Journal of Structural Geology, Vol. 11, No. 7, pp. 787 to 798, 1989
Printed in Great Britain

0191-8141/89 $03.00 + 0.00

A new theory for the static contact between rough, unmated surfaces in non-elastically deforming rock and its implications for rock friction

R. M. STESKY and S. S. HANNAN

Tectonics Study Group, Geological Sciences, University of Toronto, Erindale Campus, Mississauga, Ontario, Canada L5L 1C6

(*Received* 23 *November* 1988; *accepted in revised form* 28 *April* 1989)

Abstract—The closure behavior of fractures in marble and alabaster is markedly different from that in quartzite. The aperture decreases considerably more under normal stress and remains permanently reduced, for the same ratio of normal stress to unconfined compressive strength. Also, a larger permanent relative contact area develops between the surfaces of marble and alabaster than it does between surfaces of quartzite. The permanent contact area increases at an increasing rate with normal stress in marble and alabaster, unlike the nearly linear increase in quartzite. The failure of surface asperities of calcite and gypsum during closure accounts for these differences.

We modeled this process by considering the surfaces to consist of paraboloids lying on a flat plane and having a range of initial heights. Closure occurs by pressing a plane rigid surface against the 'hills', flattening their peaks, keeping the base area of the hills constant. To allow for a changing resistance to deformation, the contact stress is assumed to vary linearly with the shortening strain, to a first approximation.

This model was tested against measurements of fracture closure and contact area of rough surfaces of calcite marble with a known initial height distribution of surface peaks. The fit to the data is quite good. In all cases, the model shows that closure is accompanied by a decrease in contact strength of deforming asperities, suggested also by the cataclastic deformation observed petrographically. The number of contact spots and the total length of contact seen in profile are also reasonably well modeled.

These results have important implications for our understanding of frictional strength of fractures. The overall resistance to shear along rough surfaces depends upon the product of the shear strength and true area of the contacts, both of which are affected by normal stress. Application of this model approach shows that the initial frictional resistance of some fractures in ductile rocks can be directly related to the topographic characteristics of the surfaces and the rheological properties of the asperities.

INTRODUCTION

FRACTURES are very complex structures. They have rough surfaces and contact is made at only a few spots. Many physical processes, such as friction, wear, adhesion, electrical and thermal contact, fluid and electrical flow, and contact stiffness, depend on the detailed nature of these contacts. In particular, the shear stress to cause slip on these surfaces significantly depends on the area and load-carrying capacity of these contacts. Prior to slip, the surfaces undergo compression due to normal stress, a process that may modify the contacts. In other words, the normal stress *conditions* the surface prior to slip.

In order to understand the initial frictional resistance of fractures, we need to know more about this conditioning that prepares the surface for slip. If the surfaces are normally loaded, the area of the contact points grows and more contacts are made. This process has been adequately modeled where the surfaces can be considered to deform elastically (Greenwood & Williamson 1966, Walsh & Grosenbaugh 1979, Brown & Scholz 1985, 1986) and plastically (Pullen & Williamson 1972). Many materials, however, do not deform in these ideal ways, so these theories cannot be used. For these materials, the asperities, or surface peaks, deform by complex processes that change their load-carrying capacity by work hardening or softening. The resistance to shear then depends markedly on the normal load applied.

Calcite and gypsum are two minerals that deform in such a way (Stesky & Hannan 1987). Fractures containing these minerals develop a permanent closure that is a larger fraction of the total closure and a more rapid growth of relative contact area under initial loading than is predicted by elastic and plastic theories. The purpose of this brief paper is to show the deformation of surfaces in calcite marble and present a new theory to account for it. The theory incorporates the geometry of the rough surfaces and the rheology of the deforming asperities. The insights gained from this simple model will be applied to the problem of shear resistance under normal load. The deformation of another rock, alabaster, composed of fine-grained gypsum, was also studied (Hannan 1988); the results support the conclusions presented here and will be published elsewhere.

FRACTURE CLOSURE

The rocks used for our study were two varieties of calcite marble. For comparison, we also studied a quartzite, a rock that has been found to deform elastically under the conditions of our tests (Brown & Scholz 1985, 1986).

Cores, 15 mm in diameter, were cut to a length of 25 mm and polished. Each sample was axially loaded several times to a maximum uniaxial stress of about half of the unconfined compresive strength (i.e. to a stress of 19

MPa for the marbles, and 39 MPa for the quartzite). We chose the maximum load to be approximately that for the onset of dilatancy, to minimize any permanent damage to the rock during cyclic loading. Repeated loading produced a repeatable stress–shortening curve, even after removing and replacing the specimen in the loading frame. We used the axial stress and shortening displacement for the final cycle as our measure of deformation of the intact part of the core.

The conditioned cores were then cut in half normal to the core axis and the new faces were polished with grinding compound up to No. 1000 grit. At the final stage the surfaces were lapped with either Nos 36- or 80-grit powder to produce the required roughness. Repeated tests showed that the final roughness was statistically reproducible. Typical profiles are shown in Fig. 1.

We used two methods to obtain these profiles. In one, the ground surfaces were traversed with a digital sampling profilometer, a diamond-tipped stylus and motor controller with amplifier and digitizing electronics. In the other, two surfaces were coated with a low viscosity, cold-setting epoxy and lightly pressed together. Following hardening of the epoxy (generally after 10–15 min), the glued core was cut axially and petrographic thin sections prepared. Photomicrographs of the sectioned surfaces were then sampled with a digitizing tablet.

The latter method involved considerably more effort, but allowed us to examine and measure the surface geometry and permanent deformation after stressing, by applying a known load to the core during epoxy curing. For the quartzite samples, release fractures and birefringence appeared in the epoxy (normally intact and optically isotropic) suggesting that little or no stress remains in the epoxied contacts after load removal. In contrast, the unloaded epoxy was optically isotropic and intact in the marble samples, indicating that little elastic recovery occurred. Thus in both cases the observed contacts were permanent ones and few elastic contacts are preserved.

Following the procedure originally used by Greenwood & Williamson (1966), *composite* profiles were calculated from pairs of surface profiles, grouped as in Fig. 1. The profiles were combined by calculating, point by point, the gap or aperture between them and subtracting the result from the maximum aperture found. The new rough surface has a height measured relative to the deepest composite valley (corresponding to the widest gap).

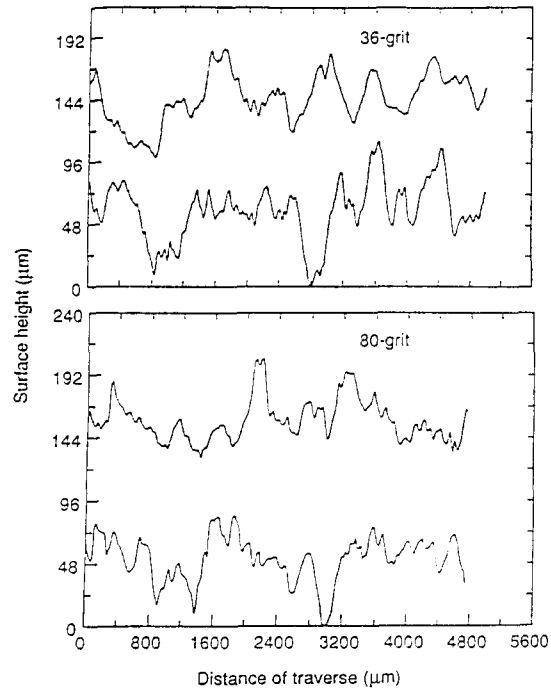

Fig. 1. Representative topographic profiles of No. 80- and No. 36-grit surfaces in calcite marble. The two profiles of each set represent the top and bottom surfaces. For the No. 36-grit surface, the profiles were measured by a profilometer and the data of one were inverted and positioned above the other at an arbitrary distance. The No. 80-grit profiles were measured from a thin section of a glued sample and show the true separation of the faces. In both cases, the heights are measured relative to that of the deepest valley on the bottom surface.

To evaluate the surface roughness, a computer procedure was used to select peaks on the profiles. The composite profiles were scanned using a moving window of 19 data points (i.e. a width of 95 μm). The center point of the window was tested against the nine points on either side; if it was a maximum, then a peak was counted and its position stored. Otherwise, the window was moved by one position and the test repeated. The width of the window was chosen rather subjectively by comparing the computer peaks with those estimated by visual inspection. With the 95 μm window the computer picked the major peaks without including the small irregularities on the sides of these peaks. The wide

Table 1. Characteristics of the rocks studied

Rock	Composition	Grain size (mm) (average)	Porosity (%)
Georgia marble	Calcite with about 3–4% of quartz, tremolite, phlogopite	0.7–2.2 (1.4)	0.5
Carrara marble	Calcite with less than 0.5% muscovite	0.12–0.35 (0.2)	0.2
Bar River quartzite	Quartz with 2–3% muscovite	0.25–0.65 (0.35)	0.2

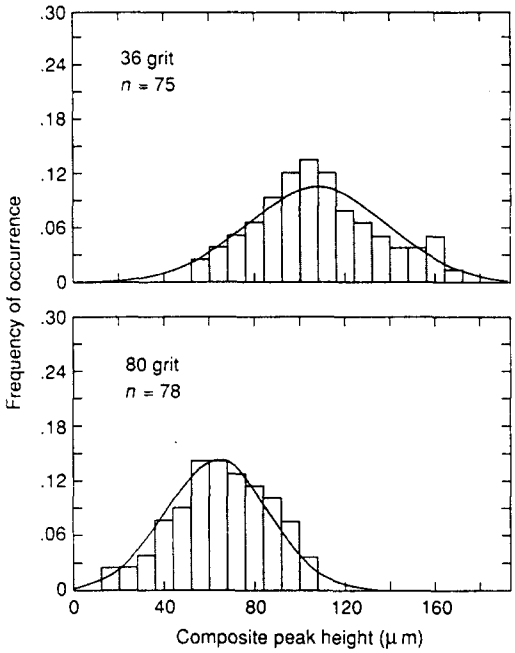

Fig. 2. Histograms of the heights of peaks of the composite surfaces calculated from several pairs of profiles such as those shown in Fig. 1. The smooth curves are the Gaussian normal distribution functions fitted to the experimental data. The function parameters are listed in Table 2. The number of peaks sampled, n, is shown for each surface.

window also eliminated the horizontal sampling problem described by Whitehouse (1978).

We used a different method to characterize the surface roughness from that used by most workers. Unless this difference is recognized, it may lead to some misunderstanding about our results. Most workers have used each surface sample point as though it were an individual mechanically isolated asperity of whatever shape is assumed (e.g. hemispherical, rod-shaped, conical). Each of these 'asperities' has a height corresponding to the height of that sample point on the surface. We find this approach physically unreasonable. Rather, we sampled the surface and identified the actual peaks in the profile, within the width of our chosen window. Such asperities are more likely to behave as mechanically isolated features.

Figure 2 shows histograms of composite peak heights, averaged over at least four profiles. Because of the relatively small number of peaks measured, the histograms are not smooth. Nevertheless, the distribution is approximately symmetric and normal. A Gaussian normal function was fitted to the data and the smoothed distribution was used for all subsequent calculations. The peaks measured in these profiles do not, in general, correspond to the true asperity summits, since a random section would only rarely intersect an asperity at its true tip. The peaks are somewhat lower than the actual tips, giving a lower mean height than is actually present. Since the height used in the calculation is arbitrary, because the choice of base level is arbitrary, we do not think this is a major problem; only the estimate of surface separation should be affected. The dispersion of heights will be different, however; the sectioning effect

Table 2. Topographic properties of peaks on the composite surfaces

Rock	Grit size	Mean peak height, h_m (μm)	Standard deviation of peak height, Δ (μm)	Radius of curvature of peak tip, β (μm)
Marble	36	107	31	28
	80	64	23	24
Quartzite	36	72	19	33
	80	48	10	31

will tend to bias the distribution towards lower heights, skewing the distribution slightly. Because of the low number of peaks counted, such skewing is not obvious in our data and will be neglected.

As others have observed, the surfaces lapped with the coarser powder had a higher mean peak height, h_m, and larger standard deviation, Δ. As well, for the same grit size, quartzite had lower h_m and Δ than did marble. In fact, the No. 36-grit quartzite surface had about the same roughness as did No. 80-grit marble. Table 2 shows these measurements in detail. The mean radius of curvature of peak shown in Table 2 is calculated by fitting a quadratic curve to the nine data points centered on the peak and computing the derivatives of the fitted curve. Using nine data points for the fitting produces minimal quantization error (Whitehouse 1978).

After roughening the cut surfaces, the cores were reassembled and reloaded to the same peak stress as were the intact cores. We calculated the fracture closure as the difference in displacement between the cut and intact cores for the same stress. This technique results in greater uncertainty than for a single direct measurement, in part because of the two required measurements and because of the necessary assumption that the intact portion of the cores and the steel–rock interfaces did not change in any way between measurements. We had no way to verify this assumption, but found that repeating the lapping and loading procedures produced closure values within about 2–3 μm, for maximum closures of 50–100 μm. Several such runs were averaged for the final closure curves shown in Fig. 3.

All rocks show the same general behavior: a large initial displacement under light loads, but less displacement at higher loads. The rocks differ in the magnitude of total displacement at a given load and in the unloading behavior. Quartzite suffered the least closure displacement and almost totally recovered its displacement upon unloading. The amount of permanent closure was less than 14% of the total closure. In contrast, marble underwent large closures, about 80% of which were permanent.

Because the roughness was about the same for all rocks, these differences reflect the relative strength or hardness of the minerals comprising these rocks and represent the extremes of a range of behavior. In an earlier study Tanoli (1982) found that the amount of permanent closure is inversely correlated with fracture stiffness, for rocks lapped with the same grit compound. Fracture stiffness is the rate of change of stress with

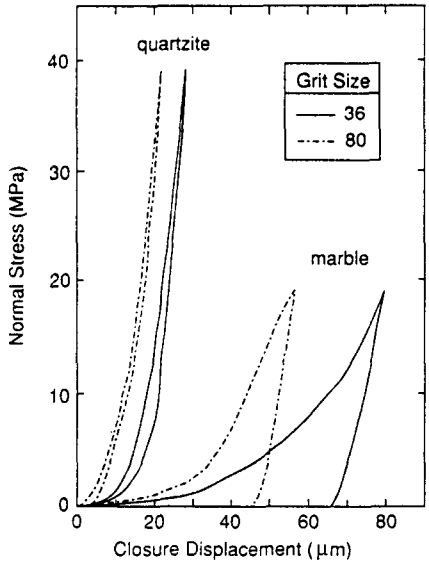

Fig. 3. Closure curves for ground surfaces in marble and quartzite with two roughnesses. These curves are averaged over several runs.

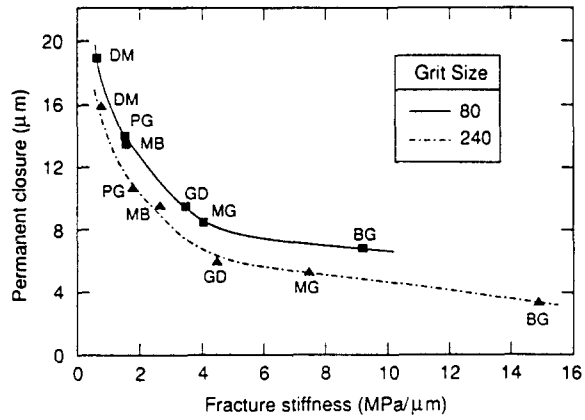

Fig. 4. Comparison of the fracture stiffness of several rocks at 25 MPa to the permanent closure after releasing the load. The curves through the data are drawn by eye. The rock type codes are: BG = Barre granite, MG = metagabbro, GD = granodiorite, MB = metabasalt, PG = altered pyroxene granulite, DM = dolomite marble. These rocks were used in studies of seismic velocity, electrical conductivity and permeability in fractured rock (Stesky 1985, 1986). (After Tanoli 1982.)

closure displacement. Figure 4 shows these measurements. Although the roughness is probably not the same for all rocks, the presence of weaker minerals, such as dolomite in DM and chlorite and sericite in PG and MB, gave greater permanent closures and lower stiffnesses than did the presence of the stronger quartz and feldspar, which dominate GD, MG and BG.

This difference between hard and soft minerals is quite striking. During the past 20 years, considerable effort has been made to understand and model the closure of rough surfaces. For elastic and plastic materials, the effort has been successful (for example, Greenwood & Williamson 1966, Pullen & Williamson 1972, Walsh & Grosenbaugh 1979, Brown & Scholz 1985, 1986). Our quartzite samples clearly behaved largely elastically, as others have found with similar rocks. Although we do not present the results here, the elastic model of Greenwood & Williamson fits our measurements quite well (Hannan 1988).

Marble definitely is not elastic at the stresses used in this study. But is a plastic model, such as that of Pullen & Williamson (1972), adequate to explain our results? In order to answer this question, we needed to understand the processes occurring on the deforming surfaces. Two further studies were undertaken: petrographic examination of deformed surfaces and measurement of the permanent relative contact area developed during closure. We found that calcite does not deform as a plastic material, so a new rheological model was needed to account for fracture closure in rocks containing minerals such as calcite.

PETROGRAPHIC OBSERVATIONS

Thin section views of contacting surfaces at various loads are shown in Figs. 5 and 6(a) & (b). At the lightest load, no contacts appear. Since very few contacts are likely at first, the probability of seeing one in any random section is small. The walls of the 'fracture' are only slightly damaged by the lapping process. At higher loads an increasing number of contacts are evident. The asperities are flattened mainly by brittle cracking, crystalline twinning and bending of pre-existing twins. The damage zone tends to be limited to the region very close to the contact; occasionally, however, it extends into the wallrock (Fig. 6a & b). Brittle cracking would lead to a loss of load-carrying ability of the asperities, at a rate that should depend on the extent of cracking and hence on the amount of flattening.

CONTACT AREA

The observation of the increasing number of contacts with load indicates that the contact area increases with load. We determined the actual contact area in two ways. Firstly, we measured the length of contact as seen in thin section as a fraction of the total section length. This length ratio is an estimate of the relative contact area, the ratio of true contact area to the nominal surface area.

Secondly, the epoxied marble cores were dissolved in dilute hydrochloric acid, leaving a cast of the fracture, here called a *resin film* (Fig. 6c). Where the two surfaces touched each other, contacts appeared as holes in the film. We measured the number of contact holes within a 13 × 13 mm grid, as well as the hole diameters in two perpendicular directions. The area of each hole was calculated assuming a simple elliptical shape. The ratio of total hole area to grid area is the relative contact area.

The results of these two kinds of measurements are given in Fig. 7, showing good agreement between the two methods. As we found earlier (Stesky & Hannan 1987), the relative contact area increased with stress, at a rate that increased with stress, reaching about 15% at 19

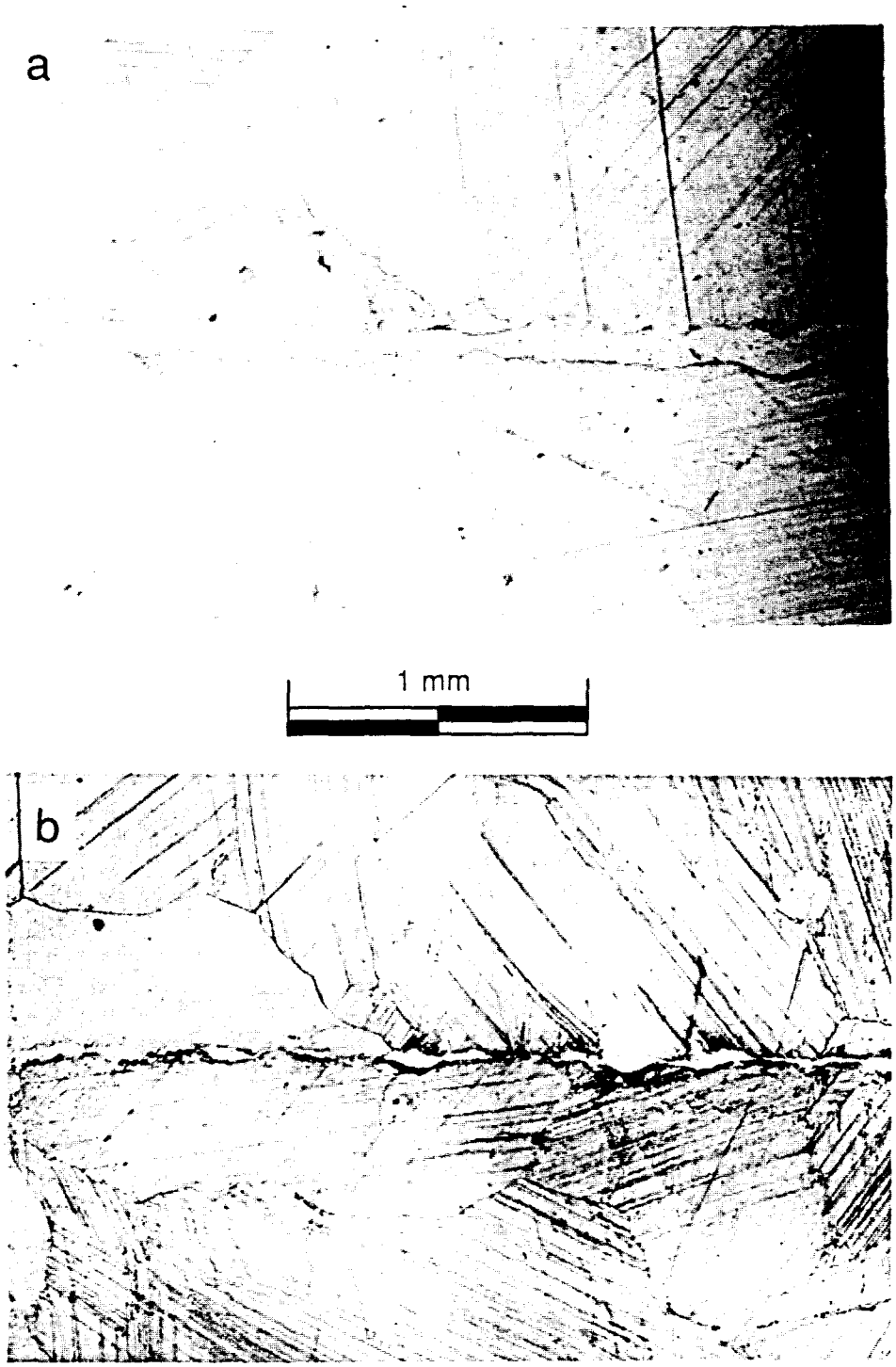

Fig. 5. Photomicrographs of thin sections of epoxied No. 80-grit surfaces of calcite marble (a) under very light loading, and (b) under a stress of 19 MPa. Note the large amount of contact in the loaded sample. The chance presence of a small quartz grain to the right of center in the loaded sample illustrates the indentation that occurs when materials of markedly different hardness come into mutual contact; because of the low concentration of quartz in this rock, such deformation is rare. Plane light; scale bar shown.

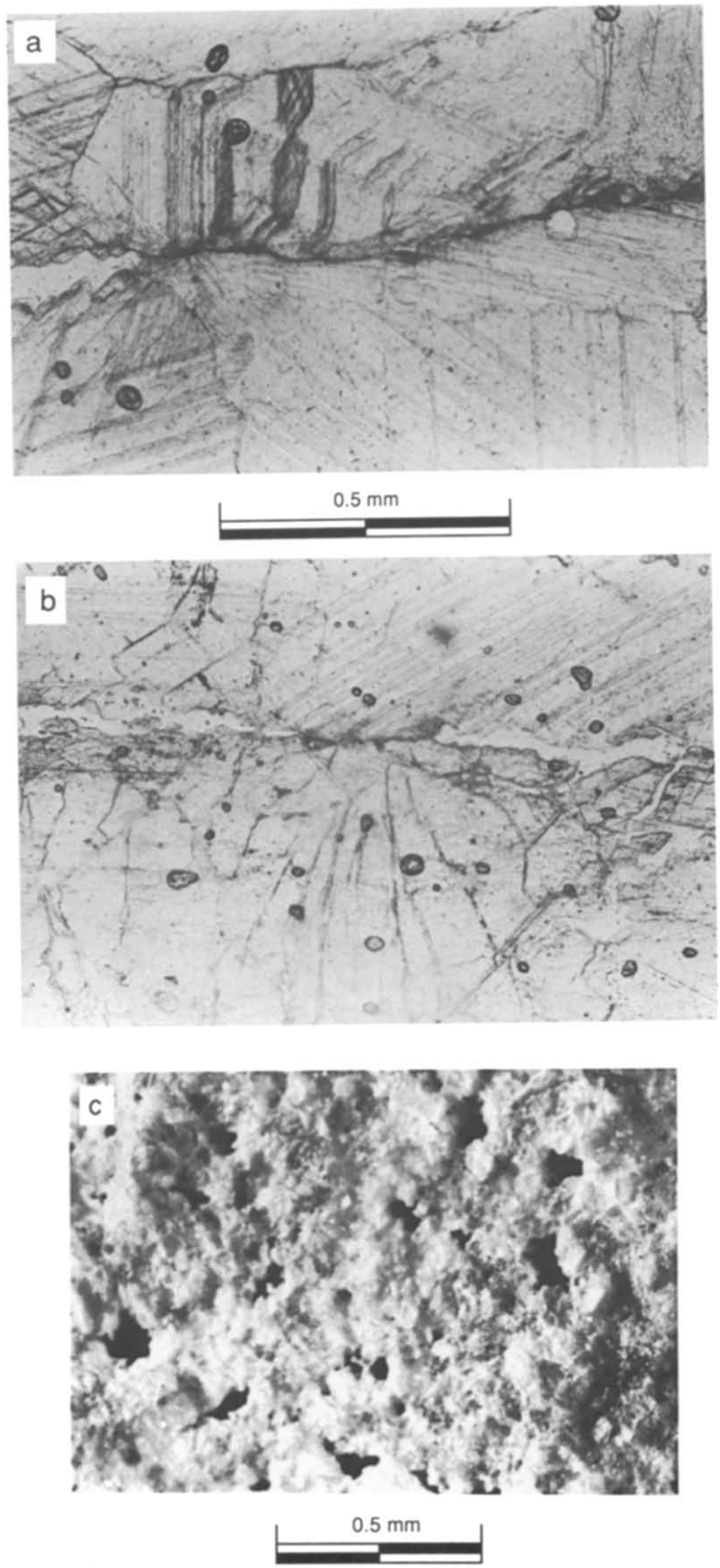

Fig. 6. (a & b) Detailed views of deformation at calcite–calcite contacts at 19 MPa stress, plane light; scale bar shown between (a) and (b). (a) Cracking and the formation of twin lamellae and bending of pre-existing lamellae. (b) Cracking. (c) Surface topography of a *resin film*, an epoxy cast of a 'fracture' surface in marble, formed under 19 MPa stress. The solid black areas are holes where the two surfaces were in contact. Reflected light; scale bar shown below.

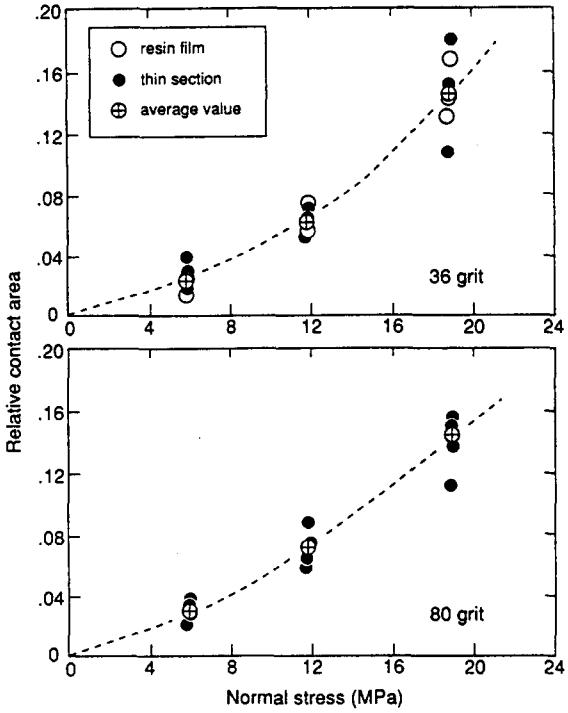

Fig. 7. Measurements of relative contact area at various stresses in marble. The value for each thin section is calculated from measurements over the entire section. Note the good agreement of the two methods. The dashed lines are smooth curves drawn through the average points. These lines are distinctly concave upward.

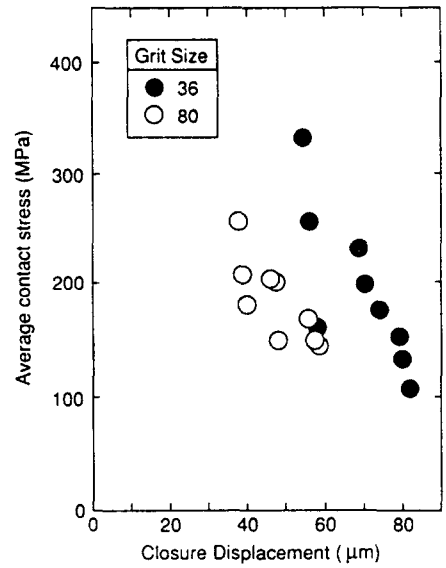

Fig. 8. The variation with closure displacement of the calculated mean contact stress, the product of the nominal applied stress and the relative contact area, showing the decrease in contact stress with closure.

MPa. The growth in contact area is even more striking for alabaster, reaching almost 30% at only 12 MPa (Stesky & Hannan 1987, Hannan 1988). In contrast, quartzite had a permanent relative contact area of less than 1% at 49 MPa, indicating that most of the deformation was elastic. Our results agree well with those of Logan & Teufel (1986) who measured 18% contact area developed after a small amount of slip in limestone at 25 MPa and about 2% contact area in sandstone at 50 MPa.

This acceleration in growth of contact area for marble and alabaster differs from that predicted for elastic and plastic materials. Elastic theory predicts a near linear increase (Greenwood & Williamson 1966), while plastic theory suggests an initial linear increase followed by a diminishing increase at higher stresses (Pullen & Williamson 1972). Again we find evidence that a different deformation behavior must be assumed for materials such as calcite and gypsum. We suggest that our results are compatible with a strain-weakening deformation process associated with the collapse of asperities by cataclasis. The deformation is localized at the asperities and the weakening requires a more rapid growth in contact area to accommodate the increase in stress. This effect is offset to some extent by the increase in number of contacts at higher loads.

Another way to see this behavior is to follow the variation of average contact stress, $\bar{\sigma}_c$,with fracture closure (Fig. 8). $\bar{\sigma}_c$ is the ratio of applied stress to the relative contact area and is a measure of the strength of the contacts. As shown in the figure, $\bar{\sigma}_c$ decreases with increasing closure. The large gap in data for low closures is present because considerable closure occurred before the applied stress reached the lowest value used for the contact area measurement. Our values for No. 80-grit surfaces agree very well with those found by Logan & Teufel (1986) during frictional sliding. Their measurements at 25 MPa indicate that the average contact stress levels out at normal stresses above those used in our study.

NEW MODEL FOR FRACTURE CLOSURE

The logic behind our model is similar to that of the Greenwood–Williamson model, but the details differ. As in their case, the two rough surfaces are combined into one composite surface, as described earlier. The problem then becomes that of a composite rough surface deforming against a flat rigid one. The first contact is made by the highest composite asperity, followed successively by the next lowest asperities. Each asperity is assumed to deform independently, a reasonable assumption judging by the wide separation of contacts seen in the resin film. By modeling the geometry of the deforming asperities and the rheology of the material comprising the asperities, we can calculate the increase in contact area and applied stress with fracture closure.

Contact area

As each asperity deforms it develops a flat top. Assuming that the asperity volume remains constant, neglecting the porosity created by fracturing, and that the deformation does not extend beyond the limits of the asperity, as suggested by the petrographic observations, we can calculate the area of flattening, that is the contact area. To do so we need to assume the geometry of the asperities. The mathematically simplest shape to use is

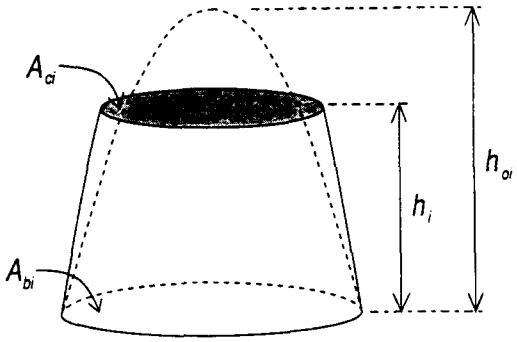

Fig. 9. Paraboloid model for a single asperity. The dashed shape is the undeformed asperity of original height, h_{oi}, and base area, A_{bi}. The solid shape is the flattened asperity with height, h_i, base area, A_{bi}, and contact area, A_{ci}.

the paraboloid (Fig. 9). By equating the volumes of the undeformed and deformed asperities, we can calculate the contact area.

The volume, V_{oi}, of an undeformed paraboloid of height, h_{oi}, and base area, A_{bi}, is $\frac{1}{2}A_{bi}h_{oi}$. The volume, V_i, of the flattened paraboloid with area of flattening, A_{ci}, height, h_i, and base area, A_{bi}, is $\frac{1}{2}(A_{bi} + A_{ci})h_i$. Equating these volumes gives:

$$A_{ci} = A_{bi}\frac{h_{oi} - h_i}{h_i}. \tag{1}$$

The total contact area is $\Sigma_{i=1}^{n} A_{ci}$, where n is the number of asperities that have made contact. We assume that all the asperities lie at the same base level and, in total, fill the space available, these being reasonable assumptions considering the wide variety of sizes available. Then the total fracture surface area is approximately given by $\Sigma_{i=1}^{N} A_{bi}$. The total relative contact area is the ratio of these two quantities, or:

$$\alpha \approx \frac{\Sigma_{i=1}^{n} A_{bi}(h_{oi} - h)/h}{\Sigma_{i=1}^{N} A_{bi}}. \tag{2}$$

Here we set the height of the flat, $h_i = h$, the same for all asperities. We express the contact area as a discrete summation, because of both computational convenience and the discrete nature of the peaks on the surfaces. If the number of asperities is very large, the summations can be replaced by integrals over the continuous height distribution function.

For a paraboloid,

$$A_{bi} = 2\pi\beta_{oi}h_{oi}, \tag{3}$$

where β_{oi} is the radius of curvature of the asperity summit.

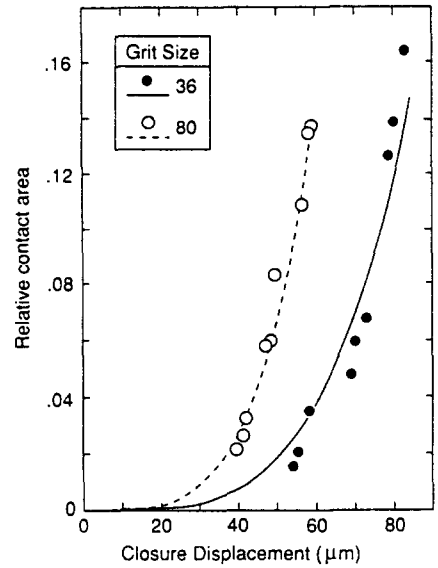

Fig. 10. The variation of relative contact area with closure for marble. The experimental data are shown by closed and open symbols and the theoretical fits are shown by solid and dashed lines.

For simplicity we assume that β_{oi} is the same for all asperities, giving:

$$\alpha \approx \frac{\Sigma_{i=1}^{n} h_{oi}(h_{oi} - h)/h}{\Sigma_{i=1}^{N} h_{oi}}. \tag{4}$$

This equation is a relationship between the relative contact area, α, and height of the flat, h, above the asperity base level. Since closure displacement, δ, rather than h is measured, we make use of the relation: $h = S - \delta$, where S is the initial height of the flat, also called the fracture separation. S is not known independently; rather, we must calculate it from experimental data.

Equation (4) can be tested using our contact area data. We use the known distribution of composite peak heights (Fig. 2), where the heights, h_{oi} and S, are relative to the deepest valley found. Figure 10 shows the experimentally determined relative contact area, α, plotted against fracture closure, δ. By adjusting S, we found the best value which minimizes χ^2, the mean squared deviation of the experimental contact areas. The fits are quite good and give initial separations of 128 and 192 μm for the Nos 80- and 36-grit surfaces, respectively (Table 3). These separations correspond well to the high end of the distribution of asperity peaks found from profiling (Fig. 2).

To prevent any confusion with the concept of initial separation defined in other ways, we emphasize that

Table 3. Final fitting parameters for the paraboloid model applied to the marble data

Grit size	Initial separation (μm)	Goodness-of-fit parameter, χ^2 (area)	(stress)	Initial contact stress, σ_{co} (MPa)	Strain-dependence coefficient, m (MPa)
36	192	0.002	0.6	234	−350
80	128	0.014	0.8	415	−937

here we mean the maximum gap between the two surfaces at the time of first contact. This definition does not correspond to that for the maximum closure, when the two surfaces are in complete contact; deformation of the neighboring asperities will 'fill in' the gap long before the approaching face could make contact. This value of S certainly should *not* be used to calculate the permeability of such fractures.

Applied stress

In order to determine the stress needed for a given closure displacement, we must assume a rheology or constitutive law for the material comprising the asperities. Because marble showed large permanent closure, with relatively minor elastic recovery, we assume that the elastic deformation is negligible. It is not possible to separate out the elastic and inelastic portions of the closure, since the surfaces undergo major alteration during loading. Using the unloading curve or reloading the surfaces for a second cycle would not give the elastic deformation for the original undeformed surface. The assumption that all closure is permanent introduces a closure error of about 10%.

Our evidence suggests that the asperity contact stress decreased with deformation. A simple way to model this is to assume that the contact stress, σ_{ci}, for the ith asperity varies linearly with shortening, or:

$$\sigma_{ci} = \sigma_{co} + m\frac{(h_{oi} - h)}{h_{oi}}, \tag{5}$$

where σ_{co} is the initial contact stress and m is the strain-dependence coefficient.

A more complex rheology, with non-linear effects and strain rate dependence, is unnecessary at this stage of our knowledge. Such complexities can be included, if needed. Both of these material properties, σ_{co} and m, are assumed to be the same for all asperities, a reasonable assumption for monomineralic rocks.

The normal force, F_{ni}, acting across the contact of the ith asperity is $\sigma_{ci}A_{ci}$. The total normal force, F_n, acting on all n contacting paraboloid asperities is then:

$$F_n = \sum_{i=1}^{n} F_{ni} = \sigma_{co}\sum_{i=1}^{n} A_{ci} + m\sum_{i=1}^{n} A_{ci}\frac{(h_{oi} - h)}{h_{oi}}. \tag{6}$$

Using equations (1) and (3), the nominal applied normal stress, σ_a, acting over the entire surface is found to be:

$$\sigma_a \approx \sigma_{co}\frac{\Sigma_{i=1}^{n} h_{oi}(h_{oi} - h)/h}{\Sigma_{i=1}^{N} h_{oi}} + m\frac{\Sigma_{i=1}^{n}(h_{oi} - h)^2/h}{\Sigma_{i=1}^{N} h_{oi}}. \tag{7}$$

As for the contact area, the summations can be replaced by integrals over the appropriate continuous height distribution function. Despite the complexity of this equation, it can be tested quite easily. The only unknowns are σ_{co} and m, the rheological parameters, both of which appear as linear variables with known coefficients. We use the calculated height distribution function, the estimated value of S, and the measured values of σ_a at various closures to find the best fit values of the rheological parameters. The experimental and fitted curves are shown in Fig. 11. The theoretical curves reproduce the general character of the experimental data quite well. The final topographic and rheological parameters used to fit our data, along with the χ^2 values, are shown in Table 3. For computational simplicity, the data for the two roughnesses were fitted independently, although the rheological parameters should be the same for both.

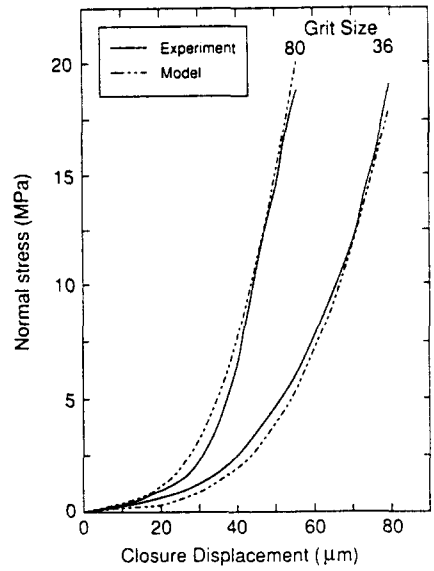

Fig. 11. The loading portion of the closure curves for marble of two different roughnesses, showing the fitted theoretical curves.

Despite the differences in the calculated parameters, two results are worth noting. Firstly, the initial contact stress, σ_{co}, for the marble has an average value of about 325 MPa, approximately one-third of the Vickers indentation hardness for calcite (1050 MPa, Brace 1960). This ratio of yield stress to hardness is that found for plastically deforming metals in tension and compression (Tabor 1951, Bowden & Tabor 1964), suggesting that the calculated value of σ_{co} is at least a reasonable one. σ_{co} is a grain-scale property and cannot easily be related to bulk strength properties such as the unconfined compressive strength of the whole rock. Indeed, σ_{co} is about 10 times the unconfined compressive strength of the marble rock, although it may be close to the differential stress needed for deformation of this rock under sufficient confinement for bulk cataclastic flow to occur. Further work is needed to establish the significance of the contact stress value.

Secondly, for both roughnesses, the strain-dependence coefficient, m, is negative, supporting our earlier suggestion that asperity failure is accompanied by strain weakening. Although the number of contacts increased with load, the loss of load-carrying capacity of the asperities dominated, giving a decrease in the average contact stress (Fig. 8). However, as noted earlier, the average contact stress becomes approximately con-

stant at higher stresses (Logan & Teufel 1986). The number of new contacts diminishes rapidly and the average single asperity area increases. Evidently, at high stresses, m is not constant with closure strain, as we assumed, but may become zero or positive; that is, the asperities may strain harden at large amounts of flattening. At this stage, many of the asperities are thin discs confined between undeforming plates, leading to a much higher strength. This effect is a commonly observed one and is the principle used when halite or pyrophyllite are employed in a solid-medium apparatus to achieve pressures much above the normal strength of the medium. Clearly, then, our simple model is limited to situations where such extreme flattening does not occur. A way to extend the model to include such non-linearity in asperity contact stress is to use a quadratic, or higher order polynomial, equation for the rheological function. The statistical problem becomes one of multilinear regression.

SOME PREDICTIONS

Two additional quantities were measured that were not included in the model: the total number of contacts and the total contact length observed in the profiles. Both quantities can be calculated from the model and provide a check on the validity of the model.

The number of contacts, n_c, at a given load is a function of the distribution of composite peak heights, but not of their shape or material properties:

$$n_c = N \sum_{h_{oi}=h}^{S} f(h_{oi}), \tag{8}$$

where N is the total number of asperities, h_{oi} is the initial height of the ith asperity, h is the current fracture separation, S is the initial separation, and $f(h_{oi})$ is the initial height distribution function of the composite surface.

Since our data are based on profiles, n_c and N are given per unit traverse length; both can be measured from the thin sections of surfaces epoxied under load. The initial height distribution is the Gaussian function fitted to the histogram of peak heights of undeformed surfaces. The measurements and model curves are shown in Fig. 12. Over the range of stresses used, the number of contacts increases approximately linearly with stress and is reasonably well accounted for by the theory.

The total contact length observed in profiles of contacting surfaces can be calculated by assuming an asperity shape. Using equations (1) and (3), we express the contact area for the ith paraboloid asperity as:

$$A_{ci} = 2\pi\beta \frac{h_{oi}(h_{oi} - h)}{h}.$$

Since the contact radius, $R_{ci} = \sqrt{A_{ci}/\pi}$, then

$$R_{ci} = \sqrt{2\beta \frac{h_{oi}(h_{oi} - h)}{h}}. \tag{9}$$

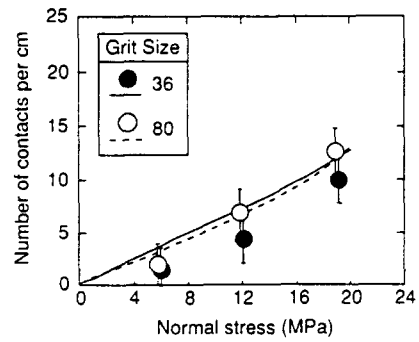

Fig. 12. The variation with normal stress of the number of contacts per cm of profile in marble, as seen in thin section. The range of variation at each stress is shown by bars. The solid and dashed lines are the calculated number of contacts, based on the fitted height distribution function and initial separation derived from the theoretical fit to the relative contact area data.

For any random traverse of the surface, the contacts are crossed along a chord to the circular contact area. For a large number of intersections, the expected length of the chord, l_{ci}, is $\sqrt{2}R_{ci}$. Thus the total contact length, l_c, is

$$l_c = 2\sqrt{\beta} \sum_{i=1}^{n} \sqrt{\frac{h_{oi}(h_{oi} - h)}{h}}. \tag{10}$$

As before, the radius of curvature, β, of all paraboloid asperities is assumed constant, equal to the average for all asperities. β was calculated from the profiles by fitting a circular arc to groups of five data points centered on the peaks and averaging the results.

The measured and calculated contact lengths are shown in Fig. 13. The fit to the data is reasonably good, considering the added uncertainty in the value of β. The paraboloid model tends to overestimate the contact length, a conclusion drawn from data for other rocks, as well (Hannan 1988). The actual asperities are generally more peaked than assumed for the paraboloid, giving smaller contact lengths than was calculated. We developed the model with a cone geometry as well, and found that the actual measurements are bracketed by the two

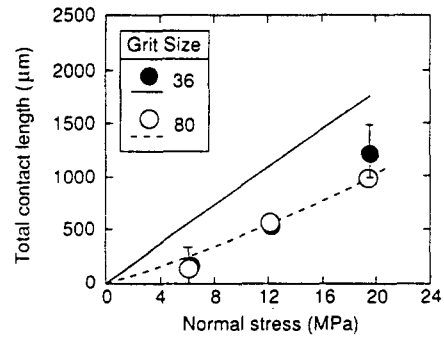

Fig. 13. The variation with normal stress of the total length of contact in marble seen in thin section. Representative sample variations are shown with bars. The solid and dashed lines are calculated from the model, based on the theoretical fits to the contact area and closure data, as well as the asperity tip curvature estimated from the composite profiles.

model predictions (Hannan 1988). These results will be presented in a later paper.

Both the number of contacts and total contact length increased approximately linearly with normal stress. This result means that the average size of contact is nearly constant over the range of stresses used. Consequently, the growth in contact area is accomplished mainly by an increase in the number of contacts. The number of new and, hence, small contacts increases with load, counteracting the growth of the existing contacts and keeping the average size constant. However, at some stress, not much higher than was used in our experiments, the number of new contacts would not increase as rapidly, since the peaks on the lower side of the distribution mean would begin to make contact. At this stress the mean contact size would increase, but the total contact area would increase less rapidly with stress. These predictions describe precisely the observations of Logan & Teufel (1986) of contacts developed during frictional sliding at normal stresses between 10 and 25 MPa. Thus despite the difference in procedure, the results of these two studies agree very well.

At a high percentage of contact area, the large contacts would begin to interfere with each other, a process we call *coalescence*, leading to a further lowering of the area growth rate. Thus, the contact area–normal stress curve (Fig. 7) would be sigmoidal in shape over a larger range of stresses than we used. Further measurements at high stresses will be needed to confirm the validity of our predictions.

IMPLICATIONS FOR ROCK FRICTION

As we stated at the outset, the shear force needed to cause slip on rough surfaces in point contact depends on both the contact area and the compressive and shear strengths of the contacting asperities. The normal stress applied prior to slip will affect these quantities. The contact area increases rapidly with stress and the contacts weaken but at a diminishing rate. The result is an increase in applied shear stress with normal stress, as is commonly observed.

To model this process we assume that the surfaces are unmated and in point contact, with no significant interlocking of opposing asperities. In practice interlocking may occur, but our petrographic observations suggest that it is relatively rare for the roughnesses used in our study. After closure, most contacts on the fairly smooth surfaces become flat and approximately parallel to the overall surface (e.g. Fig. 5b). Where indentation of rigid grains into softer ones occurs (e.g. Fig. 5b) there will be significant additional shear resistance to side motion. In that case, the deformation becomes one of ploughing, a style not modeled here.

Where there is significant mating of the two surfaces, such as in a natural joint, interlocking may become very significant. For example, Reeves (1985) and others before him have shown that the friction coefficient of such surfaces at low normal stresses depends on the contact gradient, that is, the asperity slope. This has been called the dilation angle. Such interlocking may also develop after some slip on unmated surfaces, although we have no direct evidence of it. Since we are only concerned with the onset of slip, such later effects will not be considered.

Although other approaches to modeling the frictional resistance of rough surfaces have been used, we chose an analysis similar to that for normal stress. The nominal shear stress, τ, acting between two rough surfaces in contact is $\tau_c\alpha$, where τ_c is the total contact shear stress supported by the deformed asperities and α is the relative contact area. By analogy with the case for normal stress, τ_{ci}, the contact shear stress for the ith asperity, may vary with the amount of asperity deformation, as

$$\tau_{ci} = \tau_{co} + m^* \frac{(h_{oi} - h)}{h_{oi}}, \tag{11}$$

where τ_{co} is the initial contact shear stress and m^* is the strain-dependence coefficient for shear.

The total shear force, F_s, acting on n contacting paraboloid asperities, each contact having area, A_{ci}, is:

$$F_s = \tau_{co} \sum_{i=1}^{n} A_{ci} + m^* \sum_{i=1}^{n} A_{ci} \frac{(h_{oi} - h)}{h_{oi}}. \tag{12}$$

Using equations (1) and (3), the nominal applied shear stress, τ_a, acting over the entire surface is found to be:

$$\tau_a \approx \tau_{co} \frac{\Sigma_{i=1}^{n} h_{oi}(h_{oi} - h)/h}{\Sigma_{i=1}^{N} h_{oi}} + m^* \frac{\Sigma_{i=1}^{n} (h_{oi} - h)^2/h}{\Sigma_{i=1}^{N} h_{oi}}. \tag{13}$$

The initial frictional resistance can thus be related to the topographic characteristics of the fracture surfaces and the strength of the contacting asperities. We need to know the closure and initial shear strength at various normal stresses for a surface with a known distribution of initial asperity heights.

We can carry the model one step further, by noting the similarity in form of equations (7) and (13). Using equation (4), we have

$$\sigma_a = a\sigma_{co} + bm \tag{14}$$

$$\tau_a = a\tau_{co} + bm^*, \tag{15}$$

where b represents the second summation factor in equations (7) and (13).

Eliminating b gives

$$\tau_a = \frac{m^*}{m}\sigma_a - a\frac{(m^*\sigma_{co} - m\tau_{co})}{m}. \tag{16}$$

Equation (16) shows that initially at low normal stress where the contact area is very small, τ_a increases approximately linearly with σ_a, with a coefficient of friction equal to m^*/m, the ratio of shear and normal strain-dependence coefficients. At higher normal stresses, a increases, so that the second term becomes significant. Countless observations of rock friction show

that this term *must* be positive because τ_a increases less rapidly at high normal stresses. Since α is positive and m is negative, a necessary condition is:

$$m^*\sigma_{co} < m\tau_{co}. \quad (17)$$

Since $\sigma_{co} > \tau_{co}$, then m^* is negative, corresponding to a loss of single-asperity shear strength by collapse during normal closure, as we would expect. Thus the model is in agreement with friction observations and physical intuition.

To use equation (16), we need values for the material properties, τ_{co} and m^*. They can be obtained from equation (13), in the same manner as we determined σ_{co} and m. Another way is to use friction data, along with measurements of contact area. Rearranging equation (16), we get:

$$\alpha m\tau_{co} + (\sigma_a - \alpha\sigma_{co})m^* = m\tau_a. \quad (18)$$

This is a linear equation in τ_{co} and m^*, with the other parameters known from closure and friction experiments. For calcite-bearing rocks, the data from Logan & Teufel (1986) are not sufficient to calculate these two shear parameters, since only two of the three friction experiments are in the range of conditions where our theory appears to be valid. Their two low-pressure measurements give a friction coefficient of about 1, indicating that m^* is about the same as m (from equation 16) and that τ_{co} is approximately equal to σ_{co}.

Without experimental data we cannot carry the model any further. Nor, unfortunately, does the model appear to contribute to the puzzle of *Byerlee's law*, the curious observation that at high normal stresses, almost all rocks have about the same friction strength for the same normal stress (Byerlee 1978). However, the model does provide a framework for understanding the influence of environmental factors on the frictional strength of fractured rock. The geometric characteristics of the surface are isolated in the model from the material properties of the deforming asperities. Time, temperature and reactive fluids will modify the material properties in ways that can be studied without the complications of surface geometry.

CONCLUSION

Despite the simplistic nature of the geometry and the somewhat arbitrary assumption of rheology, this model does a creditable job of accounting for the closure of unmated, rough surfaces under stress and the contact area that develops during that closure. It includes information about the distribution of peak heights, as well as the petrographic evidence of the deformation mechanisms.

This model can form the basis for further study of the normal and shear deformation of fractures in rocks and other materials that deform non-elastically. Although in this case, the deformation was largely by brittle processes, so that strain weakening occurred, the model is sufficiently general that other strain-dependent rheologies can be studied, at least to the extent that a linear strain dependence is approximately correct. It can be easily extended to include more complex dependencies on strain.

The physical significance of the rheological parameters is not known at this time. Application of this model to other materials requires that these parameters be determined for each material. Further study is needed to relate these parameters to other more easily measured properties, such as mineral hardness.

Acknowledgements—We particularly thank Dr J. B. Currie, now retired, for his advice and encouragement in the early developmental stages of this study. Drs P.-Y. F. Robin, H. C. Halls, R. C. Bailey, G. W. Pearce, and the anonymous reviewers contributed many useful comments and discussions for which we are appreciative. M. Mittelstaedt helped design the loading apparatus and profilometer, while his workshop gang built them; A. Sosin made the thin sections, a job made difficult by the presence of the epoxied cuts; S. Jaunzems printed our photographs. Finally, we thank the Earth Sciences Grant Selection Committee of the Natural Sciences and Engineering Research Council of Canada for providing many years of financial support.

REFERENCES

Bowden, F. P. & Tabor, D. 1964. *The Friction and Lubrication of Solids, Part I.* Oxford University Press.

Brace, W. F. 1960. Behavior of rock salt, limestone, and anhydrite during indentation. *J. geophys. Res.* **65**, 1773–1788.

Brown, S. R. & Scholz, C. H. 1985. Closure of random elastic surfaces in contact. *J. geophys. Res.* **90**, 5531–5545.

Brown, S. R. & Scholz, C. H. 1986. Closure of rock joints. *J. geophys. Res.* **91**, 4939–4948.

Byerlee, J. D. 1978. Friction of rocks. *Pure & Appl. Geophys.* **116**, 615–626.

Greenwood, J. A. & Williamson, J. B. P. 1966. Contact of nominally flat surfaces. *Proc. R. Soc. Lond.* **A295**, 300–319.

Hannan, S. S. 1988. An experimental study of fracture closure in elastically and non-elastically deformable rocks. Unpublished Ph.D. thesis, University of Toronto.

Logan, J. M. & Teufel, L. W. 1986. The effect of normal stress on the real area of contact during frictional sliding in rocks. *Pure & Appl. Geophys.* **124**, 471–485.

Pullen, J. & Williamson, J. B. P. 1972. On the plastic contact of rough surfaces. *Proc. R. Soc. Lond.* **A327**, 159–173.

Reeves, M. J. 1985. Rock surface roughness and frictional strength. *Int. J. Rock Mech. Miner. Sci. & Geomech. Abs.* **22**, 429–442.

Stesky, R. M. 1985. Compressional and shear velocities of dry and saturated jointed rock: a laboratory study. *Geophys. J. R. astr. Soc.* **83**, 239–262.

Stesky, R. M. 1986. Electrical conductivity of brine-saturated fractured rock. *Geophysics* **51**, 1585–1593.

Stesky, R. M. & Hannan, S. S. 1987. Growth of contact area between rough surfaces under normal stress. *Geophys. Res. Lett.* **14**, 550–553.

Tabor, D. 1951. *The Hardness of Metals.* Oxford University Press.

Tanoli, S. K. 1982. A study of the mechanics of fracture closure. Unpublished M.Sc. thesis, University of Toronto.

Walsh, J. B. & Grosenbaugh, M. A. 1979. A new model for analyzing the effect of fractures on compressibility. *J. geophys. Res.* **84**, 3532–3535.

Whitehouse, D. J. 1978. The digital measurement of peak parameters on surface profiles. *J. Mech. Engng Sci.* **20**, 221–227.