1 **Short title: Neutral parameter inference**

2

3 **Title: A novel inference of the fundamental biodiversity number for**

4 **multiple immigration-limited communities**

5

6 **Champak Beeravolu Reddy.[1*], Pierre Couteron [1] and François Munoz [2]**

7

8 [1] IRD, UMR AMAP, Bd de la Lironde, TA A-51/PS2, 34398 Montpellier Cedex 05, France

9 [2] Université Montpellier 2, UMR AMAP, Bd de la Lironde, TA A-51/PS2, 34398 Montpellier

10 Cedex 05, France

11

12

13 *Corresponding author*:

14 Champak Beeravolu Reddy

15 UMR CBGP (INRA/IRD/Cirad/Montpellier SupAgro),

16 Campus International de Baillarguet CS 30 016 F-34988, Montferrier-sur-Lez cedex, France

17 Tel.: +33 4 99 62 33 31; Fax: +33 4 99 62 33 45

18 Email: champak_br@yahoo.com

19

20

24

1   **Keywords:** parameter estimation, neutral theory, island biogeography, coalescence theory,

2   community ecology, similarity statistics, tropical trees, spatial scale.

3

4   **Type of paper:** Letter

5   **Word count (abstract):** 148

6   **Word count (whole manuscript):** 6284

7   **Word count (main text):** 4501

8   **No. of references:** 50

9   **No. of figures:** 4

10   **No. of tables:** 0

11

12

13

1    **Abstract**

2    Neutral community theory postulates a fundamental quantity, $\theta$, which reflects the

3    species diversity on a regional scale. While the recent genealogical formulation of community

4    dynamics has considerably enhanced quantitative neutral ecology, its inferential aspects have

5    remained computationally prohibitive. Here, we make use of a generalized version of the

6    original two-level hierarchical framework in order to define a novel estimator for $\theta$, which

7    proves to be computationally efficient and robust when tested on a wide range of simulated

8    neutral communities. Estimating $\theta$ from field data is also illustrated using two tropical forest

9    datasets consisting of spatially separated permanent field plots. Preliminary results also reveal

10   that our inferred regional diversity parameter based on community dynamics may be linked to

11   widely used ordination techniques in ecology. This paper essentially paves the way for future

12   work dealing with the parameter inference of neutral communities with respect to their spatial

13   scale and structure.

14

1 **Introduction**

2      In his now well known contribution, Hubbell (2001) introduced a neutral theory of

3 biodiversity, building upon the mainland-island model borrowed from the theory of island

4 biogeography (MacArthur & Wilson 1967). His theory basically describes a model of

5 interacting communities where the slow and regional scale dynamics such as speciation and

6 extinction occur on a metacommunity level and the relatively faster local demographic events

7 such as the birth and death of individuals are the mainstay of the local community dynamics.

8 Hubbell (2001) further imagined equilibrium situations for these two entities which are

9 dictated by the parameter $\theta$ (the fundamental biodiversity number) for the metacommunity

10 dynamics and an immigration parameter $I$ (Etienne & Olff 2004) for the local community

11 dynamics. The interaction between these entities is modelled through the arrival of immigrant

12 individuals from the regional pool of species (i.e. the metacommunity) towards a single or

13 multiple local communities. It can be further interpreted as a measure of the isolation that a

14 local community undergoes from the regional species pool due to some form of limited

15 dispersal or more exactly immigration-limitation (Beeravolu *et al.* 2009). Up till now, much

16 of the interest on neutral models in ecology has been either based on this two-level spatially

17 implicit hierarchical framework (Hubbell 2001; Vallade & Houchmandzadeh 2003; Etienne

18 2005) or on spatially explicit models which do not make any such hierarchical distinctions

19 (Chave & Leigh 2002; Zillio *et al.* 2005; O'Dwyer & Green 2009). In addition to model

20 structure, details of the metacommunity processes such as speciation have been explored

21 (Haegeman & Etienne 2008; Kopp 2010) though a consensus neutral approach is still lacking

22 in community ecology (Gravel *et al.* 2006) and work is still in progress (Haegeman & Loreau

23 2011).

24      One explanation for the discord among ecologists, especially on the practical

25 relevance of the two-level spatially implicit neutral model (hereafter denoted 2L-SINM), has

4

1   been the issue of neutral parameter inference. Most of the current methods consist of either

2   fitting a species abundance distribution (Purves & Pacala 2005; Dornelas *et al.* 2006; McGill

3   *et al.* 2006) or finding quantitative point estimates of the neutral parameters (Etienne *et al.*

4   2006; Munoz *et al.* 2007). In the former, the general binning process of the species

5   information into abundance classes, among other concerns (McGill *et al.* 2007), has raised

6   several issues (Gray *et al.* 2006). At the same time the reliability of point estimates also

7   remains doubtful (Leigh 2007), particularly as $I$ and $\theta$ are known to be "hyperbolically

8   correlated" when estimating them simultaneously (Beeravolu *et al. Submitted manuscript*;

9   Etienne *et al.* 2006; Munoz *et al.* 2007; Beeravolu *et al.* 2009; Jabot & Chave 2009). Also, a

10  basic contention underlying all these critiques has been the inherent insufficiency of the

11  abundance information to fully elucidate ecological and evolutionary processes that neutral

12  models aim to combine (Harte 2003).

13      In this paper, we build upon a recently discussed neutral framework which resembles a

14  decoupling of scales (Levin 1992) and further enhances the spatial structure of the 2L-SINM.

15  This is accomplished by introducing an intermediate level of regional process which

16  generalizes the classical 2L-SINM (Hubbell 2001) into a 3L-SINM (Munoz *et al.* 2008) and

17  enables the relaxation of the speciation-drift (or speciation-extinction) equilibrium assumption

18  on the pool of available immigrant individuals (Beeravolu *et al.* 2009). In a previous paper,

19  Munoz *et al.* (2008) introduced the 3L-SINM and used the steady state results to improve on

20  the independent estimation of the neutral immigration parameter $I$, whereas we use similar

21  approaches to establish an independent inference of $\theta$ for the first time.

22

23  **Methodological background**

24  <u>The hierarchical SINM</u>

1    Under the original 2L-SINM model, local communities are defined as panmictic

2    patches of species assemblages sufficiently isolated from each other so as not to receive an

3    immigrant individual directly from another local community (Fig. 1). Besides, from a regional

4    perspective, local communities are presented as immigration-limited samples of the same

5    metacommunity and subject to a top-down arrival of immigrants depending upon the

6    immigration probability $m$ (which is the scaled version of $I$ on the unit scale). One possible

7    generalization of the 2L-SINM can then be defined as a regional scale common to all local

8    communities and whose species composition is allowed to differ considerably from that of the

9    metacommunity, *sensu* Hubbell (2001, pg. 122), which is supposed to be panmictic and at

10   speciation-drift equilibrium (Munoz *et al.* 2008). In more practical terms, immigrant species

11   originating from a common source pool may very well correspond to a particular sub-region

12   of the larger hypothetical metacommunity. For instance, let a mountain range be characterized

13   by a coherent biogeographical history and represent a large-scale metacommunity. If a deep

14   gorge runs through this range at some point, for all practical purposes, we can consider a part

15   of the range to be under some sort of isolation from immigrants coming from the rest of the

16   metacommunity, especially for the case of sessile biota such as plants even though their

17   overall floristic composition remains specific to the mountain biotope.

18   Subsequently, the intermediate pool of species under the 3L-SINM can be modelled as

19   a sample drawn from the metacommunity under the influence of non-neutral processes or

20   simply as an immigration-limited sample (see Jabot *et al.* 2008 for a similar approach).

21   Moreover, as for local communities, this intermediate pool can be described analytically using

22   a genealogical approach (for each individual of a local community) similar to the coalescent

23   theory of population genetics (Etienne & Olff 2004). The composition of this intermediate

24   pool is then defined as the set of ancestor species having immigrated at some point of time

25   into the local communities. This 3L-SINM also "collapses" back to the particular case of a

1    2L-SINM for the case of an immigration-unlimited random sample of the panmictic

2    metacommunity (see Fig. 1). Besides, this collapse and vice versa forms the backbone of

3    Munoz *et al.*'s (2007, Appendix) coalescence simulation strategy of multiple local community

4    samples which is briefly described later (see "Simulating multiple local community

5    samples").

6

7    <u>Conditional similarity under the 3L-SINM</u>

8         Following Munoz *et al.* (2008), we refer to the similarity of individuals belonging to a

9    same species within a local community sample ($F_{intra}$) using the Simpson concentration

10   (Simpson 1949). Let $S$ be the total number of species among $N$ local community samples. For

11   a given sample $k$ (varying from 1 to $N$), this index represents the probability of randomly

12   drawing (without replacement) two conspecific individuals (using Munoz *et al.*'s (2008) exact

13   estimator of sample similarity):

14

$$F_{intra}(k) = \sum_{s=1}^{S} \frac{n_{sk}}{n_k} \frac{n_{sk}-1}{n_k-1} \qquad (1)$$

16

17   where $n_k = \sum_{s=1}^{S} n_{sk}$ and $n_{sk}$ is the number of individuals of the $s^{th}$ species found in the $k^{th}$

18   sample. We define the time dependant version, $F_{intra}(k,t)$, of the intra sample similarity (or

19   conspecific at time $t$) and in keeping with the 3L-SINM framework, the theoretical similarity

20   at $t$ between the intermediate (regional) pool of available immigrating species and a given

21   local community $k$ is $F_{inter}(pool,k,t)$. If each local sample is assumed to reasonably

22   approximate the composition of its respective local community, it is possible to derive, as

23   shown in the following, the time independent expectation $F_{intra}(k)$ (Munoz *et al.* 2008,

24   Appendix C).

1    Let us assume that, between a time step $t$ and $t+1$, there occurs a random death in the

2    $k^{th}$ local community, such that two randomly chosen individuals may or may not contain the

3    dead individual with probability $2/n_k$ and $1 - 2/n_k$ respectively. Following a coalescent

4    approach (see Etienne & Olff 2004, eqn 1), for every dead individual, the replacing individual

5    is either the offspring of a local individual or an immigrant individual of a lineage currently

6    not present in the community. In the former, the replacement probability is given by $m_k$ and 1

7    – $m_k$ for the latter, where $m_k$ represents the immigration probability into the $k^{th}$ local

8    community. Consequently, the conspecific probability (hereafter denoted CnP) at $t+1$ of two

9    randomly chosen individuals from a community which include the dead individual is the sum

10   of three different transition probabilities.

11   When the replacement is an immigrating replacing individual, the CnP of the chosen

12   couple is $F_{inter}(pool,k,t)$ which is the time conditional version of $F_{inter}(pool,k)$. If the

13   replacement is a local event, it could be the descendant of the other individual, with

14   probability $1/n_k$, in which case the CnP would be 1. If the replacing individual belongs to an

15   offspring of an individual from the rest of the community (with probability $1 - 1/n_k$) the CnP

16   is given by $F_{intra}(k,t)$. We could also consider the possibility that the dying individual

17   produces offspring (a modification of Moran's (1958) model), thus adding to the competition

18   for the vacant spot as suggested by Hubbell (1979; 2001), though this wouldn't affect the final

19   result (Munoz *et al.* 2008). In sum, considering all the transition probabilities defined above,

20   we can write the full CnP at $t+1$ for any two individuals in a local community as (Munoz *et al.*

21   2008, Appendix C)

22

23

$$
\begin{aligned}
F_{intra}(k,t+1) = & \left(1 - \frac{2}{n_k}\right) F_{intra}(k,t) \\
& + \frac{2}{n_k}\left\{ m_k F_{inter}(pool,k,t) + (1-m_k)\left[\frac{1}{n_k} + \left(1-\frac{1}{n_k}\right)F_{intra}(k,t)\right]\right\}.
\end{aligned}
$$

(2)

1

2    Here, $m_k$ can be defined either as $I_k/(I_k + n_k)$ or $I_k/(I_k + n_k - 1)$ which corresponds to the

3    unmodified and modified Moran's models respectively and $I_k$ stands for the number of

4    immigrating individuals into $k^{th}$ local community (Etienne & Olff 2004). At steady state,

5    $F_{intra}(k,t+1) = F_{intra}(k,t) = F_{intra}(k)$, which reduces eqn (2) to

6

7    $$F_{intra}(k) = 1 - \frac{I_k}{I_k + 1}\left(1 - F_{inter}(pool,k)\right). \qquad (3)$$

8

9        Let us assume that the $k$ local community samples are far enough from each other in

10    order to represent distinct communities and denote $F_{inter}(k)$ to be the CnP of an individual

11    from the $k^{th}$ community and an individual from one of the $k - 1$ other communities.

12    Consequently, the coalescence approach dictates that $F_{inter}(k)$ is equal to the CnP of their

13    respective ancestors who are distinct immigrating individuals from the regional pool (i.e.

14    $F_{intra}(pool)$) which can be written down as (Munoz *et al.* 2008, eqn 3):

15

16    $$F_{inter}(k) = F_{inter}(pool,k) = F_{intra}(pool). \qquad (4)$$

17

18    Local similarity under the 2L-SINM

19        In this section, we develop the other key idea which pertains to the model "collapse"

20    from the 3L-SINM to the 2L-SINM (see Fig. 1.) which entails that the results of the previous

21    section also apply to the particular case of the 2L-SINM. For the sake of completeness, we

22    also detail the analytical relationship linking $F_{intra}(k)$ to the immigration parameter $I_k$ and the

23    biodiversity parameter $\theta$ for the 2L-SINM (see also Etienne 2005, eqn 8).

9

1    Let $F_n(k)$ be the CnP of drawing a sample of $n$ individuals from the $k^{th}$ local

2    community and $j$ the corresponding number of ancestors in the metacommunity for the same

3    sample. In other words, the quantity $j$ corresponds to the number of those individuals who

4    were the first of every lineage to have immigrated into the community sample, which, for a

5    sample of size $n$, is $j < n$ original community lineages. Accordingly, we write the $n$-sample

6    CnP as

7

8
$$F_n(k) = \sum_{j=1}^{n} p_{meta}(1/j) \, p_{comm}(j/n) \tag{5}$$

9

10   where $p_{comm}(j/n)$ is the probability of drawing a sample of $n$ individuals from the $k^{th}$ local

11   community which are the progeny of exactly $j$ different immigrating individuals and $p_{meta}(1/j)$,

12   the probability that all of these $j$ ancestors belong to the same species.

13       The 2L-SINM describes a metacommunity at speciation-drift equilibrium whose

14   sample abundance distribution can be described using well known multinomial formulas from

15   the field of population genetics (see Hubbell 2001, 119). Thus a random sample from the

16   metacommunity containing $j$ individuals belonging to $\sigma$ different species has the probability

17

18
$$p_{meta}(\sigma/j) = \frac{\overline{s}(\sigma, j)\theta^{\sigma}}{\theta^{(j)}} \tag{6}$$

19

20   where $\overline{s}(\sigma, j)$ is the unsigned Stirling number of the first kind and $\theta^{(j)}$ is the Pochhammer

21   notation for the rising factorial defined as $\theta^{(j)} := \prod_{i=1}^{j}(\theta+i-1)$ (Ewens 1972; Tavaré & Ewens

22   1997, eqn 41.5). This can also be extended to an immigration-limited local community (at

1   immigration-drift equilibrium) where the immigration process replaces the speciation process

2   and

3

$$p_{comm}(j/n) = \frac{\overline{s}(j,n)(I_k)^j}{(I_k)^{(n)}} . \tag{7}$$

5

6   Thus, eqn (5) can be rewritten as

7

$$F_n(k) = \sum_{j=1}^{n} p_{meta}(1/j) p_{comm}(j/n)$$
$$= \sum_{j=1}^{n} \frac{\overline{s}(1,j)\theta}{\theta^{(j)}} \frac{\overline{s}(j,n)(I_k)^j}{(I_k)^{(n)}} \tag{8}$$

9

10   which, for the special case of a conspecific sample consisting of two individuals (i.e. $n = 2$)

11   reduces to (Etienne 2005, eqn 8)

12

$$F_2(k) = F_{intra}(k) = 1 - \frac{\theta}{\theta+1} \frac{I_k}{I_k+1} . \tag{9}$$

14

15   Using eqns (3), (4) and (9), the expression for $\theta$ simply reduces to

16

$$F_{inter}(k) = \frac{1}{\theta+1} . \tag{10}$$

18

19   <u>Estimator $\theta$ based on the conditional inter-sample similarity</u>

11

1    Here, we define a statistic (previously used by Munoz *et al.* 2008) which measures the

2    $k^{th}$ sample's similarity with respect to the rest of the samples thereby providing an estimate of

3    the quantity $F_{inter}(k)$. This can be written in terms of sampling without replacement as

4

5
$$\hat{F}_{inter}(k) = \sum_{s=1}^{S} \frac{n_{sk}}{n_k} \frac{n_s - n_{sk}}{N_T - n_k} \qquad (11)$$

6

7    where $N_T = \sum_{k=1}^{N} \sum_{s=1}^{S} n_{sk}$ and $n_s = \sum_{k=1}^{N} n_{sk}$. Using eqn (10), we can now write down our novel

8    estimator for the biodiversity parameter $\theta$ as,

9

10
$$\hat{\theta} = E_k \left[ \theta_k \right] = E_k \left[ \frac{1}{\hat{F}_{inter}(k)} - 1 \right] \qquad (12)$$

11

12    where $E_k$ denotes the expectation over the $k$ local community samples.

13    While eqn (12) can be applied directly to simulated local community samples with a

14    known theoretical $\theta$, it needs to be adapted for a field dataset which may belong to a single or

15    several metacommunities. In the following we attempt to identify a subset of the field dataset

16    which is most likely to correspond to a 2L-SINM framework (i.e. a speciation-drift

17    equilibrium at the metacommunity level). One possible way to go about this task is to

18    measure the spread of the $\hat{\theta}_k$ distribution and attempt to reduce it by using a sequential

19    elimination scheme which opts out field samples one at a time.

20    Let us assume that our field dataset consists of $Y$ community samples of variable size.

21    Our method consists of calculating the $\hat{\theta}_Y$ value of a given dataset along with a measure of

22    statistical deviation (denoted $COD_Y$, see below) and repeating the same by randomly pulling

12

1 out one sample at a time and computing $\hat{\theta}_{Y-1}$ and $COD_{Y-1}$. Among the $Y$ values of $\hat{\theta}_{Y-1}$ thus

2 obtained, we eliminate the sample whose absence produced the smallest $COD_{Y-1}$ value and

3 then proceed with the sequential elimination scheme with the remaining $Y-1$ samples. We

4 define our coefficient of deviation or COD (a robust analogue of the coefficient of variation)

5 as the ratio of the mean absolute deviation (MAD) over the average where

6

$$COD_k\left[\hat{\theta}_k\right] = \frac{MAD_k\left[\hat{\theta}_k\right]}{E_k\left[\hat{\theta}_k\right]} = \frac{E_k\left[\left|E_k\left[\hat{\theta}_k\right] - \hat{\theta}_k\right|\right]}{E_k\left[\hat{\theta}_k\right]}. \tag{13}$$

8

9 The MAD is a well known robust statistic of the sample variation when the population

10 distribution is unknown or for highly skewed curves commonly known as heavy-tailed

11 distributions. For a normal distribution, the standard error can be roughly calculated as 1.253

12 times its MAD value (MathWorks 2008). This descriptor should not be confused with the

13 closely related median absolute deviation (also known as MAD) which uses a median instead

14 of a mean. The elimination scheme was finally stopped with the appearance of a consistent

15 asymptotic pattern of the sequentially obtained COD values which indicated a stable estimate

16 for a network of samples.

17

18 **Applications**

19 <u>Simulating multiple local community samples</u>

20 We simulated steady state local community samples following the modification of the

21 sequential construction scheme of Etienne (2005, Appendix S2). The simulation algorithm

22 basically follows the 3L-SINM structure with $m$ for the intermediate pool set to 1 (refer to

23 Fig. 1.) thereby producing samples strictly congruent with the 2L-SINM. Accordingly, we

1    create an explicit link between the ancestry information of the community samples and a large

2    predefined metacommunity (see Munoz *et al.* 2007, Appendix).

3          Every simulation consists of a set of $N$ local community samples, each having a

4    randomly chosen immigration parameter $I_k$ (varying from 3 to 300) and sample size $n_k$

5    (varying from 200 to 600). Our criterion for the sample size corresponds approximately to the

6    size of a hectare of tropical forest (i.e. $n_k = 400$ trees above 10 cm of diameter) typically found

7    in field studies involving several permanent sampling plots (Pyke *et al.* 2001; Ramesh *et al.*

8    2010b). We also varied the number of samples (i.e. $N$) by simulating 5, 10, 20, 30 and 50

9    samples. These simulations also need a biodiversity parameter to be defined for which we

10    simulated sets of scenarios where $\theta = (10, 50, 100, 200, 300)$. In the following, we shall to a

11    simulated sampling protocol (or SSP hereafter) as a simulation generated using the

12    information provided by the couplet $(N, \theta)$ as the other parameters are chosen to vary at

13    random (i.e. $I_k$ and $n_k$). Thus, we have considered a total of 5 (values for $N$) x 5 (values for $\theta$)

14    = 25 SSPs, each of which was in turn replicated 200 times whereby we obtained a grand total

15    of 5000 estimates (denoted $\hat{\theta}$) of the theoretical biodiversity parameter $\theta$. We assessed the

16    performance of the estimation of $\theta$ by studying the histograms of the Relative Bias (RB) given

17    by $(\hat{\theta} - \theta)/\theta$ and the COD (cf. eqn (13)).

18

19    <u>Inferring $\theta$ using field data</u>

20          Apart from simulations, we estimated $\hat{\theta}$ using two tropical forest datasets each

21    consisting of multiple small permanent field plots. Both these datasets consist of the

22    abundance information of individuals above 10 cm of diameter at breast height which have

23    been identified to the species level. The first dataset consists of 50 plots (mainly 1ha) along

24    the Panama Canal Watershed (PCW) area set up by the CTFS (Pyke *et al.* 2001), a subset of

25    which was previously used by Jabot *et al.* (2008) for inferring immigration parameters. This

14

1    freely available dataset is part of a larger study (the Marena dataset) and the field plots used in

2    this paper are originally referred to using the following numbering 1-41, C1-C4 and S0-S4

3    (see Condit *et al.* 2002, Appendix; Chave *et al.* 2004, Appendix B). Our second dataset also

4    consists of 50 field plots (each of 1ha in size) of the wet evergreen forest type from the

5    Western Ghats (WG) region of South India (extracted from Ramesh *et al.* 2010a). This dataset

6    has been discussed previously by Munoz *et al.* (2007, Appendix) for the purpose of inferring

7    neutral parameters.

8

9    **Results**

10    We studied the RB (relative bias) and the COD (coefficient of deviation, see eqn (13))

11    histograms for the various SSPs (simulated sampling protocols). The 5000 SSP estimates of

12    $\hat{\theta}$ which make up the histograms were obtained in the matter of a few minutes using the

13    MATLAB® software (MathWorks 2008). In both the five and fifty SSPs (Figs. 2 and 3

14    respectively), the best fit normal curve clearly emphasized the increasing symmetry of the

15    distribution of the RB of $\hat{\theta}$ with increasing theoretical $\theta$ and suggested that the estimator is

16    unbiased. In general, the RB distribution of $\hat{\theta}$ showed a tendency to be skewed for low values

17    of $\theta$ while it became more symmetric and always remained centred around zero as $\theta$ increased

18    (see also Figure S1 in the Supporting Information). At the same time, the COD distribution

19    was skewed for a low number of samples (e.g. $N = 5, 10$) and a high theoretical $\theta$ and vice

20    versa for $N > 10$, while the COD skewness varied little (for low $\theta$) compared to the RB

21    skewness (Figure S1). However, note that the COD histograms (Figs. 2 and 3) rarely exceed a

22    maximum value of 0.2, which was used as a benchmark when estimating $\theta$ on field data for

23    large $N$.

24    We also applied our estimator to the two tropical forest datasets presented above (Fig.

25    4). When using all the 50 field samples of the PCW data, we obtained a high $COD_{50}$ value ($\approx$

1    0.8) in comparison to the COD histogram for the $N = 50$ strictly neutral SSPs (Fig. 3). In

2    contrast, the $COD_{50}$ for the WG data ($\approx$ 0.2) was well within the range (not shown).

3    Subsequently, we applied the sequential elimination scheme on both these datasets (cf.

4    previous section) in order to identify the network of plots composing the ideal (i.e. panmictic)

5    2L-SINM metacommunity. The estimation of the neutral biodiversity parameter for the WG

6    dataset proved to be comparatively stable while its respective COD values fluctuated slightly

7    (between 0.1 – 0.2) below the maximum value observed for strictly neutral simulations.

8    Furthermore, our estimation of $\hat{\theta}$ for the WG data was well bounded by the values 62.33 and

9    50.99 which are those found by Munoz *et al.* (2007) for the very same plots. For the PCW

10   dataset, $\hat{\theta}$ estimates reached a qualitatively stable estimate which corresponded to a COD $<$

11   0.2. This was obtained after having sequentially eliminated 15 plots (COD $<$ 0.2), though we

12   continued to eliminate samples in order to check for its stability (Fig. 4). Besides, a closer

13   look at the remaining PCW plots revealed that the first eight eliminated plots were part of the

14   Outer PCW region (numbered 31-39, see Pyke *et al.* 2001, Fig. 1).

15

16   **Discussion**

17          In this paper, we have basically introduced a new estimator for multiple field samples

18   of the neutral biodiversity parameter $\theta$, first formulated by Hubbell. Subsequently, this

19   estimator has been tested on wide-ranging simulations of multiple neutral local community

20   samples at migration-drift equilibrium. A general conclusion from our simulation study is that

21   the relative bias of $\hat{\theta}$ seems to be relatively well distributed around zero with a progressive

22   tightening of the same as $\theta$ increased. This property is highly desired given that currently

23   available estimators of $\theta$ for a single local community sample present an increasing bias with

24   increasing $\theta$ (Munoz *et al.* 2007). As for the estimation variance, measured using the COD

25   values, we note that the number of community samples used (i.e. $N$) is an important factor as

1   a fewer number (e.g. 5 and 10) leads to an increase in the spread of the COD distribution (Fig.

2   2). Our results can also be seen as a significant improvement in contrast to likelihood

3   approaches (Etienne 2009b) which become computationally intractable for more than five

4   samples of small size like the ones used in this paper (Beeravolu *et al. Submitted manuscript*).

5        However, the technique presented here needs to be further compared to existing

6   methods (Munoz *et al.* 2007; Etienne 2009a) that estimate $\theta$ from a single panmictic sample

7   and also make use of the analytical expectations for the Ewens Multivariate Distribution

8   (Ewens 1972) of population genetics literature. In particular, by randomly sampling an

9   individual from several spatially separated samples these authors (Munoz *et al.* 2007; Etienne

10  2009a) constitute a metacommunity sample which is repeated a number of times in order to

11  hone their estimates of $\hat{\theta}$. But this approach can be unreliable when the number of local

12  community samples is small as it would provide a (relatively small) metacommunity sample

13  with insufficient information for inferential purposes. Though, an added advantage of using

14  the Ewens estimation for $\theta$ is the availability of an analytical expression for the induced bias

15  which decreases with increasing sample size and increases with an increase in $\theta$ (Donnelly &

16  Tavaré 1995, 414; Tavaré & Ewens 1997, 236).

17       Moreover, as our estimates using the WG data seem to be relatively robust to the

18  sequential elimination scheme and coherent with Munoz *et al.*'s (2007) estimates, the 50-

19  sample evergreen forest dataset from the WG seems to strongly corroborate a 2L-SINM at

20  this particular sampling scale. At the same time, Ramesh *et al.* (2010b) have found that some

21  environmental variables had a strong predictive power on the plots' floristic composition.

22  This implies that while the data seem to agree with the neutral model, violations from neutral

23  assumptions might not hinder a sound estimation of $\hat{\theta}$ as a phenomenological descriptor of

24  the overall diversity of the region. Instead, estimating $\hat{\theta}$ may be a good basis to compare the

25  overall diversity between biogeographic regions as Fisher's $\alpha$ is known to be asymptotically

1    identical to $\theta$ (Hubbell 2001, 165). This approach could also be extended to other forest types

2    and biogeographic regions in order to identify the extents of the different metacommunities as

3    discussed above and measure their relative diversity.

4        To pursue this idea further, we can use the distribution of the $\hat{F}_{inter}(k)$ values in order

5    to identify the local communities whose top-down linkage to a common metacommunity may

6    be unlikely. Conversely, it is also a simple approach which delineates a subset of samples that

7    look "floristically homogeneous" with respect to the 2L-SINM. Such a group of field plots

8    whose taxonomic composition complies with the 2L-SINM of a metacommunity could be

9    used to spatially delimit or map forest types over a regional sampling design. While the

10   elimination technique presented in this paper can be seen as a simple top-down clustering

11   scheme, it produces results which bear a close resemblance to well known bottom-up

12   ordination schemes such as hierarchical agglomerative clustering. For example, Chust *et al.*

13   (2006) used spatially explicit information in relation to a field plot's ecology such as its

14   elevation, remotely sensed data and the geographical distance to predict the forest types for

15   almost all of the PCW data used in this paper. They perform a hierarchical agglomerative

16   clustering of the sample abundances (or occurrences) using a proportional-link linkage

17   algorithm and further extrapolate each cluster's spatially explicit characteristics using a

18   multiple regression model thereby mapping a forest type. The most "distant" clusters found in

19   their study (Chust *et al.* 2006, Appendix 2) correspond perfectly to some of the first few plots

20   eliminated using our sequential elimination technique. Although, note that Chust *et al.* (2006,

21   Figs. 1 and 4) exclude some of the plots used in this paper (numbered 38 and 39) and instead

22   use other plots (numbered P1, P2, G1, G2 and SH) which were absent from the present (see

23   Chave *et al.* 2004, Appendix B for a corrected account of the PCW field plot numbering).

24       For the case of a single large local community sample, Jabot & Chave (2009), echoing

25   Harte (2003), contend that species abundances contain a limited amount of useful information

1    and supplement it by the use of species phylogenetic information in order to resolve

2    inconsistencies in the estimation of neutral parameters. Our results set in a more complex

3    context raise the question whether species phylogenies are truly needed when multiple

4    spatially separated field samples are available. Nevertheless, the $F_{inter}(k)$ statistic (eqn (11))

5    presented here can easily be developed into an efficient Bayesian framework (*sensu* Jabot &

6    Chave 2009), which remains a very powerful method, for injecting additional information

7    such as species phylogenies or even the demographic history of the communities (Beaumont

8    & Rannala 2004). Moreover, recent developments in the field of theoretical population

9    genetics make use of a comparable inter-sample similarity metric (Gaggiotti & Foll 2010)

10   under the island model of Wright (1931), which is a classic model of population subdivision.

11   Interestingly, note that the island model comes conceptually close to the mainland island

12   model of MacArthur & Wilson (1967) for the case of an infinite number of islands, in which

13   case it is known as the continent-island model (Wilkinson-Herbots 1998, 574) or an island-

14   mainland metapopulation (Rannala & Hartigan 1995) although there are subtle differences to

15   be taken into account (in terms of demographic assumptions).

16   Finally a major weakness of almost all neutral approaches is that it is an equilibrium

17   theory which nevertheless has greatly facilitated its mathematical development. Though, truly

18   dynamic neutral modes are desperately lacking in community ecology (but see Leigh *et al.*

19   1993; Gilbert *et al.* 2006) and some initial steps have been taken in this direction

20   (Vanpeteghem & Haegeman 2010), much needs to be done before we are able to infer the

21   parameters of a dynamic model from field data. However, the main improvement presented in

22   our paper is a simple and computationally efficient approach for estimating the biodiversity

23   parameter $\theta$ (in the case of a multiple sample 2L-SINM framework) which is in many ways

24   complementary to the estimation of the multiple sample $I_k$ parameter (Munoz *et al.* 2008).

25

1 **Acknowledgements**

5

6 **References**

7
8 Beaumont M.A. & Rannala B. (2004). The Bayesian revolution in genetics. *Nature Reviews*

9       *Genetics*, 5, 251-261.

10 Beeravolu C.R., Couteron P., Pélissier R. & Munoz F. (2009). Studying ecological

11       communities from a neutral standpoint: A review of models' structure and parameter

12       estimation. *Ecol. Model.*, 220, 2603-2610.

13 Beeravolu C.R., Munoz F. & Couteron P. On the simultaneous maximum likelihood inference

14       of neutral parameters from several local community samples. *Submitted manuscript*.

15 Chave J., Condit R., Aguilar S., Hernandez A., Lao S. & Perez R. (2004). Error propagation

16       and scaling for tropical forest biomass estimates. *Philosophical Transactions of the*

17       *Royal Society of London. Series B: Biological Sciences*, 359, 409.

18 Chave J. & Leigh E.G. (2002). A spatially explicit neutral model of beta-diversity in tropical

19       forests. *Theor. Popul. Biol.*, 62, 153-168.

20 Chust G., Chave J., Condit R., Aguilar S., Lao S. & Perez R. (2006). Determinants and spatial

21       modeling of tree beta-diversity in a tropical forest landscape in Panama. *Journal of*

22       *Vegetation Science*, 17, 83-92.

1   Condit R., Pitman N., Leigh E.G., Chave J., Terborgh J., Foster R.B., Nunez P., Aguilar S.,

2       Valencia R., Villa G., Muller-Landau H.C., Losos E. & Hubbell S.P. (2002). Beta-

3       diversity in tropical forest trees. *Science*, 295, 666-669.

4   Donnelly P. & Tavaré S. (1995). Coalescents and genealogical structure under neutrality.

5       *Annu. Rev. Genet.*, 29, 401-421.

6   Dornelas M., Connolly S.R. & Hughes T.P. (2006). Coral reef diversity refutes the neutral

7       theory of biodiversity. *Nature*, 440, 80-82.

8   Etienne R.S. (2005). A new sampling formula for neutral biodiversity. *Ecol. Lett.*, 8, 253-260.

9   Etienne R.S. (2009a). Improved estimation of neutral model parameters for multiple samples

10      with different degrees of dispersal limitation. *Ecology*, 90, 847-852.

11  Etienne R.S. (2009b). Maximum likelihood estimation of neutral model parameters for

12      multiple samples with different degrees of dispersal limitation. *J. Theor. Biol.*, 257,

13      510-514.

14  Etienne R.S., Latimer A.M., Silander J.A. & Cowling R.M. (2006). Comment on "Neutral

15      ecological theory reveals isolation and rapid speciation in a biodiversity hot spot".

16      *Science*, 311.

17  Etienne R.S. & Olff H. (2004). A novel genealogical approach to neutral biodiversity theory.

18      *Ecol. Lett.*, 7, 170-175.

19  Ewens W.J. (1972). The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.*, 3,

20      87-112.

1  Gaggiotti O.E. & Foll M. (2010). Quantifying population structure using the F-model.

2      *Molecular Ecology Resources*, 10, 821-830.

3  Gilbert B., Laurance W.F., Leigh E.G. & Nascimento H.E.M. (2006). Can neutral theory

4      predict the responses of amazonian tree communities to forest fragmentation? *Am.*

5      *Nat.*, 168, 304-317.

6  Gravel D., Canham C.D., Beaudet M. & Messier C. (2006). Reconciling niche and neutrality:

7      the continuum hypothesis. *Ecol. Lett.*, 9, 399-409.

8  Gray J.S., Bjorgesaeter A. & Ugland K.I. (2006). On plotting species abundance distributions.

9      *J. Anim. Ecol.*, 75, 752-756.

10  Haegeman B. & Etienne R.S. (2008). Relaxing the zero-sum assumption in neutral

11      biodiversity theory. *J. Theor. Biol.*, 252, 288-294.

12  Haegeman B. & Loreau M. (2011). A mathematical synthesis of niche and neutral theories in

13      community ecology. *J. Theor. Biol.*, 269, 150-165.

14  Harte J. (2003). Tail of death and resurrection. *Nature*, 424, 1006-1007.

15  Hubbell S.P. (1979). Tree Dispersion, Abundance, and Diversity in a Tropical Dry Forest.

16      *Science*, 203, 1299-1309.

17  Hubbell S.P. (2001). *The unified neutral theory of biodiversity and biogeography*. Princeton

18      University Press, Princeton, N.J.

19  Jabot F. & Chave J. (2009). Inferring the parameters of the neutral theory of biodiversity

20      using phylogenetic information and implications for tropical forests. *Ecol. Lett.*, 12,
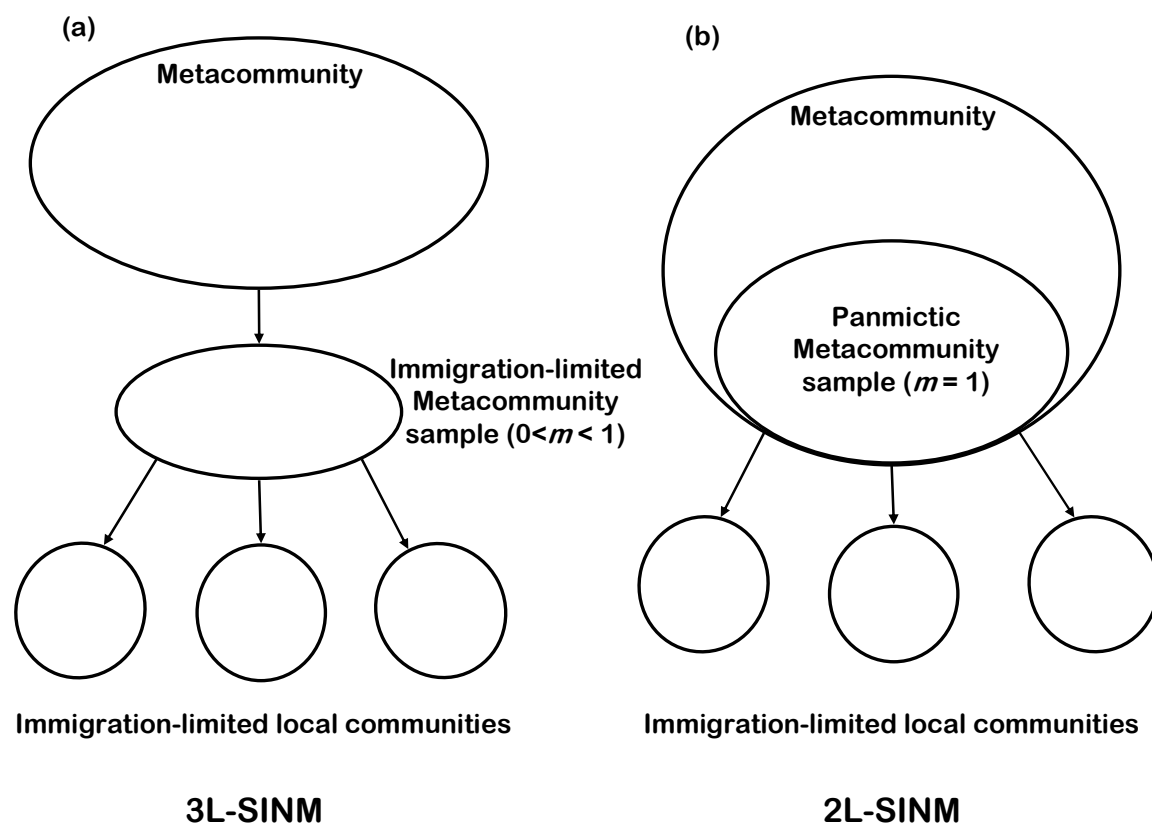
21      239-248.

1   Jabot F., Etienne R.S. & Chave J. (2008). Reconciling neutral community models and

2       environmental filtering: theory and an empirical test. *Oikos*, 117, 1308-1320.

3   Kopp M. (2010). Speciation and the neutral theory of biodiversity. *Bioessays*, 32, 564-570.

4   Leigh E.G. (2007). Neutral theory: a historical perspective. *J. Evol. Biol.*, 20, 2075-2091.

5   Leigh E.G., Wright S.J., Herre E.A. & Putz F.E. (1993). The decline of tree diversity on

6       newly isolated tropical islands: a test of a null hypothesis and some implications. *Evol.*

7       *Ecol.*, 7, 76-102.

8   Levin S.A. (1992). The problem of pattern and scale in ecology: the Robert H. MacArthur

9       award lecture. *Ecology*, 73, 1943-1967.

10  MacArthur R.H. & Wilson E.O. (1967). *The theory of island biogeography*. Princeton

11      University Press, Princeton, N.J.

12  MathWorks (2008). Matlab R2008a, version 7.6.0.324. In. The MathWorks, Inc. Natick, MA.

13  McGill B.J., Etienne R.S., Gray J.S., Alonso D., Anderson M.J., Benecha H.K., Dornelas M.,

14      Enquist B.J., Green J.L., He F.L., Hurlbert A.H., Magurran A.E., Marquet P.A.,

15      Maurer B.A., Ostling A., Soykan C.U., Ugland K.I. & White E.P. (2007). Species

16      abundance distributions: moving beyond single prediction theories to integration

17      within an ecological framework. *Ecol. Lett.*, 10, 995-1015.

18  McGill B.J., Maurer B.A. & Weiser M.D. (2006). Empirical evaluation of neutral theory.

19      *Ecology*, 87, 1411-1423.

20  Moran P.A.P. (1958). Random processes in genetics. *Proc. Camb. Phil. Soc.*, 54, 60-71.

1    Munoz F., Couteron P. & Ramesh B.R. (2008). Beta Diversity in Spatially Implicit Neutral

2           Models: A New Way to Assess Species Migration. *Am. Nat.*, 172, 116-127.

3    Munoz F., Couteron P., Ramesh B.R. & Etienne R.S. (2007). Estimating parameters of neutral

4           communities: From one single large to several small samples. *Ecology*, 88, 2482-

5           2488.

6    O'Dwyer J.P. & Green J.L. (2009). Field theory for biogeography: a spatially explicit model

7           for predicting patterns of biodiversity. *Ecol. Lett.*, 13, 87-95.

8    Purves D.W. & Pacala S.W. (2005). Ecological drift in niche structured communities: neutral

9           pattern does not imply neutral process. In: *Biotic Interactions in the Tropics* (eds.

10          Burslem D, Pinard M & Hartley S). Cambridge University Press Cambridge, pp. 107–

11          138.

12   Pyke C.R., Condit R., Aguilar S. & Lao S. (2001). Floristic composition across a climatic

13          gradient in a neotropical lowland forest. *Journal of Vegetation Science*, 12, 553-566.

14   Ramesh B.R., Swaminath M.H., Patil S.V., Pélissier R., Venugopal P.D., Aravajy S., Elouard

15          C. & Ramalingam S. (2010a). Forest stand structure and composition in 96 sites along

16          environmental gradients in the central Western Ghats of India (Ecological Archives

17          E091-216). *Ecology*, 91, 3118-3118.

18   Ramesh B.R., Venugopal P.D., Pélissier R., Patil S.V., Swaminath M.H. & Couteron P.

19          (2010b). Mesoscale Patterns in the Floristic Composition of Forests in the Central

20          Western Ghats of Karnataka, India. *Biotropica*, 42, 435-443.

21   Rannala B. & Hartigan J.A. (1995). Identity by Descent in Island-mainland Populations.

22          *Genetics*, 139, 429-437.

1  Simpson E.H. (1949). Measurement of Diversity. *Nature*, 163, 688-688.

2  Tavaré S. & Ewens W.J. (1997). Multivariate Ewens distribution. In: *Discrete Multivariate*

3     *Distributions* (eds. Johnson NL, Kotz S & Balakrishnan N). Wiley New York, pp.

4     232-246.

5  Vallade M. & Houchmandzadeh B. (2003). Analytical solution of a neutral model of

6     biodiversity. *Phys. Rev. E*, 68.

7  Vanpeteghem D. & Haegeman B. (2010). An analytical approach to spatio-temporal

8     dynamics of neutral community models. *J. Math. Biol.*, 1-35.

9  Wilkinson-Herbots H.M. (1998). Genealogy and subpopulation differentiation under various

10    models of population structure. *J. Math. Biol.*, 37, 535-585.

11 Wright S. (1931). Evolution in Mendelian Populations. *Genetics*, 16, 97-159.

12 Zillio T., Volkov I., Banavar J.R., Hubbell S.P. & Maritan A. (2005). Spatial scaling in model

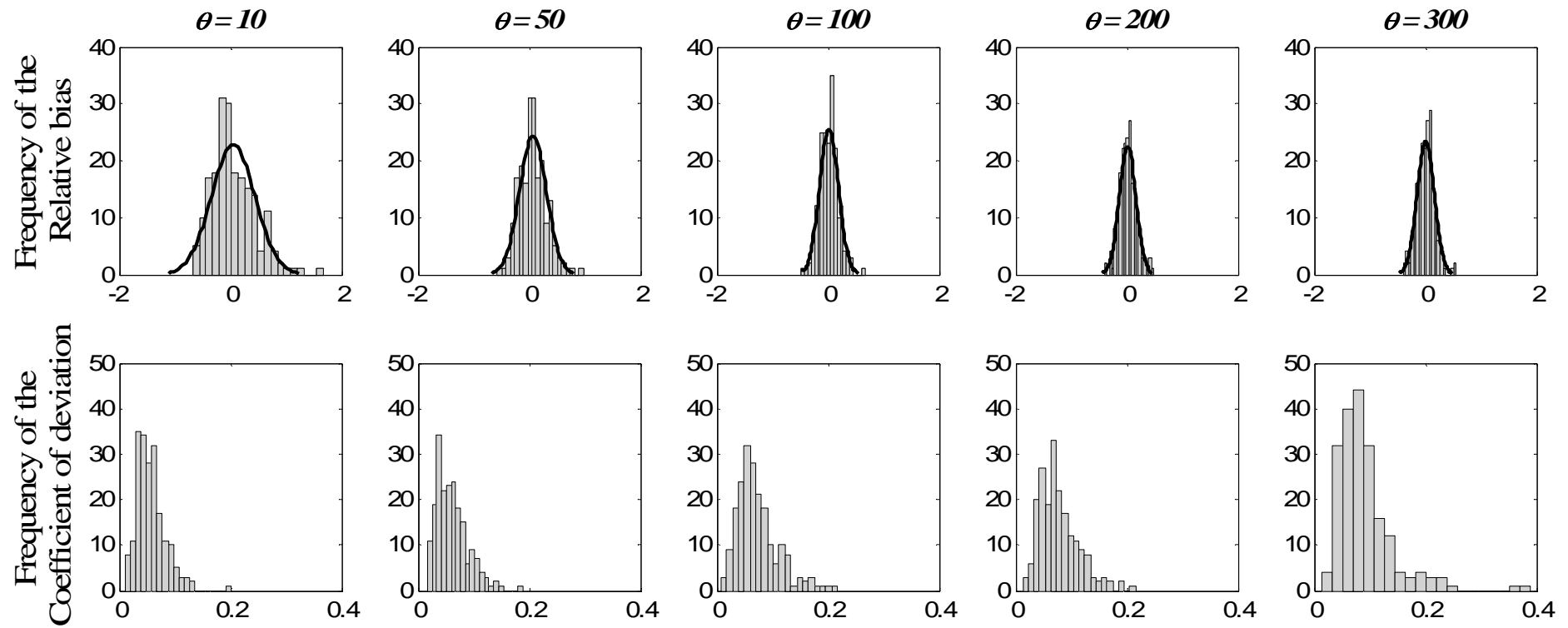13    plant communities. *Physical review letters*, 95, 98101-98101.

14

15

16

1 **Figures**

2

3 Figure 1

4



(a)

Metacommunity

Immigration-limited Metacommunity sample ($0 < m < 1$)

Immigration-limited local communities

**3L-SINM**

(b)

Metacommunity

Panmictic Metacommunity sample ($m = 1$)

Immigration-limited local communities
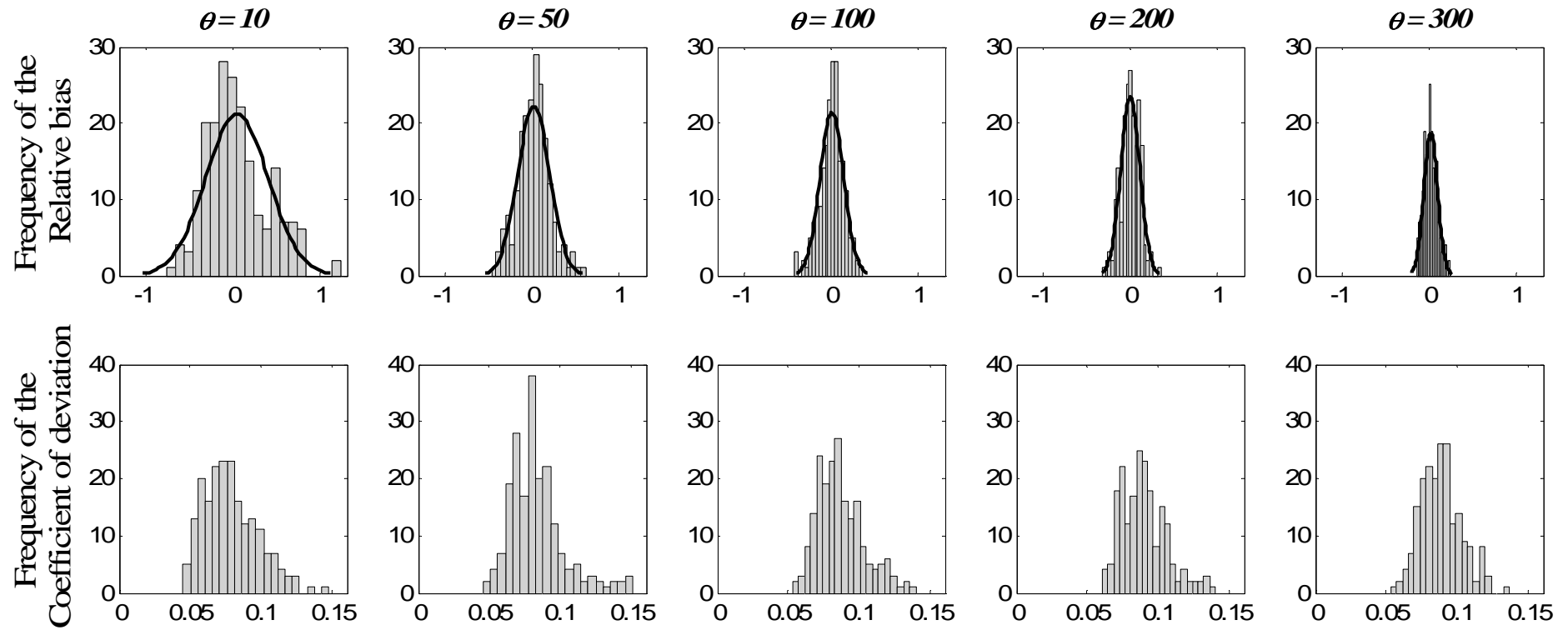
**2L-SINM**

5

6

7

2

3
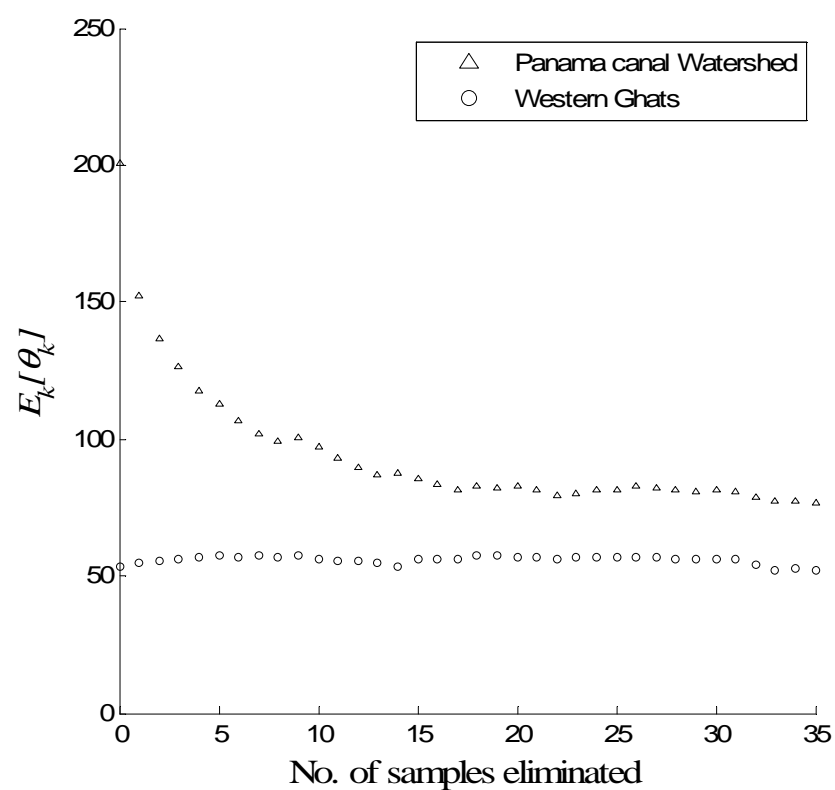
1    Figure 3

2

3

4

1    Figure 4

2



3

4

1    **Figure legends**

2

3    Figure 1 The hierarchical coupling of scale oriented processes in spatially implicit neutral models (SINM).

4    Processes such as speciation and extinction of species exist in the metacommunity and the relatively faster

5    demographic processes take place at the local community level. For Hubbell's (2001) two level model (2L-

6    SINM) an ideal panmictic metacommunity sample can be considered to be the source pool of multiple local

7    communities. Without loss of generality, we can assume that the pooled species composition of the dispersal-

8    limited local community samples is itself a immigration-limited metacommunity sample (the 3L-SINM)

9    corresponding to some sub-region of the same which gives rise to a three level hierarchical model (Beeravolu *et*

10    *al.* 2009). The value of the parameter *m* then acts as the degree of isolation from the metacommunity.

11

12    Figure 2 The distribution of the relative bias and the coefficient of deviation values of $\hat{\theta}$ (the estimated

13    biodiversity parameter) for a sampling protocol containing 5 simulated plots each and for five different

14    theoretical values of $\theta$ found in the literature. The solid black line is the best fit normal curve and emphasizes the

15    increasing symmetry of the distribution of the relative bias with increasing $\theta$.

16

17    Figure 3 The distribution of the relative bias and the coefficient of deviation values of $\hat{\theta}$ (the estimated

18    biodiversity parameter) for a sampling protocol containing 50 simulated plots each and for five different

19    theoretical values of $\theta$ found in the literature. The solid black line is the best fit normal curve and emphasizes the

20    increasing symmetry of the distribution of the relative bias with increasing $\theta$.

21

22    Figure 4 Estimating $\hat{\theta}$ from field data using eqn (13) and following a sequential elimination scheme (see main

23    text) in order to identify the network of plots composing the ideal 2L-SINM metacommunity. The sample

24    elimination criterion is determined by the maximum decrease in the coefficient of deviation due to the absence of

25    a particular plot. Here we sequentially eliminated 35 samples which appeared sufficient to highlight the

26    asymptotic pattern which indicated the stable $\hat{\theta}$ estimate for a network of plots.

27