEDGEMENTS. The Biotechnology programme at Pune is supported by the Department of Biotechnology, Government of India. Presence of a good statistics department within a stone's throw on the campus proved to be mutually beneficial over the years in both the teaching programmes and research: we obtained the procedures and they, the calculi. This paper was prepared during the sabbatical year of the author at Rio de Janeiro, supported by a grant (#300917/95-3) from Brazilian National Research Council – CNPq – towards a visiting professorship. Those desirous of more details of the procedures and extending these to routine biochemical experiments may contact the author at Pune.

V. Sitaramam is in the Department of Biotechnology, University of Pune, Pune 411 007, India.

SCIENTIFIC CORRESPONDENCE

Experimental annual forecast of all-India mean summer monsoon rainfall for 1997 using a neural network model

Accurate long-range forecast of rainfall can help in improved planning of agricultural strategies as well as in crisis (e.g. drought) preparedness. However, conventional methods for such forecasts, based on either statistical or dynamical methods, have limited skill in long-range prediction of monsoon rainfall. This is particularly true for prediction for longer than a season in advance. Although some of the statistical models have shown considerable skill in forecasting Indian summer monsoon rainfall (ISMR)[1], their range does not exceed a few months. Besides, most of these statistical (regression) models require a number of observed parameters, many of which are not available until before onset. This has been one of the major obstacles in attempts to generate very long range (VLR, one year or more) forecasts of monsoon rainfall.

In an attempt to develop an alternative successful methodology for generating VLR forecast of ISMR, the technique of neural networks was explored[2]. The usefulness of neural networks for simulating atmospheric process has been emphasized in recent years[3]. However, extensive experimentation and evaluation revealed the conventional neural networks to be inefficient to generate VLR forecasts of ISMR with skill and consistency. However, it was shown in a recent work that a generalization of the structure of a conventional neural networks, termed cognitive network (CN), has the potential for generating VLR forecast of ISMR with good accuracy[2]. Subsequently an extensive statistical evaluation of the performance of CN was carried out by generating 73 hindsets for the period 1821 to 1993 for different test cases. Results of these analyses, presented elsewhere[4], show that the CN can generate annual forecast with average (over 73 cases) error less than half the standard deviation of the data, with absolute error about 36 mm. Encouraged by these results, an experimental annual forecast of ISMR was generated for 1996 using this method as 866 mm.

These forecasts were generated using ISMR data for 50 years prior to 1995 taken from Parthasarathy *et al.*[6]. The forecast for 1996 was 866 mm, which is about 102% of the mean of the 120-year data used in our study. The corresponding observed value is also 102%, indicating an excellent agreement of our experimental forecast with observation. However, it should be mentioned that the (yearly) normal used by the India Meteorological Department is somewhat different from the simple 120-year mean considered by us[7]. The purpose of this note is to record experimental forecasts generated by us for 1997 and 1998, as 846 and 945 mm respectively.

The purpose for generating forecasts for 1997 and 1998 is that it will contribute to an objective evaluation of both the skill and the range of the forecasts.

It should be mentioned that both the forecasts for 1997 and 1998 (and previously for 1996 as well) were generated as an average of an ensemble of forecasts for the respective year. These ensembles were computed by generating a large